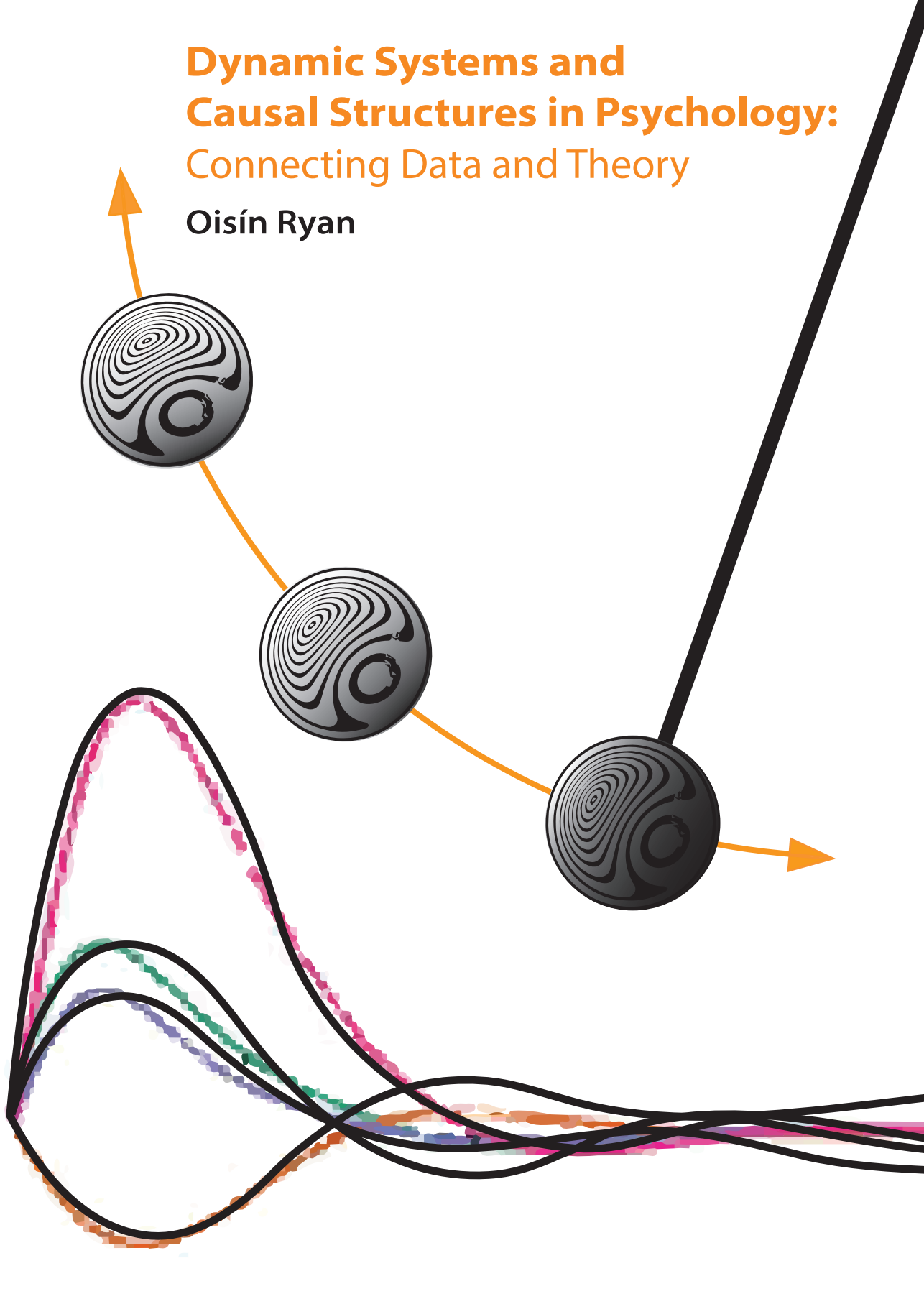
Dynamic Systems and Causal Structures in Psychology: Connecting Data and Theory

Oisín Ryan



Dynamic Systems and Causal Structures in Psychology

Connecting Data and Theory

Dynamische systemen en causale structuren in de psychologie:
De verbinding tussen data en theorie
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

vrijdag 30 oktober 2020 des avond te 6.00 uur

door

Oisín Ryan

geboren op 23 november 1991
te Kilkenny, Ierland

Promotor: Prof. dr. E.L. Hamaker
Copromotor: Dr. R.M. Kuiper

The studies in this thesis were funded by the Netherlands Organization for Scientific Research (Onderzoekstalent Grant 406-15-128)

Beoordelingscommissie:

Prof. dr. E. Ceulemans	Katholieke Universiteit Leuven
Prof. dr. P. de Jonge	Rijksuniversiteit Groningen
Prof. dr. I. Klugkist	Universiteit Utrecht
Prof. dr. H. L. J. van der Maas	Universiteit van Amsterdam
Prof. dr. M. C. Völkle	Humboldt Universität zu Berlin

Dynamic Systems and Causal Structures in Psychology:
Connecting Data and Theory
Proefschrift Universiteit Utrecht, Utrecht.
- Met samenvatting in het Nederlands.

ISBN: 978-94-6416-145-8

Cover design: Luke Keeshan

Print: Ridderprint | www.ridderprint.nl

Contents

Chapter 1

Introduction

Psychological phenomena are best understood as *dynamic processes*: Political beliefs become more or less conservative as we age, the mathematics abilities of children develop over the school year, and the duration and frequency of extreme moods from day-to-day and week-to-week distinguish healthy from unhealthy emotion regulation. As such, the key to understanding psychological phenomena lies in understanding how behaviors, cognition, perceptions, emotions, dispositions, abilities, and all other relevant facets of the mental world evolve, vary and interact with each other over time, within an individual.

This process perspective has witnessed a tremendous growth in popularity, bordering on consensus, in the psychological science literature in the past two decades (Boker, 2002; Van Der Maas et al., 2006; Hamaker, 2012; Molenaar, 2004). In clinical psychology and psychiatry this idea has been particularly transformative, with the traditionally static disease-based conceptualization of mental disorder supplanted in recent years by the view that psychopathologies are inherently complex, multi-dimensional, and dynamic entities (Borsboom, Cramer, Schmittmann, Epskamp, & Waldorp, 2011; Borsboom, 2017; Kendler, 2019; Nelson, McGorry, Wichers, Wigman, & Hartmann, 2017; Wichers, 2014). From this perspective, to understand and eventually control and treat mental disorders we must uncover the mechanistic relationships between psychological processes that underlie psychopathology.

Researchers who subscribe to this theoretical perspective have collected a variety of different types of empirical data with which they hope to study psychological processes. Due to the subject matter of clinical psychology, much of this empirical data comes from non-experimental studies. Two popular categories of non-experimental data can be distinguished. The first of these is *cross-sectional data*, consisting of measurements of psychological processes across many individuals at a single point in time. The second is *intensive longitudinal data*, consisting of repeated measurements of psychological processes over time, for one or more individuals, typically in natural everyday settings. This latter type of data has witnessed a surge in popularity in recent years, in part due to the advent of smartphone technology, and in part motivated by concerns over the difficulty of making inferences from between-person data to within-person processes (Molenaar, 2008; Hamaker & Wichers, 2017).

The core goal of this dissertation is to investigate how researchers can best use non-experimental data on psychological processes to investigate the dynamic mechanisms that give rise to psychopathology. Doing so will depend critically on the *methodology* and more broadly the *methodological framework* that is used to analyze these data.

1.1 Current Methodological Practice and Problems

Researchers who wish to gain insights into psychological processes from non-experimental data primarily do so through the estimation of relatively simple *statistical models* from data. The current dominant methodological framework for this type of modeling in psychology is the *structural equation modeling* (SEM) framework. In the SEM framework, researchers begin by specify their theoretical beliefs in the form of a graphical model describing which variables they believe directly influence which others (Bollen, 1989). These graphical representations imply a set of (usually linear) regression equations, the parameters of which can be estimated directly from data. Two closely related statistical modeling techniques are *statistical network modeling* (Borsboom & Cramer, 2013; Epskamp, Rhemtulla, & Borsboom, 2017; Epskamp, Waldorp, Mõttus, & Borsboom, 2018; Lauritzen, 1996), in which all variables are allowed relate to all others, and *time-series modeling* (Hamaker, Dolan, & Molenaar, 2005; Hamilton, 1994; Molenaar & Campbell, 2009), which is used to investigate relationships between time-lagged variables from longitudinal data.

In each case, researchers working within these frameworks typically aim to garner a greater understanding of the underlying mechanism through the interpretation of estimated model parameters. Typically, dependencies between variables are interpreted as indicative of *direct causal effects* between processes (Granger, 1969; Bollen, 1987; Cole & Maxwell, 2003; Hamaker, Kuiper, & Grasman, 2015; Bulteel, Tuerlinckx, Brose, & Ceulemans, 2016), and the overall mechanism is understood in terms of combinations or summaries of these parameters. However, two critiques of these approaches cast doubt on their suitability for capturing the mechanisms underlying mental disorder.

First, these commonly used statistical models are limited in their ability to capture or represent dynamic relationships, that is, the evolution of processes over time. While we may expect this to be the case for models based on cross-sectional data, even time-series models of longitudinal data typically treat the passage of time only with respect to the ordering of measurements, ignoring for instance the length of the time-interval that elapses between measurement waves. Gollob and Reichardt (1987) describe a classic example of why this is problematic. Imagine that we wish to study the effect of taking aspirin on reducing headache pain. This effect may be zero two minutes after ingestion, substantial after thirty minutes, strong after two or three hours, reduced after five hours and zero again twenty-four hours later. To truly understand the relationship between aspirin and headache levels necessitates that we understand the full dynamic picture of how that relationship evolves over time. Interpreting a single parameter, exclusive to a particular interval, as *the* effect of one process on another, as we would using common statistical modeling approaches, leaves us with at best a critically incomplete understanding of the underlying mechanism. This problem will necessarily be exacerbated when researchers wish to model multidimensional systems of dynamic processes, and so the suitability of these methods for that purpose is questionable.

The aspirin example is also helpful in highlighting the second problem with

current practice, that is, the difficulty of using statistical models based on observational data to infer causal relationships. Ideally, we would study the aspirin-headache relationship with an experiment. If we randomly assign headache-suffering participants to take either an aspirin or a placebo, and we observe that aspirin takers have lower headache levels two hours later, we can be reasonably certain that this reduction was *caused* by the aspirin. However, if we are unable to conduct such an experiment, we must instead rely on observations of aspirin intake and headache levels in everyday life. Suppose that the analysis of this data returns a negative statistical dependence between aspirin and headaches. When (if ever) can we draw the same causal conclusion from this information as we would have based on the experimental study? Unfortunately, the traditional statistical modeling literature provides few if any reasonable answers to this question. Intuitive approaches which equate the fit of statistical models with their causal veracity have variously been described as "radically suboptimal", yielding "garbage-can" models, and the very intuition behind this notion has been labeled an urban-myth (Spirtes et al., 2000; Achen, 2005; Spector & Brannick, 2011; Rohrer, 2018). Without a principled way to link statistical with causal information, the utility of standard approaches for uncovering causal mechanisms is left on somewhat uncertain footing.

1.2 Alternative Methodological Frameworks

In this dissertation I will explore how we can improve current statistical modeling approaches in psychology by addressing the limitations set out above. To achieve this, I will look to other disciplines that have grappled with similar issues from different methodological perspectives. In particular the chapters of this dissertation focus on exploring, borrowing and adapting ideas from two distinct methodological frameworks.

First, fields as diverse as physics, climatology, ecology, biology, chemistry and engineering have used *dynamical systems theory* to understand and describe phenomena that vary over time (Strogatz, 2015). In this framework, dynamic mechanisms are described using the language of *differential equations*, breaking down the evolution over time of processes into the fundamental building blocks of moment-to-moment interactions. In so doing differential equations allow a tremendous explanatory depth and breadth: Relatively simple differential equations can capture complicated patterns of dynamics, a modeling strategy that has been successful in helping us to understand phenomena as diverse as the motion of the earth around the sun (Newton, 1687), the relationship between predator and prey populations in the wild (Volterra, 1931) and the reaction of the immune system to the HIV virus (Ho et al., 1995). In comparison to the simple statistical models described above, differential equations provide us with a rich and flexible formalism for modeling dynamic processes, allowing us to describe and explore how they evolve and vary continuously as a function of time.

Second, in fields such as epidemiology, econometrics and computer science, researchers have long struggled with the issue of inferring causal relationships

from observational data. In these fields, the different approaches which have been developed to tackle this issue can broadly be described as the *interventionist causal inference* framework, so called due to the definition of causal effects in terms of hypothetical experimental interventions (Rubin, 1974; Greenland & Robins, 1986; Angrist, Imbens, & Rubin, 1996; Pearl & Verma, 1991; Pearl, 2009). This framework provides a mathematical language to describe causal structure, formalizing the notion that statistical relationships (which describe patterns in data) exist on a different level of explanation than causal relationships (which describe the processes responsible for producing data). This in turn allows us to answer the question of how and when statistical dependencies can be used to infer causal structure, but moreover provides a new way to approach statistical analyses when causal relationships are the target of inference.

Although these two approaches are distinct from one another, and indeed tackle distinct shortcomings present in current practice, both the dynamical systems and interventionist causal inference frameworks offer promising avenues by which we might improve our understanding of the mechanisms underlying mental disorder. The chapters in my dissertation form an exploration of what we can learn from these different approaches, and how we can use these lessons to improve current practice.

1.3 Outline and Summary

The remainder of this dissertation consists of five chapters, each addressing some problematic aspect of how current statistical modeling approaches are used to investigate psychological processes, primarily based on non-experimental data, and with a focus on the domains of clinical psychology and psychiatry. In each chapter potential solutions to problems are offered inspired by either the methodological frameworks of dynamical systems theory, interventionist causal inference, or a mix of approaches borrowed from both traditions. The chapters themselves are organized with respect to the type of data and approach used as well as the complexity with which the underlying mechanism is modeled.

Chapter 2 critically evaluates the use of statistical network models, fit on cross-sectional data, in order to infer patterns of directed causal relationships between variables. This practice is evaluated from the interventionist causal inference perspective, revealing the inherent difficulties that are present in making this type of inference even under ideal conditions. Moreover, this chapter introduces a newly developed tool that aids researchers in exploring the different graphical causal structures which may underlie a given statistical network model, in principle aiding the generation of causal hypotheses. The discussion section of this paper highlights the necessity of properly accounting for the time dimension if researchers wish to infer patterns of causal dependencies between dynamic processes. In the following chapters this is addressed by focusing largely on methods of modeling dynamic relationships based on intensive longitudinal data.

Chapter 3 and Chapter 4 examine how statistical models based on differen-

tial equations, which are referred to throughout as *continuous-time* models, can be used to improve the modeling of psychological processes in comparison to standard time-series approaches. Chapter 3 focuses on a very simple continuous-time model, providing a short treatment of the key practical and conceptual advantages of this model. Some core concepts from dynamical systems theory are introduced, and the interpretation of the continuous-time model is illustrated with a bivariate empirical example. Chapter 4 introduces a continuous-time approach to dynamical network analysis, and develops new centrality measures specifically for these networks. Inspired by the interventionist causal inference literature, these measures allow researchers to identify the optimal target for different types of interventions, either acute or continuous in nature.

The remaining chapters deal with the use of differential equations to model more complex dynamics we may expect to operate between psychological processes. Chapter 5 revisits the idea of using statistical models based on observational data to infer patterns of causal relationships between variables (as in Chapter 2) when the structure of the underlying system is unknown. Here, however, the data-generating mechanism takes the form of a bistable dynamical system, a particular type of system which has received considerable theoretical attention in clinical psychology, and which can be formalized by a set of non-linear differential equations. It is shown that, for highly idealized simulated data, some statistical models can be used to recover so-called global dynamics (i.e. more stable characteristics of the system) but that it is difficult to correctly recover the micro-dynamics (i.e. moment-to-moment relationships). Repeating the same analyses for simulated data with a realistic sampling frequency showed that while global dynamics may still be recovered, the recovery of micro-dynamics was unsuccessful. These results highlight both the difficulty of making inferences from statistical models without a strong theory, as well as the fundamental role that sampling frequency plays in statistical modeling of dynamic processes.

In Chapter 6, it is argued that *formal theories* are critically necessary to facilitate the study of mental disorders as complex dynamical systems. Formal theories, expressed in the language of differential equations, are common in fields that apply dynamical systems theory but almost absent from the clinical psychology literature. The necessity of formal theories is supported by a short review of the contemporary philosophy of science literature. Following this, three possible routes by which statistical models could be used to obtain formal theories are investigated. The first route, interpreting statistical models directly as formal theories (as in Chapters 3 and 4) and the second route, using common statistical models to infer characteristics of the underlying mechanism (as in Chapters 2 and 5), are both shown to have fundamental shortcomings. The third route, using statistical models to help further develop an existing formal theory, is shown to be most promising. The chapter concludes by proposing a framework for the generation, development and testing of formal theories, detailing the role that statistical models play at each step of this process.

This last chapter represents the culmination of developments in the dissertation, synthesizing the different challenges and approaches described in the previous chapters, and utilizing these to inform a framework for formal theory con-

struction. The dissertation ends with a clear message for the fields of clinical psychology, psychiatry and the methodologists who work within these fields: If we hope to understand the dynamic processes that give rise to psychopathology, a radical reorientation of current research practice towards formal theory development is urgently needed.

Chapter 2

The Challenge of Generating Causal Hypotheses Using Network Models

Abstract

The network approach to psychopathology is a theoretical framework in which mental disorders are viewed as arising from direct causal interactions between symptoms. To investigate such networks, researchers typically estimate undirected *network models* from empirical data, called *Pairwise Markov Random Fields* (PMRFs), or for normally distributed variables, *Gaussian Graphical Models* (GGMs). In this paper, we critically evaluate the use of PMRF-based methods to *generate causal hypotheses* about an underlying directed causal structure. We argue that hypothesis generation is critically dependent on the specification of a *target causal structure*: This is generally absent from applications of PMRFs, researchers instead taking a *causally-agnostic* approach. We show that the agnostic approach is fundamentally problematic, since the heuristics typically used for hypothesis generation do not hold for all types of causal structure. The specification of a target structure, however, allows a principled approach to hypothesis generation and yields novel insights. We illustrate this using the (weighted) *Directed Acyclic Graph* (DAG) as the target structure. We review the relationship between PMRFs and DAGs, showing that hypothesis generation using heuristics alone is non-trivial: Many different DAGs can result in the same PMRF, and these differ in their substantive interpretations. With an empirical example, we illustrate how the GGM can be used to generate more informed causal hypotheses, by exploring the *equivalence set* of weighted DAGs. This is aided by a novel tool implemented in an *R* package. Finally, we discuss additional barriers to discovering causal relationships in practice, and possible alternative formalisms for causal structure.

This chapter has been adapted from: Ryan, O., Bringmann, L. F. & Schuurman, N. K. (under review). The Challenge of Generating Causal Hypotheses Using Network Models. Pre-print: https://psyarxiv.com/ryg69/. Author contributions: OR conceptualized the initial project, wrote the paper and *R* code and ran the analyses. LFB and NKS helped further develop the ideas in the project, discussed progress and provided textual feedback.

2.1 Introduction

The network approach to psychopathology is a theoretical framework in which mental disorders are viewed as arising from direct causal interactions between symptoms (Borsboom & Cramer, 2013; Borsboom, 2017). In practice, researchers often aim to uncover aspects of the underlying network structure by estimating network (i.e. graphical) models, generally from cross-sectional data (Van Borkulo et al., 2014; Epskamp, Waldorp, et al., 2018; Epskamp, Borsboom, & Fried, 2018). In these instances, researchers typically estimate a *Pairwise Markov Random Field* (PMRF); for normally distributed variables this takes the form of a *Gaussian Graphical Model* (GGM). These are network models with undirected connections between variables, representing their conditional relationship (i.e., partial correlation) controlling for all other variables in the network.

The PMRF is often promoted as an exploratory method of *generating causal hypotheses* about the underlying network structure (Borsboom & Cramer, 2013; Epskamp, van Borkulo, et al., 2018). In many instances, this data-generating structure is conceptualized as consisting of *directed causal relationships*, and the PMRF is taken to reflect the *causal skeleton*, identifying the presence, but not the direction, of direct causal links (e.g., van Borkulo et al., 2015; Boschloo, Schoevers, van Borkulo, Borsboom, & Oldehinkel, 2016; Haslbeck & Waldorp, 2018). Typically researchers are agnostic regarding what specific form the underlying causal structure takes, for instance, whether it consists of uni-directional or bi-directional relationships, and whether the structure is cyclic or acyclic (Cramer, Waldorp, van der Maas, & Borsboom, 2010; Epskamp, Waldorp, et al., 2018; Isvoranu et al., 2016; McNally et al., 2015; Robinaugh, Millner, & McNally, 2016; Van Borkulo et al., 2014; Costantini et al., 2015).

However, this agnostic approach to causal structure means that the task of causal hypothesis generation is fundamentally intractable: If the class of casual structure we wish to make inferences about is not clearly defined, it is impossible to know how to go about generating hypotheses about that structure. In other words, the causal information conveyed by the absence or presence of connections in a PMRF depends *entirely* on the precise mapping from that network to the underlying directed causal structure.

The uncertainty resulting from this causal-agnosticism is typified by the discussion surrounding Directed Acyclic Graphs (DAGs) in the network analysis literature. DAGs are a popular approach to conceptualizing directed causal structures in the causal inference literature, and one for which the relationship with PMRFs is both relatively simple and well-known (Spirtes et al., 2000; Pearl, 2009; Lauritzen, 1996). Typically, the hypothesis generation heuristics suggested by users of PMRFs in psychology are consistent with, and even seem to be derived from, the DAG as an underlying structure (see Epskamp, Waldorp, et al., 2018 p.457-458; Epskamp, van Borkulo, et al., 2018 p.420, and Borsboom & Cramer, 2013 p.105). Simultaneously, however, the DAG is rejected as a plausible target structure, for example due to the absence of cyclic effects. This represents a fundamental problem for researchers wishing to generate causal hypotheses: It is unclear whether these same heuristics still apply when mapping PMRFs to some

other undefined class of causal structure.

In this paper, we critically evaluate the use of PMRFs as a causal hypothesis generating tool. To this end, we outline the importance of specifying a target causal structure when generating such hypotheses, and describe how critically hypothesis generation relies on the relationship between the target structure and the PMRF. Following this, we evaluate how PMRFs can be used to generate causal hypotheses once a target structure is defined, using the DAG as an example of such a target. In taking this approach, we show that even in the context of the simple mapping between DAGs and PMRFs, generating causal hypotheses from a PMRF is not trivial, as a variety of distinct DAG structures can lead to the same PMRF (Lauritzen, 1996; Andersson, Madigan, Perlman, et al., 1997; Raykov & Marcoulides, 2001; MacCallum, Wegener, Uchino, & Fabrigar, 1993). This complicates the generation of causal hypotheses for two reasons. First, variables may be connected in the PMRF *even when there is no directed causal relationship between them*; Second, different hypotheses researchers make based on connections in the PMRF may not be compatible with any one underlying structure.

The remainder of the paper is structured as follows. First, we introduce the necessary background on the two graphical models, PMRFs and DAGs, which will be discussed in the remainder of the paper. Second, we detail the necessity of specifying a target causal structure by a) arguing in principle against taking a causally-agnostic approach, and b) evaluating the utility of PMRFs when a target structure *is* specified, using the DAG as that target. Third, we illustrate how PMRF-based methods can be used to generate causal hypotheses once a target structure is chosen, using an example based on a published GGM analysis. This empirical illustration is aided by a novel tool which allows us to generate the set of structures which map to a given GGM under ideal conditions. Finally, we discuss some routes forward in using network models to explore causal structure.

2.2 Background

In this section we will give an overview of network models, also known as graphical models, two terms which we will use interchangeably in the current article. Specifically we will review two instances of graphical models under consideration in the current paper: The Pairwise Markov Random Field (PMRF), which is used in the network analysis literature to generate causal hypotheses, and the Directed Acyclic Graph (DAG), which we will use as an example target causal structure. The remainder of this paper will make use of this background to allow for an informed evaluation of PMRFs as hypothesis-generating tools, under two scenarios: First, when the target casual structure is unspecified (i.e., the causally-agnostic approach); Second, when the target causal structure is a DAG.

For both the PMRF and DAG we describe a special case which can be obtained when the variables in question have a joint Gaussian (normal) distribution: the Gaussian Graphical Model (GGM) and a weighted DAG based on linear regression. Finally, we describe the relationship between PMRFs and DAGs, as described in the graphical modeling literature (Lauritzen, 1996).

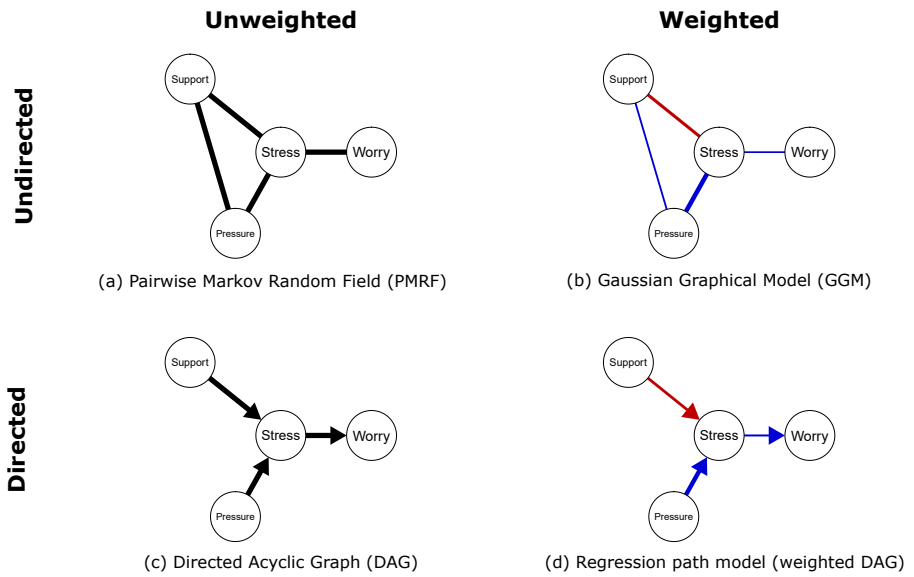


Figure 2.1: Four different types of network (graphical model), arranged by their edge-characteristics.

2.2.1 Networks and Graphs: A Primer

A network (or graph) is made up of a set of *nodes* (or vertices), and a set of *edges* which encode the connections between pairs of nodes (Lauritzen, 1996; Newman, 2018). In the current paper, we will focus on applications of empirical network models in psychology, in which the nodes represent a set of p variables, denoted $\boldsymbol{X}$, and the edges in the network represent statistical relationships between these variables. Typically these relationships are not directly observed, but must be estimated from data. In Figure 2.1, the nodes in each network are four variables related to burnout: levels of social support (Support), work pressure (Pressure), Stress, and Worry.

Graphs can be described as either directed or undirected, and weighted or unweighted, depending on the characteristics of the edges in the graph. In unweighted graphs, edges are either present or absent (as in Figure 2.1(a) and (c)) and in weighted graphs edges have a particular value attached to them. In Figure 2.1(b) and (d) the thickness of each edge represents the absolute value of the weight, and the color represents the sign (blue for positive, red for negative). The edges in a given graph are collected in an adjacency matrix (for unweighted graphs) or weights matrix (for weighted graphs).

In each type of graphical model considered here, the edges represent a particular type of *conditional dependence* relationship between pairs of variables in the network, that is, the relationship between two variables conditional on (also referred to as "controlling for" or "keeping constant") a particular subset of other variables in the graph. The exact nature of the relationship described by the edges

depends on the graphical model under consideration.

2.2.2 The Pairwise Markov Random Field and Gaussian Graphical Model

A PMRF is an undirected, unweighted graph in which the absence of an edge between a pair of variables X_i and X_j denotes that this pair is *independent* when conditioning on the set of all other variables in the network $\boldsymbol{X}^{-(i,j)}$

$$X_i \perp\!\!\!\perp X_j | \boldsymbol{X}^{-(i,j)} \tag{2.1}$$

where $\perp\!\!\!\perp$ denotes independence between a pair of random variables (Dawid, 1979; Lauritzen, 1996).[1] An example of a PMRF is shown in Figure 2.1(a). In this graph, Worry is only connected to Stress. This denotes that Worry is independent of Pressure when we condition on Stress and Support (Worry $\perp\!\!\!\perp$ Pressure | Stress, Support). Similarly, we can say Worry is dependent on Stress when we condition on Pressure and Support (Worry $\not\perp\!\!\!\perp$ Stress | Pressure, Support).

The GGM is a particular type of weighted PMRF for variables following a Gaussian distribution, where the edge weights indicate the strength of the linear conditional relationship between a pair of variables, controlling for all others (Dempster, 1972; Cox & Wermuth, 1996; Lauritzen, 1996). An example of a GGM is shown in Figure 2.1(b). The positive edge connecting Stress and Worry denotes that, keeping Support and Pressure constant, high scores for Stress tend to co-occur with high scores for Worry. Typically the weights matrix of the GGM is given by the matrix of *partial correlations*.

To obtain a GGM from a set of p variables with a joint Gaussian distribution $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, researchers use the *precision matrix*, the inverse of the variance covariance matrix

$$\boldsymbol{\Omega} = \Sigma^{-1} \tag{2.2}$$

The precision matrix describes the conditional dependency relationships in Equation (2.1) for a set of Gaussian variables in the following way: If an element of this matrix is zero ($\omega_{ij} = 0$) this means that the two variables pertaining to that element are conditionally independent, given all other variables in the network ($X_i \perp\!\!\!\perp X_j | \boldsymbol{X}^{-(i,j)}$), while non-zero off-diagonal elements mean two variables are not conditionally independent ($\omega_{ij} \neq 0 \Rightarrow X_i \not\perp\!\!\!\perp X_j | \boldsymbol{X}^{-(i,j)}$). The matrix of partial correlations is found by standardizing the precision matrix and multiplying the off-diagonal elements by −1 (for further details see Dempster, 1972; Epskamp et al., 2017).[2]

In psychological applications, the GGM is most often estimated from cross-sectional data (although it is sometimes also fit on the residuals from a time series model, cf. Epskamp, Waldorp, et al., 2018). Typically the precision matrix

[1] PMRFs are also sometimes referred to as Conditional Independence Graphs (e.g., M. Kalisch & Bühlmann, 2007). As we deal with two different types of graphs which both describe particular conditional (in)dependence statements, we use the terminology of PMRF throughout.

[2] The relationship between a precision matrix and a partial correlation matrix is the similar to the relationship between a covariance matrix and a correlation matrix. The partial correlation matrix is easier to interpret directly, and so is more often used to visualize the results of a GGM analysis.

is estimated directly from the data using regularization techniques such as the graphical lasso (Friedman, Hastie, & Tibshirani, 2008; Epskamp & Fried, 2018). These techniques introduce bias in the parameter estimates in order to avoid over-fitting. The graphical lasso technique returns an estimate of the precision matrix $\hat{\mathbf{\Omega}}$ in which small values are set exactly to zero (Friedman et al., 2008). The presence of exact zero values in this matrix means that the estimated precision matrix, and thus the corresponding GGM, is often simpler (more sparse) than the true population precision matrix.

2.2.3 The PMRF in Practice

Epskamp, Waldorp, et al. (2018) outline three potential uses for PMRF models in psychology. The first is to use the PMRF to investigate purely predictive relationships: For example, in Figure 2.1(a), Support is conditionally dependent on both Stress and Pressure, but independent of Worry conditional on those two variables. This tells us that, in order to optimally predict Support, we would need information on Stress and Pressure, but that information on Worry is not needed. Second, the PMRF can be interpreted directly as an undirected data-generating mechanism. This approach has been used for the binary-variable Ising model both in the statistical physics literature (Murphy, 2012) and in the context of theoretical toy models in the psychology literature (Cramer et al., 2016; Dalege et al., 2016; Borsboom, 2017). Third, Epskamp, Waldorp, et al. suggest that the PMRF can be used as a causal hypothesis generating tool: That is, taking the undirected edges in the PMRF as indicative of the presence of a directed causal effect in the underlying causal structure (e.g., Borsboom & Cramer, 2013; Boschloo et al., 2016; Deserno, Borsboom, Begeer, & Geurts, 2017; Knefel, Tran, & Lueger-Schuster, 2016; Fried, Boschloo, et al., 2015). It is this use of the PMRF that we will critically evaluate in the remainder of this paper.

It seems intuitive that the conditional dependencies estimated by the PMRF are somehow informative about the underlying directed causal structure. However, to evaluate how well PMRFs perform in generating causal hypotheses, we need to specify the form of that underlying causal structure, and understand how the PMRF relates to it. Epskamp, Waldorp, et al. (2018) primarily discuss the relationship between the PMRF and the Directed Acyclic Graph (DAG). DAGs (also referred to as Bayesian networks) are directed graphical models consisting of unidirectional edges $X_i \rightarrow X_j$, and have proven widely popular as a way of conceptualizing causal structures in formal approaches to causal inference (Spirtes et al., 2000; Pearl, 2009; Dawid, 2002; Richardson & Robins, 2013; Galles & Pearl, 1995; Dawid, 2010). However, it is evident from the literature that many researchers use the PMRFs for hypothesis generation while taking what can be described as a *causally-agnostic* approach: Rejecting the notion of a DAG as the underlying structure, but without specifying an alternative formalism (e.g., Cramer et al., 2010; Isvoranu et al., 2016; McNally et al., 2015; Robinaugh et al., 2016; Van Borkulo et al., 2014; Costantini et al., 2015).

In the following sections we will evaluate the use of the PMRF to generate causal hypotheses, both in the causally-agnostic setting, and when a DAG is taken

as the target causal structure. To facilitate this, in the following sub-section we introduce the necessary background on DAGs, and the relationship between the DAG and PMRF.

2.2.4 Directed Acyclic Graphs

The structure of a DAG describes which variables in $\boldsymbol{X}$ are conditionally dependent and independent from one another. The DAG structure is often described in "familial" terms: Directed edges $X_i \rightarrow X_j$ connect *parent* nodes or causes (X_i) to *children* nodes or effects (X_j). Nodes which share a common child, but are not directly connected to one another, are termed *unmarried* parents. An example of a DAG is shown in Figure 2.1(c), where Support and Pressure are unmarried parents of Stress, and Stress is a parent of Worry. Both children and children *of* children (and so forth) are called *descendants*, and parents and parents *of* parents (and so forth) are called *ancestors*. In the graph shown in Figure 2.1(c) Worry is a descendant of its ancestors Support, Pressure and Stress.

A DAG describes the conditional dependencies present in a set of random variables $\boldsymbol{X}$ according to the causal *Markov* condition (Spirtes et al., 2000). This condition states that each variable X_i is conditionally independent of its non-descendants, given its parents

$$X_i \perp\!\!\!\perp \boldsymbol{X}^{-de(i)} | \boldsymbol{X}^{pa(i)}, \tag{2.3}$$

which can be used to read off conditional (in)dependencies between pairs of variables from the DAG. For example, in Figure 2.1(c), we can derive that Support and Pressure are *marginally independent* (Support $\perp\!\!\!\perp$ Pressure $|\emptyset$), because Pressure is a non-descendant of Support, and Support has no parents in this graph ($pa(\text{Support}) = \emptyset$). Any two nodes connected by an edge are dependent conditional on any subset of other variables.

Further conditional (in)dependency relationships between any pair of variables in the graph can be derived from the structure of a DAG using so called *d-separation* rules (Pearl, 2009). These rules allow us to relate DAG structures to other graphical models such as the PMRF. The most important of the d-separation rules for the current paper relates to situations in which two variables share a common child, also known as a common effect or *collider* structure $X_i \rightarrow X_k \leftarrow X_j$. According to d-separation rules, the parent variables X_i and X_j are dependent conditional on the collider variable X_k. For example in Figure 2.1(c), although Support and Pressure are marginally independent, they are dependent when conditioning on Stress (Support $\not\perp\!\!\!\perp$ Pressure | Stress). For substantive examples applying d-separation rules in social science settings, readers are referred to Glymour (2006).

A weighted DAG can be obtained from a set of Gaussian variables $\boldsymbol{X}$, assuming linear relationships between these variables, using a linear regression path model in which child nodes are predicted by their parents

$$\boldsymbol{X} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{X} + \boldsymbol{e} \tag{2.4}$$

where $\boldsymbol{B}$ is a $p \times p$ matrix of regression weights, $\boldsymbol{a}$ represents a $p \times 1$ vector of intercepts, and $\boldsymbol{e}$ represents a $p \times 1$ vector of residuals, which are assumed to have a Gaussian distribution with mean zero and diagonal variance-covariance $\boldsymbol{\Psi}$ (i.e. the residual terms are uncorrelated). This type of weighted DAG is exactly equivalent to a linear structural equation or path model. In a weighted DAG, as shown in Figure 2.1(c), the matrix of regression coefficients $\boldsymbol{B}$ serves as the weights matrix. Here we can see for instance that Support has a moderate negative effect on Stress, and Pressure has a stronger positive effect on Stress. Crucially, while the GGM can be estimated directly from data, a unique weighted DAG can typically only be obtained from data if the structure of the unweighted DAG is known (Levina, Rothman, Zhu, et al., 2008; Shojaie & Michailidis, 2010).

2.2.5 From DAG to PMRF: Moral Graphs and Skeletons

So far we have reviewed two types of graphical models, PMRFs and DAGs, and the conditional dependency relationships described by each. From the graphical modeling literature we know that there is a straightforward relationship between the structure of a DAG and the structure of a PMRF (Lauritzen, 1996; Spirtes et al., 2000). Specifically, when the underlying causal structure is a DAG, then the corresponding PMRF is equivalent to the *moral graph* of that DAG. The moral graph is an undirected graph obtained by first "marrying" (i.e. drawing an edge between) all "unmarried parents" in the DAG, and then replacing all directed edges with undirected edges. The moral graph therefore contains an undirected edge if either a) these two nodes are connected by a directed edge in the DAG or b) these two nodes share a collider (Wermuth & Lauritzen, 1983; Lauritzen, 1996; Spirtes et al., 2000).[3] The equivalence between PMRFs and moral graphs can be seen when comparing the DAG and the PMRF in Figures 2.1(a) and (c). Here, the PMRF contains an undirected version of all of the edges in the DAG, in addition to an edge connecting the unmarried parents Support and Pressure, identical to the moral graph of that DAG.

Crucially, the moral graph of a DAG should not be confused with the *skeleton* of a DAG. The *skeleton* of a DAG is the undirected graph obtained by replacing the directed edges in a DAG with undirected edges: The skeleton contains an edge between two nodes *if and only if* the underlying DAG contains an edge between those nodes (Spirtes et al., 2000). The skeleton describes exactly which variables do and do not share a connection, but does not contain information on the directionality of that connection. In general the moral graph will contain more edges than the skeleton, with additional edges induced in the moral graph whenever there is a collider with unmarried parents.

The difference between the moral graph (PMRF) and the skeleton is shown for three example DAGs in Figure 2.2. Each of the three DAGs shown in the first column share the same skeleton, as shown in the second column. The third col-

[3] A collider structure $X_i \rightarrow X_k \leftarrow X_j$, which does not contain an edge directly connecting the parents ($X_i \nrightarrow X_j$ and $X_i \nleftarrow X_j$) is called an *open v-structure*. If there are any open v-structures in the DAG, then the moral graph must contain an undirected edge between the relevant parents $X_i - X_j$. Thus, the moral graph "marries" unmarried parents.

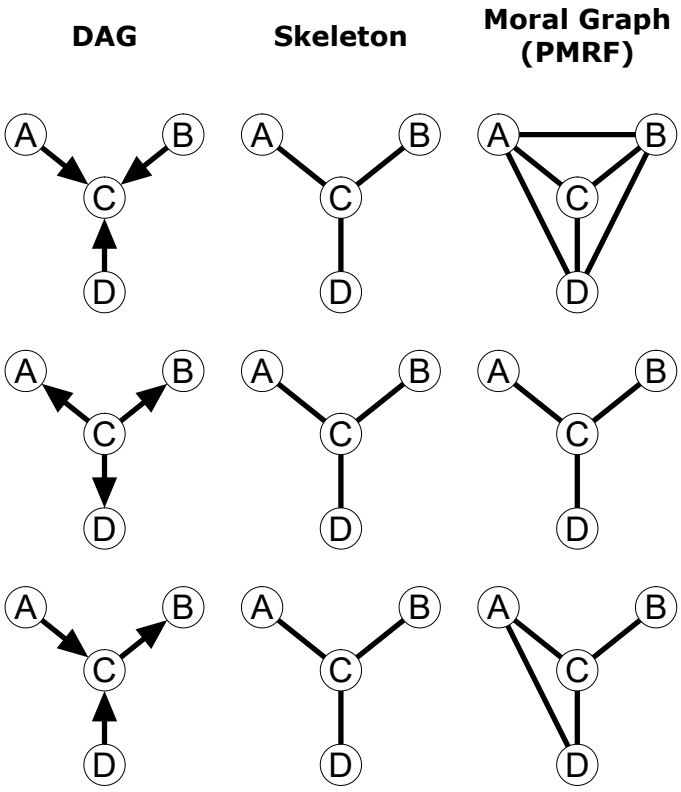


Figure 2.2: Three examples of DAGs which all have the same skeleton, but result in different moral graphs depending on the orientation of the edges in the DAG.

umn of Figure 2.2 shows the moral graph for each DAG. We see that although all three DAGs share the same skeleton, each moral graph is distinct. Furthermore, the moral graph contains additional edges connecting variables *which are not connected in the DAG*, apart from the DAG in the second row of Figure 2.2, in which there are no collider structures.

2.3 Using PMRFs to Generate Causal Hypotheses

Having reviewed the exact relationship between DAGs and PMRF models in the previous section, we can now evaluate the practice of using PMRF-based models to generate hypotheses about an underlying directed causal structure. First, we outline the fundamental problem of generating causal hypotheses from PMRFs when the form of the target causal structure is unspecified (i.e., the causally-agnostic approach). This discussion motivates our focus on DAGs as candidate causal structures. In the remainder, we make use of DAGs to outline how PMRFs

can be used to generate causal hypotheses once the target structure is specified. Specifically, we review what the mapping between DAGs and PMRFs described above means for hypothesis generation, address the particular challenges which arise when using the PMRF to infer multivariate patterns of relationships in an underlying DAG, and describe some additional assumptions which are needed to make causal hypothesis generation feasible in practice.

2.3.1 Causal Hypotheses and Unspecified Causal Structures

As outlined previously, in the psychology literature PMRF-based models are often interpreted as indicative of some true underlying causal structure, with the presence of an edge taken as a necessary but not sufficient condition for a causal relationship between two variables. The underlying causal structure itself is often described in terms of *directed* causal relationships between variables (e.g. insomnia $\rightarrow$ fatigue, support $\rightarrow$ stress) and the PMRF is promoted as a tool by which to generate hypotheses about this directed structure (Borsboom & Cramer, 2013; Epskamp, Waldorp, et al., 2018; Epskamp, van Borkulo, et al., 2018). Assuming that all variables involved in the causal system have been observed (i.e. no unobserved common causes), the heuristics which are supplied for this hypothesis generation task can be broadly summarized as follows:

> *Heuristic 1:* An edge between two variables in the PMRF ($X_i - X_j$) indicates that two variables share either a direct causal link ($X_i \rightarrow X_j$ or $X_i \leftarrow X_j$) or a common effect ($X_i \rightarrow X_k \leftarrow X_j$)
>
> *Heuristic 2:* The absence of an edge between two variables ($X_i \nsim X_j$) indicates that these two variables do not share a direct causal link ($X_i \nrightarrow X_j$ and $X_i \nleftarrow X_j$)

Notably, these heuristics are consistent with treating the PMRF as the *moral graph* of an underlying DAG, as described in the previous section. Furthermore, these heuristics are often derived explicitly with reference to relationship between PMRFs and DAGs (Borsboom & Cramer, 2013; Epskamp, Waldorp, et al., 2018).

However, these same heuristics are typically described and applied by researchers who simultaneously reject the possibility that the underlying structure is a DAG, for instance due to the hypothetical presence of causal "loops" $X_i \leftrightarrows X_j$ (e.g., Cramer et al., 2010; Isvoranu et al., 2016; McNally et al., 2015; Robinaugh et al., 2016; Van Borkulo et al., 2014; Costantini et al., 2015). Moreover, beyond the presence of such hypothetical loops, an agnostic approach to the underlying causal structure is typically taken, in that the precise form of this alternative structure is left unspecified. This represents a fundamental contradiction: Without specifying the form of the underlying causal structure, it is impossible to verify whether these heuristics apply outside of the case used to derive them in the first place.

In fact, we know for certain that for some types of causal structure, the heuristics relating conditional dependencies (in the form of PMRF edges) to causal dependencies described above do not hold. For instance, take it that the target

causal structure takes the form of directed relationships between dynamic, time-varying processes, such as described by a *Local Independence Graph* (Schweder, 1970; Aalen, 1987; Didelez, 2000). This type of structure allows us to specify causal loops in the form of time-forward relationships between processes linked by a system of differential equations. For this type of causal structure, it is well known that both heuristics described above can fail even when there are no unobserved common causes: 1) observations of two *causally independent* processes can be *conditionally dependent* (Aalen, Røysland, Gran, Kouyos, & Lange, 2016; Maxwell & Cole, 2007) and 2) observations of two *causally dependent* processes can be *conditionally independent* under certain conditions (Kuiper & Ryan, 2018).

This counter-example shows the critical necessity of specifying a target causal structure. First, it shows that we cannot simply assume that the heuristic rules described above hold for any and all types of causal structure. Second, it shows that for an intuitive time-forward interpretation of causal loops, these heuristics cannot be applied in an out-of-the-box fashion. This means that if a researcher wishes to simultaneously apply these hypothesis generation heuristics, while rejecting the notion of an underlying DAG, the burden of proof is on that researcher to provide an alternative formalism for the underlying structure, and to prove that the heuristics described above hold for that formalism. Finally, the counter-example highlights the profound advantage of using the DAG as a target causal structure, as we know for certain that these heuristics can be applied to learn about an underlying DAG structure. Given that the DAG is, to our knowledge, the only formalized causal structure for which we can currently be certain these heuristics apply, the best case scenario for using these heuristics to generate causal hypotheses is if the underlying causal structure is a DAG.[4]

In the following, we will examine the difficulties which remain in this hypothesis generation task, even when the underlying structure is a (faithful) DAG consisting only of observed variables (with these additional conditions for now left implied, and discussed at the end of the section).

2.3.2 Causal Hypotheses and DAGs: The Moral Graph

Given that we choose an underlying DAG as our target causal structure, how can the PMRF be used to help us generate hypotheses about the directed causal relationships in that (unknown) DAG? As discussed in the previous section, the equivalence of PMRFs and moral graphs means that the main information the PMRF can give us about an underlying DAG structure is *which edges are absent*. As stated in the second heuristic above, and illustrated in the examples in Figure 2.1 and Figure 2.2, the *absence* of an edge between two variables in the PMRF implies the *absence* of a directed edge between those two variables in the underlying DAG. Critically, however, the *presence* of an edge between two variables in

[4]It is possible, in theory, that these heuristics also apply equally well for some other type of causal structure, possibly including causal loops, of which the current authors are not aware, and which is not discussed by any of the researchers applying PMRF models in this way. However, we deem this unlikely, at least without specification of a wider array of additional assumptions. We address the issue of alternative causal models in the discussion.

the PMRF cannot be taken to imply the *presence* of a directed edge between those two variables in the underlying DAG: This latter statement is a property of the DAG *skeleton*, and not the moral graph.

It is pertinent to note here an overlap in terminology used in the DAG and network analysis literature. In the network analysis literature, multiple researchers refer to PMRF models explicitly as representing the *causal skeleton* of *some directed causal structure*, encoding the presence but not the direction of a causal relationship (van Borkulo et al., 2015; Borsboom & Cramer, 2013; Isvoranu et al., 2016; Haslbeck & Waldorp, 2018; Fried, Epskamp, Nesse, Tuerlinckx, & Borsboom, 2016; Armour, Fried, Deserno, Tsai, & Pietrzak, 2017). We reiterate that, if the underlying causal structure is a DAG, then in general the PMRF will not be equivalent to the skeleton of that DAG.

This point bears repeating because it lies at the heart of the difficulties which arise when using PMRFs to hypothesize about DAGs. The value of PMRFs for discovering causal structure is often justified by a comparison to networks where edges represent marginal dependencies (for instance, full rather than partial correlations). In contrast to these marginal dependency graphs, by conditioning on all variables, the PMRF omits what Costantini et al. (2015) and Heeren and McNally (2016) (amongst others) refer to as "spurious" edges: That is, connections between variables which do not reflect the presence of a direct connection in the causal data-generating structure (but which reflect either a common cause or indirect relationship). However, what is perhaps under-appreciated by users less familiar with the graphical modeling literature is that, when the underlying causal structure is a DAG, conditioning on all other variables in the network also has the effect of inducing additional "spurious" connections between variables which are unconnected in the DAG, resulting from conditioning on collider variables (as discussed by, amongst others, Epskamp, Waldorp, et al., 2018; Borsboom & Cramer, 2013).

2.3.3 Challenges in Using PMRFs to Generate DAG Hypotheses

This uncertainty regarding the causal status of edges in the PMRF means that there is, in general, a one-to-many mapping from a single PMRF to a set of DAGs: While any given DAG only has one corresponding PMRF, multiple different DAGs typically lead to the same PMRF.[5] This means that, while patterns of connected variables in a part of the PMRF (i.e. a *local structure*) may seem to imply a particular sequence of directed causal relationships, there are often many different directed structures which could have led to any such pattern. This represents a fundamental challenge for causal hypothesis generation: Although the heuristics described above may hold for any one edge, it is typically less clear from the PMRF alone whether two or more of these hypothetical directed relationships can be co-present in any one DAG.

For example, in the psychological network literature *chains* of dependent variables in PMRFs, of the form $X_i - X_j - X_k$ are often interpreted as indicative of a

[5] Note that the PMRF can also represent conditional dependency structures which cannot be represented by a DAG, such as $A - B - C - D - A$

directed mediation structure, $X_i \rightarrow X_j \rightarrow X_k$. Such interpretations are present in, for example, Deserno et al. (2017) (X_i = social contacts, X_j = social satisfaction, X_k = feeling happy), Isvoranu et al. (2016) (X_i = sexual abuse, X_j = anxiety and depression, X_k = psychosis) and Fried, Bockting, et al. (2015) (X_i = bereavement, X_j = lonely, X_k = sad and happy). While these are potentially correct causal hypotheses, there are typically many other valid possibilities that are consistent with the PMRF and an underlying DAG. These possibilities may include DAGs in which one or more of the connections of interest are absent, and are only induced in the PMRF due to a collider structure. Researchers may be at risk of drawing misleading conclusions about the underlying causal structure if these possibilities are disregarded, or dismissed without a justification for why one particular structure is more plausible than the next. Even more problematic is that combinations of different causal hypotheses based on local structures may be *incompatible* with one another, as they may imply for example, a new collider structure, or imply some (in)dependence relationships which contradict those in the PMRF. The plausibility of any given hypothesis regarding a part of the DAG, must be assessed by taking the *global* structure of this system into account, that is, considering the hypothetical DAG structure as a *whole*.

The challenging nature of using PMRFs can be illustrated with an example. Figure 2.3 shows a four-variable PMRF, and in addition, all of the different DAG structures which can generate that PMRF.[6] In this case we can see that 13 distinct DAG structures result in the given PMRF: We will call such a set of DAGs the *moral-equivalent* set. It is immediately clear that the the one-to-many mapping of a PMRF to different DAG structures precludes us from making any definitive statement regarding the directionality of any particular relationships. However, exploring the moral-equivalent set can still provide quite useful information. Inspecting the commonalities and differences in the moral-equivalent set gives us a more complete idea of the causal structures that the PMRF may reflect, and aids in generating and assessing causal hypotheses.

For example, numbers one through five of these DAGs contain one less edge than is present in the PMRF, indicating that there is some uncertainty regarding whether the variables A, B and C are all linked to each other by directed causal relationships. Furthermore, we can see that the edge $C - D$ is oriented in one direction more frequently than another, as $C \rightarrow D$ in 9 out of 13 DAGs and $C \leftarrow D$ in the remaining 4. If we assume that all 13 DAG structures are equally plausible, this gives some indication that $C \rightarrow D$ is more likely than $C \leftarrow D$. Furthermore, in those DAGs in which either A and/or B are parents of C, D cannot be a parent of C: This would result in a collider structure ($A \rightarrow C \leftarrow D$ and/or $B \rightarrow C \leftarrow D$) which is not allowed by the PMRF. This means we have some indication that C is not a common child of A, B and D, but must be a parent of at least one other variable.

By using Figure 2.3 we can re-evaluate the plausibility of different hypotheses that researchers may be tempted to make from local structures in the PMRF. For example, as A, B and C are connected to one another in the PMRF, one may be

[6] Assuming sufficiency and faithfulness.

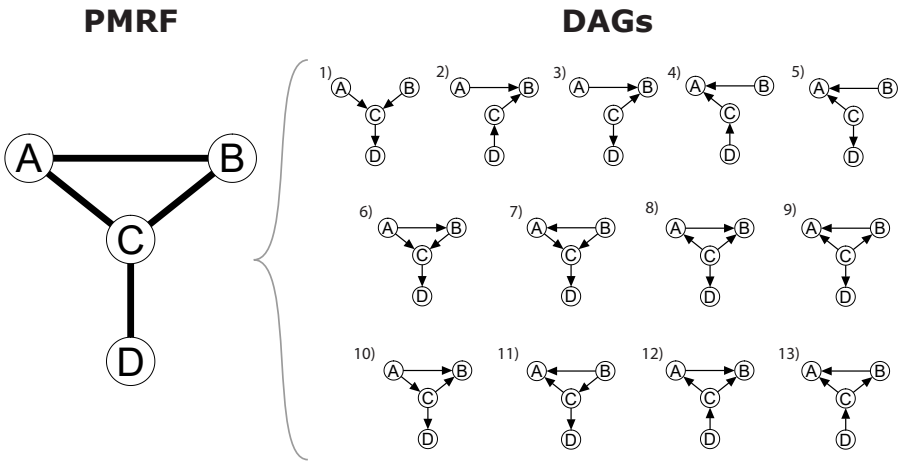


Figure 2.3: A Pairwise Markov Random Field (PMRF) and each DAG structure which generates that PMRF.

tempted to conclude that A has a direct effect on B ($A \rightarrow B$), as well as an indirect effect through C ($A \rightarrow C \rightarrow B$). Let us call this hypothesis I. To take another example, a researcher may be tempted to hypothesize that, as there is no connection between D and B, this is indicative that the effect of D on B is fully mediated by the variable C, that is $D \rightarrow C \rightarrow B$. Let us call this hypothesis II.

Inspecting Figure 2.3 we can see that hypothesis I holds in only one of thirteen DAG structures, DAG number 10. Furthermore, hypothesis II holds in exactly three DAG structures, numbers 2, 12 and 13. Strikingly, hypothesis I and II are totally incompatible with one another: although both hypotheses are reasonable explanations of two different local structures, and in fact hypothesize the same direction for one edge $C \rightarrow B$, it is impossible for both to be true in any underlying DAG structure.

This treatment highlights the difficulties which remain in generating causal hypotheses from the PMRF alone, even in the best case scenario of taking the DAG as the target causal structure. However, we have also seen that a lot of information which can be used in causal hypothesis generation is gained simply by generating different directed graphs which are consistent with a given PMRF. This entire procedure is made possible by the fact that we have specified the form of our target causal structure, and know the mapping from the PMRF to that causal structure. Only when this is done can we go about the task of causal hypothesis generation in a principled way, as illustrated here. It is pertinent to note here that in Figure 2.3 we depict a rather optimistic scenario for causal hypothesis generation - as the number of nodes and/or the number of connections in the graph grow, so too does the size of the moral-equivalent set. However, more informed causal hypotheses still can be generated by taking into account the size and sign of the relationships between variables, and by making some distributional assumptions: For instance, by relating GGMs to weighted DAGs in

the form of linear SEM models. We illustrate the generation of causal hypotheses from GGMs in the next section.

2.3.4 On Faithfulness, Sufficiency and Other Obstacles to Causal Hypotheses

Once a target structure is identified, in order to use the PMRF to generate hypotheses about that structure, at least three simplifying assumptions must be made. These additional assumptions are necessary to ensure that the data-generating DAG is contained in the moral-equivalence set, given above. But more generally, these assumptions are necessary for the two hypothesis generation heuristics outlined above to be valid, even in the best-case scenario that the underlying structure is a DAG. These assumptions bear further consideration here, as they pose additional barriers to the generation of causal hypotheses from PMRF models in practice.

2.3.4.1 Sufficiency

The first of these assumptions is known as *sufficiency*, which essentially entails that the underlying DAG consists of only the observed variables $\boldsymbol{X}$ (in the examples above, A, B, C and D) in the sense that there are no unobserved common causes of two or more observed variables (Lauritzen, 1996; Pearl, 2009; Spirtes et al., 2000).[7] This means that the conditional dependencies captured by the PMRF of $\boldsymbol{X}$ are assumed to come about due to directed causal relationships between the observed variables $\boldsymbol{X}$, and not due to relationships the observed variables share with unobserved variables (such as an unobserved common cause, or an unobserved collider variable on which we have unwittingly conditioned).

Relaxing the sufficiency assumption means that the task of generating causal hypotheses becomes much more difficult, as, depending on the number and type of missing variables we are willing to consider, there are many more DAGs which may have generated the given PMRF. For example, if we consider there to be a single unobserved variable, E, which acts as a common cause of A and B, then we must consider the DAG in panel (a) of Figure 2.4 as a plausible underlying causal structure: This DAG produces the same conditional dependencies present between the *observed variables* $A - D$ as the PMRF in Figure 2.3. If we consider only this exact type of unobserved variable, we must now consider 9 additional causal structures in which at least one of the connections in the PMRF is induced by this common cause (see Appendix 2.A).

If we relax the sufficiency assumption even further, for example allowing E to be a common cause of other variables, as in panel (b) of Figure 2.4 or allowing more than one unobserved common cause as in panel (c) of Figure 2.4, deriving a complete set of possible underlying DAGs quickly becomes infeasible. In general, without the sufficiency assumption, the causal information conveyed by the presence of an edge (as stated in *Heuristic 1*) becomes less certain, as there are

[7] More precisely we could say that there are no unblocked back-door paths, based on d-separation rules, passing through unobserved variables, that connect any pair of observed variables (Pearl, 2009).

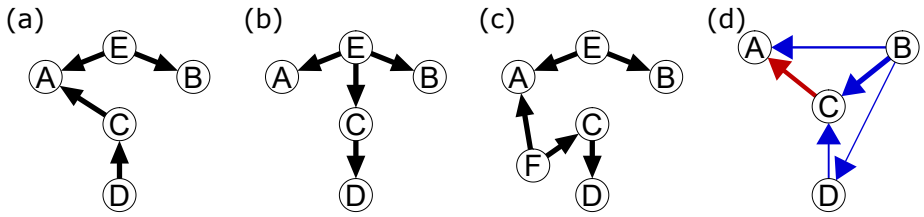


Figure 2.4: Four DAGs which generate the same PMRF between the variables $A - D$ (as shown in Figure 2.3), but which violate some assumption which is made in deriving the moral-equivalence set. The DAGs in panels (a), (b) and (c) represent sufficiency violations, with unobserved common cause(s) E (and F). The weighted DAG in panel (d) violates faithfulness, as the directed positive relationship between B and D is exactly canceled out by conditioning on C.

typically any number of possibilities involving unobserved variables which may explain a particular conditional dependency. In practice then, for any statistical analysis which hopes to capture causal relationships in some form, some assumptions regarding sufficiency and unobserved variables are necessary to make any hypothetical statements at all.

2.3.4.2 Faithfulness

The second major assumption needed to generate causal hypotheses on the basis of the PMRF is called *faithfulness* (Spirtes et al., 2000; Pearl, 2009). A DAG and associated probability distribution P meet the faithfulness condition if every conditional (in)dependence relation in P is entailed by the causal Markov condition in Equation 2.3 (Spirtes et al., 2000). For example, if two variables X_i and X_j are marginally independent, then by faithfulness the corresponding DAG should have no directed paths which can be traced from X_i to X_j, e.g. $X_i \rightarrow X_k \rightarrow X_j$. This means that we assume away the possibility that X_i and X_j are connected by two different directed pathways, which when combined in the marginal relationship between X_i and X_j, exactly cancel one another out. This would happen if, for example, there was a negative direct pathway $X_i \rightarrow X_j$ as well as a positive indirect pathway of equal size through X_k.

In panel (d) of Figure 2.4 we show a weighted DAG which would result in a violation of faithfulness. In this DAG, both B and D have a positive direct effect on C, making it a collider between them. Conditioning on this collider induces a *negative* conditional relationship between B and D. However, simultaneously, B has a *positive* direct effect on D: When combined, this negative and positive relationship cancel one another out, and so B and D appear to be conditionally independent given C - the partial correlation of B and D given C is zero, so they are unconnected in the PMRF (see Appendix 2.A for details). In general, without the faithfulness condition, the causal information conveyed by the absence of an edge (as in *Heuristic 2*) becomes less certain. In theory, the assumption of faithfulness is often justified in the context of relating the true probability distribution to a DAG; the probability of the true DAG having two pathways which exactly cancel out is said to be negligible (Spirtes et al., 2000).

2.3.4.3 Population vs Estimated Conditional Dependence

However, the faithfulness assumption bears further consideration in combination with the third simplifying assumption: Namely, that the given PMRF captures exactly the true conditional dependencies in the population, and not some estimate thereof. In other words, to enable a straightforward use of any estimated PMRF, we must assume that any edge represents that two variables are conditionally independent in the population. In practice of course, we typically don't have access to population statistics, and instead must try to account for uncertainty about these estimates in some way. In the context of sampled data Uhler, Raskutti, Bühlmann, and Yu (2013) have shown that violations of faithfulness have a non-negligible probability, as conditional dependence relationships are inferred on the basis of estimated parameters, rather than directly observed. In the context of the mediation example given above, it may be the case that while in the population the direct and indirect effects between X_i and X_k do not perfectly cancel out, the estimated marginal dependency (i.e. total effect) from a given sample may not meet some decision criteria to be considered non-zero (e.g., may not be significantly different from zero based on a t-test). This is a well-known phenomena in the SEM literature known as a *suppression* effect (Tzelgov & Henik, 1991).

It is pertinent to note again that, in practice, regularization techniques are typically used to estimate weighted PMRFs such as the GGM in psychology (Epskamp & Fried, 2018). These techniques introduce bias in parameter estimates, setting parameters representing weak conditional dependencies exactly to zero in the interest of parsimony. As such, our ability to make confident statements about the absence of an edge in such a regularized PMRF implying the absence of a direct causal relationship in the underlying DAG is on even less sure footing.

We wish to note here that, although these assumptions are framed with respect to the DAG, at least some similar assumptions are likely to hold no matter what the target causal structure we wish to make some inference about. One benefit of choosing the popular DAG as the target causal structure is that the types of assumptions which need to be met to infer DAG structures from conditional dependency information in different settings are well documented and well understood (Pearl, 2009; Spirtes et al., 2000; Dawid, 2010; Robins, 1999). By delineating these assumptions, it becomes possible to study their validity, enabling researchers to make better and more informed causal hypotheses, and facilitating the accumulation of knowledge. In contrast, taking a causally-agnostic approach precludes us in principle from understanding under what conditions and what assumptions causal hypotheses can be made.

2.4 Empirical Illustration

To illustrate the value of specifying a target causal structure in practice, we will make use of an empirical example taken from Hoorelbeke, Marchetti, De Schryver, and Koster (2016). This example will allow us to examine in more

detail the type of inferences which researchers typically aim to make based on an estimated PMRF, and how specifying a target structure influences these inferences.

In this illustration, we will make use of a novel tool which takes a precision matrix, as estimated in the GGM method, and produces *the set of statistically equivalent* weighted DAGs (hereby referred to as the *SE-set*) under ideal conditions. This can be considered the weighted equivalent of the moral-equivalent set, described in the previous section. In deriving the SE-set, we make use of the sufficiency assumption, outlined above, but we do not need to invoke the faithfulness assumption. Since we are deriving a weighted DAG equivalent to a linear path model, we also assume linear relationships between variables (as captured by the GGM), and uncorrelated error terms. Full details on the *SE-set* algorithm can be found in Appendix 2.B and it can be downloaded as an R-package from the github page of the first author.[8]

Hoorelbeke et al. (2016) analyzed the network structure of cognitive risk and resilience factors in a cross-sectional sample of 69 remitted depression patients. This network consisted of six variables: self-reported cognitive control (BRIEF_WM); performance on a behavioral cognitive control measures (PASAT_ACC); adaptive emotion regulation strategies (Adapt ER), maladaptive emotion regulation strategies (Maladapt ER); resilience levels (Resilience); and residual depression symptoms (Resid Depres).

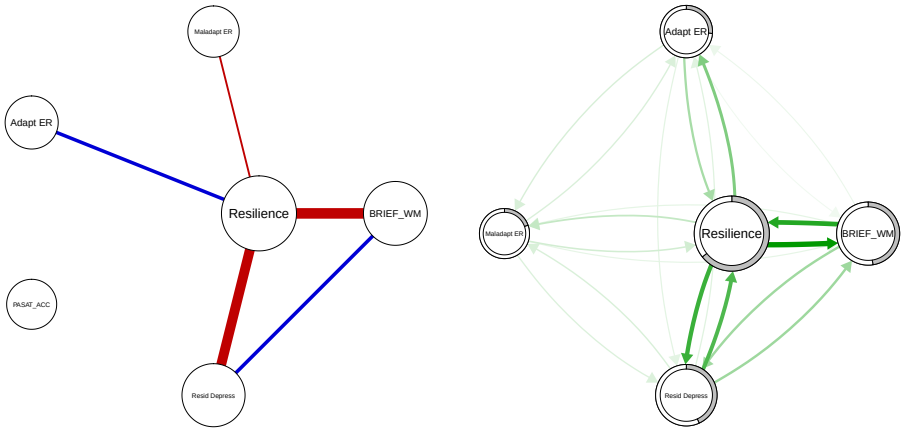
In the original article, the authors estimated a GGM, and additionally a directed *relative importance* network. The relative importance network is another way of encoding how well one variable is predicted by another, based on explained variance, and is often used in conjunction with an estimated PMRF model in order to further inform the generation of causal hypotheses (Robinaugh, LeBlanc, Vuletich, & McNally, 2014; McNally et al., 2015; Heeren & McNally, 2016). We begin by first reproducing their analysis and examining the conclusions they make on the basis of this analysis. Full details of the re-analysis is given in Appendix 2.C. Following this, we use the SE-set algorithm to re-examine and expand upon the causal hypotheses generated from the original analysis.[9]

2.4.1 GGM Analysis

The precision matrix was estimated with the *qgraph* package (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012), using the marginal correlations reported by Hoorelbeke et al. (2016). The corresponding GGM is shown in Figure 2.5(a). From this GGM, we see that Resilience has strong negative connections to both working memory and residual depressive symptoms; a weak negative connection to maladaptive emotion regulation; and a weak positive connection to adaptive emotion regulation. Furthermore, we see that the cognitive control

[8] https://github.com/ryanoisin/SEset

[9] The code to reproduce this analysis is available at https://github.com/ryanoisin/CausalHypotheses



(a) GGM using EBICglasso

(b) Relative Importance Network

Figure 2.5: Networks of cognitive risk and resilience factors based on re-analysis of Hoorelbeke et al. (2016)

measure is unconnected to the rest of the variables.[10]

2.4.1.1 Relative Importance and Predictability

Relative importance is a measure of unique predictive strength, based on explained variance. The relative importance network is a directed, weighted network, in which each edge $X_i \rightarrow X_j$ represents the *unique variance explained* in X_j by X_i in a multiple regression where X_j is the outcome variable and $\boldsymbol{X}^{-j}$ are the predictors. For each node in the relative importance network, the sum of incoming pathways (i.e., in-strength centrality) is related to the *predictability* of that node. Predictability is calculated as the *total variance explained* (R^2) in a given outcome variable from the above regression. This is measure which, in network analysis, is often interpreted with respect to an underlying directed causal structure: The predictability of a node X_i is typically interpreted as an *upper-bound* on *controllability*, that is, the variance explained in X_i by its causes (parents) in the true underlying causal structure (Haslbeck & Fried, 2017; Haslbeck & Waldorp, 2018; Fonseca-Pedrero et al., 2018). Predictability is typically depicted as a ring around each node, with the proportion of the ring filled denoting the R^2.

Relative importance was calculated using the *lmg* metric in the *R* package *relaimpo* (Grömping, 2006). As the PASAT_ACC item is unrelated to the other variables in the GGM, it is omitted from this analysis. The resulting relative importance network is shown in Figure 2.5(b). Hoorelbeke et al. note that Resilience

[10] Due to the different methods used to create a GGM, there are slight differences in the GGM network depicted in Hoorelbeke et al. (2016) and the current paper. Apart from small numerical differences in parameters, the main difference is the presence of a weak positive partial correlation between working memory and residual depressive symptoms. Further details can be found in Appendix 2.C

shows the highest level of betweenness, closeness and strength centrality, and is the only node with higher out-strength values (0.77) than in-strength values (0.64). This means that Resilience appears to have a greater value in predicting other variables than vice versa. However, the predictability of each node in Figure 2.5(b) shows that Resilience is the most predictable node in absolute terms, followed by working memory and residual depressive symptoms. In comparison, the variables relating to emotion regulation strategies have relatively lower predictability.

2.4.1.2 Interpretation of GGM Analysis

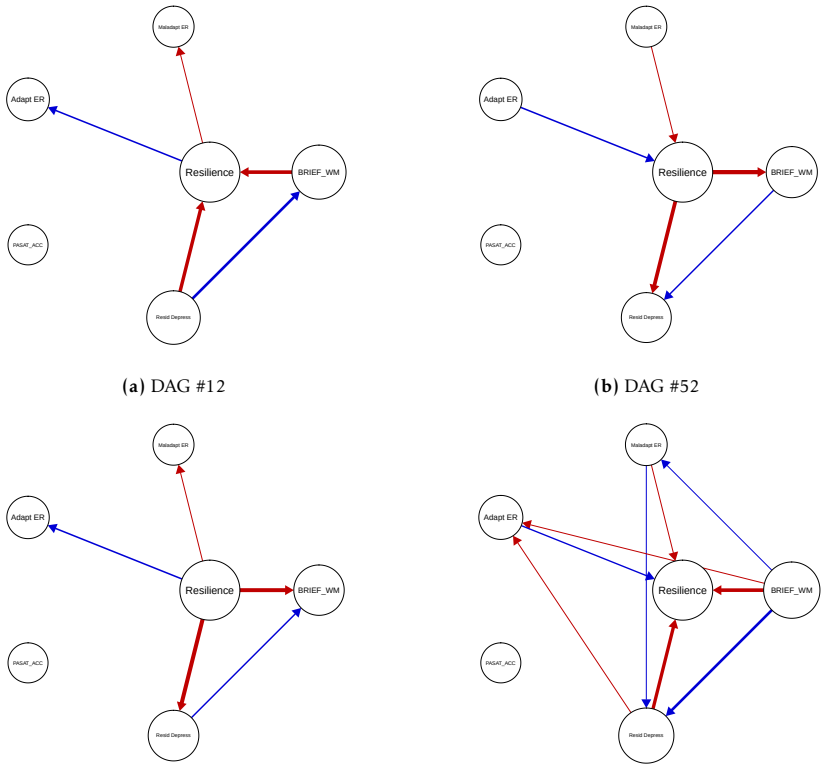
In the original paper, Hoorelbeke et al. (2016) interpret the GGM as showing that Resilience is the "main hub" of the network, connecting all other variables. Based on the relative importance network, it is tempting to draw further conclusions about the underlying causal structure. For instance, it is tempting to conclude that the undirected edges in the GGM reflect the presence of causal effects from Resilience to other variables, or that a causal structure where Resilience is the common *cause* of the other factors is more likely than a structure where Resilience is a common effect (child) of all other variables.

However, it is well known that prediction is not necessarily indicative of causation; neither is it clear why an edge in the relative importance network is evidence that a causal relationship is oriented in a particular direction. Although the original authors do note early in the paper that the weights of the directed edges here represent predictive relationships and do not imply causality, they later interpret the greater out-strength of Resilience as implying that Resilience "exerted a larger influence on the rest of the network than vice versa" (Hoorelbeke et al., 2016, p. 100). Although these interpretations feel intuitive, we will show in the following that inferring possible underlying structures in this way can be misleading.

2.4.2 Analysis of Causal Hypotheses: The SE-set

As the researchers are explicitly interested in the directionality of causal effects, this is a typical scenario in which the specification of some target causal structure can aid greatly in the exploration and generation of causal hypotheses. Here, we will take the target causal structure to be a weighted DAG. That is, we assume the data-generating structure takes the form of a linear path model. This allows us to explicitly account for the sign and strength of relationships estimated in the GGM, rather than only the presence and absence of conditional dependencies, as we did when mapping a PMRF to an unweighted DAG in the previous section.

In order to explore different possible underlying causal structures, we will make use of the SE-set algorithm. This tool allows us to generate a set of weighted DAGs which reproduce the given GGM. In doing so, we will assume sufficiency, in that we only allow DAG structures between observed variables, but we will not impose the assumption of faithfulness. The SE-set algorithm was run on the estimated precision matrix consisting of 6 variables, which would produce a



(a) DAG #12 **(b)** DAG #52

(c) DAG #14 **(d)** DAG #20

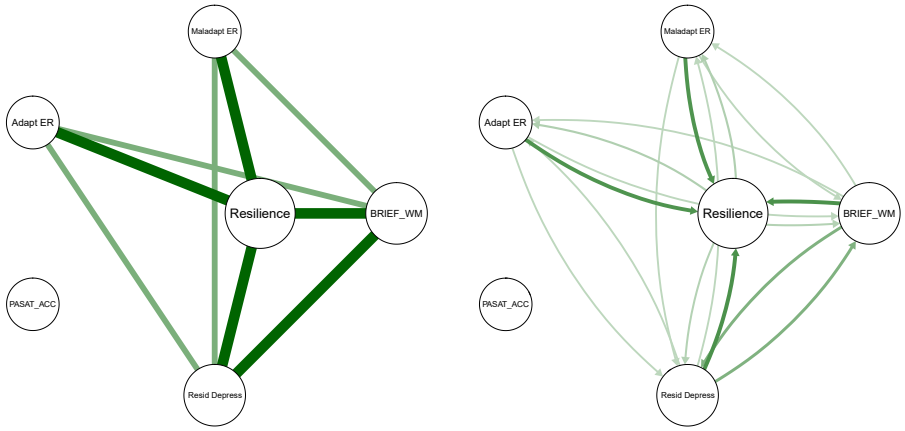
Figure 2.6: Networks representing frequency of occurrence of edges in the DAG equivalence set.

maximum of 6! = 720 different weighted DAGs, one for every possible ordering of the six variables from cause to effect. After rounding and the removal of duplicate weights matrices, we are left with only $n_{DAG} = 58$ unique weighted linear DAG structures. We can explore the SE-set by inspecting some examples of DAGs in the set and summarizing different characteristics of members of the set.

2.4.2.1 Examples of Equivalent DAGs

Four examples of weighted DAGs in the SE-set are shown in Figure 2.6. We can see that all four DAGs represent quite distinct causal structures. In Figure 2.6(a) residual depression has a negative causal effect on Resilience and a positive effect on working memory. Resilience also fully mediates the relationship between these two variables and the two emotion regulation variables. In Figure 2.6(b) the sign (positive/negative) of these relationships remains the same, but the direction of each relationship is exactly the opposite. Furthermore the size of the relationship between working memory and residual depression has decreased.

Figure 2.6(c) represents a DAG in which Resilience is the common cause of all



(a) Any edge relative frequency

(b) Directed edge relative frequency

Figure 2.7: Networks representing frequency of occurrence of edges in the SE-set.

of the other variables in the network, as may have been hypothesized on the basis of the relative importance network. Finally, Figure 2.6(d) represents an example where Resilience is the common effect, i.e. caused by all other variables in the network. Note that in this example, there are a number of edges present which are not present in the estimated GGM, even though this weighted DAG does reproduce the input precision matrix (for a discussion on how this may arise, see the discussion of the faithfulness assumption in the previous section and Appendix 2.A).

By examining different elements of the SE-set we can see that there is a variety of distinct directed structures that may have resulted in the estimated GGM. Researchers can use substantive expertise to rule out some elements of the SE-set: For instance, based on previous research we may know that depressive symptoms are a cause of working memory problems, and not vice versa, which would rule out DAGs 52 and 20 (Figure 2.6(b) and 2.6(d)) respectively as plausible data-generating structures. In the absence of such substantive knowledge, in the current paper we treat each of the 58 DAGs in the SE-set as equally plausible data-generating structures.

2.4.2.2 Relative Frequency of Edges

Rather than examining each weighted DAG individually, we can summarize the information in the SE-set. One way to do this is by calculating in what proportion of DAGs in the SE-set an edge of any direction (i.e. $A \rightarrow B$ or $B \rightarrow A$) connects two nodes. This information is shown as an undirected weighted network in Figure 2.7(a). In this network, the presence of an edge indicates that in at least one member of the equivalence set, there is a directed arrow connecting the two nodes. The thickness of the arrow indicates how frequently such an edge

is present. In this particular instance, the thick arrows denote edges which are always present in the SE-set, with the less thick edges indicating that these edges were present in only 52 percent of the SE-set.

From Figure 2.7(a) we see that every edge which is present in the estimated GGM is present in every member of the SE-set. We can further see that none of the weighted DAGs contain edges between Adapt ER and Maladapt ER, or between PASAT_ACC and any other variable. From this we can say that, for this particular example, an edge in the GGM appears to correspond to the presence of an edge in an underlying weighted DAG; none of the edges in the GGM appear to be induced due to conditioning on a collider *in this case*. Conversely, the absence of an edge in the GGM does not always correspond to the absence of an edge in the underlying weighted DAG. In half the members of the SE-set there is at least one directed relationship present which is absent in the GGM: These are weighted DAGs which are unfaithful, in the sense that two or more conditional dependencies approximately cancel one another out (as discussed in the previous section).

We can further explore the equivalence set by seeing in what proportion of the SE-set edges *of a particular direction* occur. This is represented as a weighted, directed network in Figure 2.7(b), where the edges $A \rightarrow B$ describe how often a corresponding directed relationship $A \rightarrow B$ is present in the SE-set.[11] We can see from Figure 2.7b that for most pairs of variables, edges of opposite directions occur with equal frequency. However, this is not the case for edges relating to the Resilience variable, which was more often the child (or effect) than the parent (cause). Resilience was the child of the working memory in 71 percent of the equivalence-set DAGs, and a parent in 29 percent of the cases. A similar split is present in the connections between Resilience and Resid Depress (71/29), Resilience and Adapt ER (69/31) and Resilience and Maladapt ER (69/31).

It is interesting to note the different ideas we get about a possible underlying directed causal structure from Figure 2.7(b) than the relative importance network in Figure 2.5(b). From the latter researchers may be tempted to conclude, that, because Resilience has a greater predictive value for other variables than vice versa, that Resilience is more likely to cause those variables than vice versa. From the SE-set analysis we can say that, if we think the data-generating mechanism is a linear weighted DAG, and each member of the SE-set is equally plausible, then it is more than twice as likely that Resilience is caused by the other variables in our graph than vice versa.

2.4.2.3 Controllability Distribution in the SE-set

In general, the SE-set algorithm can be used to examine how any particular characteristic of a weighted DAG varies over different members of the SE-set. Node predictability, described above, gives us an *upper-bound* on the variance explained in a node X_i by its parents in the true underlying DAG, that is, the controllabil-

[11] Pairs of directed edges in this network are mutually exclusive: if $A \rightarrow B$ is present then $A \leftarrow B$ cannot be present. The weights of mutually exclusive pairs add up to the weight of the corresponding undirected edge in Figure 2.7(a)

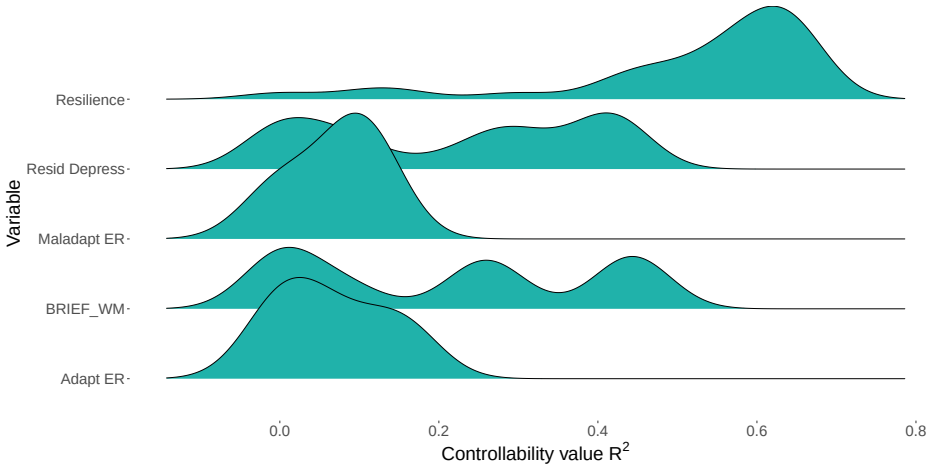


Figure 2.8: Distribution of controllability values for each variable, over DAGs in the SE-set. Controllability is here defined as the explained variance R^2 of each variable when predicted by its parents in a given DAG.

ity. However, the true controllability will differ depending on which variables are causes and effects of others. For example, Resilience will have the highest controllability in Figure 2.6(d), where it is a common effect of all other variables, and a controllability of zero in Figure 2.6(c), where it is a common cause of all others.

Using the SE-set algorithm, we can try to quantify our uncertainty about controllability, by calculating a distribution of this value over the different possible weighted DAG structures. This is shown in Figure 2.8. From this figure we can see that, in most elements of the SE-set, the controllability of Resilience is high, as we would expect given that Resilience is more often an effect than a cause of other variables. We can also see that the controllability of both maladaptive and adaptive emotion regulation strategies is quite low across different DAGs, with parents of both strategies explaining a maximum of 20 percent of the variance in each. However there is much more uncertainty about the controllability of residual depressive symptoms and working memory. The distribution of controllability for each is quite wide, ranging from zero to sixty percent, and with peaks of similar height all across this range. This analysis shows us that, while predictability in a standard GGM analysis captures the upper bound of these controllability distributions, the SE-set analysis can supplement this information by showing the variance and peaks of these distributions.

2.5 Discussion

In this paper we have critically evaluated a widespread practice in the current literature on psychological network analysis: Using estimates of PMRF-based models to generate hypotheses about an underlying directed causal structure. We

have shown that the utility of PMRF models for causal hypothesis generation is critically dependent on what target causal structure is specified. When taking a causally-agnostic approach, hypothesis generation is a fundamentally intractable problem: The heuristics used by researchers to generate these hypotheses cannot be assumed to apply to any and all causal structures. Defining a target structure, however, allows us to generate causal hypotheses in a principled way. We illustrate this procedure using the DAG and a linear path model (weighted DAG) as those target structures. In each case, we have shown that the specification of a target structure allows for detailed and novel causal hypotheses to be generated, based on the global structure of the estimated PMRF. In contrast, we have shown how the isolated application of heuristics, even in combination with other tools which encode predictive relationships, can lead to at best incomplete and at worst incorrect inferences regarding the underlying pattern of causal relationships.

Throughout the paper we have made use of DAGs and weighted DAGs, in the form of linear path models, as candidate target causal structures. D-separation rules allowed us to derive the moral-equivalent set of DAGs from a PMRF, and the SE-set algorithm introduced here allowed us to derive the statistically-equivalent set of weighted DAGs from a GGM. It should be noted that the SE-set is best considered an illustrative or exploratory tool. The benefit of using the SE-set algorithm in the current paper is that it directly relates GGMs to potential underlying causal structures. As a stand-alone DAG estimation method, however, the SE-set algorithm is limited for that same reason - inferences are entirely reliant on the GGM estimate, and uncertainty in those estimates is not accounted for. Researchers who wish primarily to estimate a (weighted) DAG structure from data would be better served in using algorithms specifically designed for that purpose, such as the PC, LiNGAM and FCI algorithms, amongst others (Spirtes et al., 2000; Shimizu, Hoyer, Hyvärinen, & Kerminen, 2006; Spirtes, Meek, & Richardson, 1995), many of which are implemented in *R* packages such as *pcalg* (M. Kalisch, Mächler, Colombo, Maathuis, & Bühlmann, 2012) and *bnlearn* (Scutari, 2010). Note that typically these algorithms are unable to uniquely identify a DAG from data, instead estimating a *Markov-equivalent* class, that is, a set of DAGs with fully equivalent conditional dependency relationships (cf. Andersson et al., 1997).[12] Numerous tutorials on the use and advantages of different DAG estimation algorithms, and the assumptions necessary for each, are available for the interested reader (e.g., Malinsky & Danks, 2018; Spirtes & Zhang, 2016).

However, the possibility of directly estimating DAG structures from data should not be taken as a panacea to the problem of inferring multivariate patterns of causal relationships from data. The focus on DAGs in the current paper is motivated by our focus on the heuristic mapping rules which are applied by researchers to generate causal hypotheses: If the underlying structure is a DAG, this represents the best case scenario for this practice, in the sense that we are guaranteed those heuristics are correct. A focus on DAGs as target causal structures also confers some additional benefits, as they have been extensively studied

[12] Furthermore, since these algorithms take disparate approaches to estimating DAG structures, we are not guaranteed that the moral graph implied by the estimated Markov-equivalent DAGs will be equal to the PMRF estimate, even when the underlying structure is a DAG.

in the causal inference literature, and can be easily used in conjunction with, for example, interventionist theories of causality (Spirtes et al., 2000; Pearl, 2009; Dawid, 2002; Richardson & Robins, 2013). That being said, it is possible that the DAG is not the optimal formalism for causal structure in psychological settings, and that some alternative causal structure may be more appropriate (for various discussions on this point, see McNally, 2016; Dawid, 2010; Cartwright, 1999, 2007). The problem however, is that it appears as though no such alternative formalism is readily available, or at least well-known in the psychology literature.

Cyclic causal models (i.e., graphs which allow causal "loops" $X_i \leftrightarrows X_j$) are in general underdeveloped, and the interpretation of cyclic causal effects without invoking a notion of time-forward dependency is notoriously difficult (e.g., Spirtes, 1995; Hayduk, 2009). As noted by, amongst many others, Borsboom et al. (2012) and Epskamp, Waldorp, et al. (2018), cyclic effects between dynamic processes can easily be represented as acyclic time-forward relationships $X_{i,t} \rightarrow X_{j,t+1} \rightarrow X_{i,t+1}$, either in a DAG or other related structure, as in time series analysis (e.g., Bringmann et al., 2013; Hamaker & Dolan, 2009). Formalized approaches to defining these time-forward causal relationships exist both in discrete-time (Eichler & Didelez, 2010; Dahlhaus & Eichler, 2003) and continuous-time (Didelez, 2000; Aalen, 1987). Alternatively, undirected graphs such as the Ising model can potentially be given a dynamic interpretation, reflecting symmetric time-forward directed relationships, by invoking Glauber dynamics (Glauber, 1963; Haslbeck, Epskamp, Marsman, & Waldorp, 2018; Marsman et al., 2018), as applied in recent theoretical models in psychology (Cramer et al., 2016; Dalege et al., 2016; Borsboom, 2017). However, if we take these types of dynamic systems to be the target structure(s), many additional assumptions must be invoked in order to ensure that PMRF edges estimated from a given data source reflect causal relationships therein. For example, assumptions about the frequency of the underlying process are needed to recover the structure of a dynamic system from time-series data (Papoulis & Pillai, 2002; Marks, 2012) and the oft-criticized *ergodicity* assumption is needed to ensure recovery of dynamic relationships from cross-sectional data (Molenaar, 2004; Hamaker, 2012; Molenaar, 2008). There may be yet other approaches to cyclic causal effects which are more promising for psychological applications: For example, cyclic effects which represent dynamic systems in an equilibrium state (Mooij, Janzing, Heskes, & Schölkopf, 2011; Forré & Mooij, 2018, 2019). However, more work is needed to assess their suitability and applicability in practice, and to establish any potential links between those types of causal structures and the conditional dependency patterns estimated by methods such as the PMRF.

Outside of their use for causal hypothesis generation, there are many attractive reasons to use PMRF-based models in practice. Amongst other things, they allow for the identification of predictive relationships, sparse descriptions of statistical dependency relationships in a multivariate density, and may be used as a variable clustering or latent variable identification method (e.g., Golino & Epskamp, 2017). However, we have shown that, when the aim is to uncover some aspects of the underlying causal structure, it is generally unclear what conclu-

sions we can draw directly from an estimated network structure alone (that is, without further specification and exploration of a target structure). Ours is not the first paper to come to such a conclusion: Dablander and Hinne (2019) showed that node centrality in a GGM is a poor indicator of causal influence in an underlying DAG; Bos et al. (2017) showed that conclusions about causal structure based on a cross-sectional GGM do not generalize to those made if a time-forward causal structure is assumed; Haslbeck and Ryan (2019) showed that it is typically unclear how to use various statistical methods common in the network approach literature, including the GGM, to draw conclusions about an underlying dynamic system from time series data. This pattern of results is concerning, as our reading of the literature suggests that discovering patterns of causal relationships is the primary motivation behind most of the applications of these methods in empirical research: These statistical methods appear to hold the promise of directly uncovering the causal interactions which are a cornerstone of the theoretical framework which motivated their development (Borsboom & Cramer, 2013; Borsboom, 2017).

In this paper we suggest one route by which the search for causal relationships in psychological networks can move forward: The investigation and specification of target causal structures. Specifying a target causal structure establishes the rules and boundaries through which we can make some inductive or abductive inference from an estimated network model to the underlying system of causal relationships. The feasibility of this approach relies on first the availability of suitable formalized causal structures and second the establishment of how information captured by different statistical methods relates to those structures. An alternative to this route would be to use methods such as the PMRF as purely deductive or confirmatory tools, to test the implications and predictions of prespecified causal theories: This in turn relies on our ability to specify these theories a-priori. In either case, the critical observation is that to use PMRFs, or in essence, any statistical method, to discover potential causal relationships, it is necessary to move beyond the causally-agnostic approach. Only when we make clear what it is we wish to make some inference about, can we hope to use any tool to make those inferences in a principled way. Only then can the power of this and other statistical approaches be fully realized: By understanding what types of inferences we can make, when we can make them, and how to go about doing this in practice.

Appendix 2.A Moral-Equivalent DAGs: Violations of Sufficiency and Faithfulness

In this appendix we provide additional insight into moral-equivalent DAGs which violate the assumptions of either sufficiency or faithfulness.

2.A.1 Sufficiency-Violating DAGs

In the main text, Figure 2.4(a), we show a single example of a DAG which generates the PMRF depicted in the left-hand column of Figure 2.3, in the presence of an unobserved variable E which acts as a common cause of both A and B. If this is the data-generating DAG, we can say that variables $A - D$ do not meet the assumption of sufficiency - to derive the true DAG, the variable E is needed.

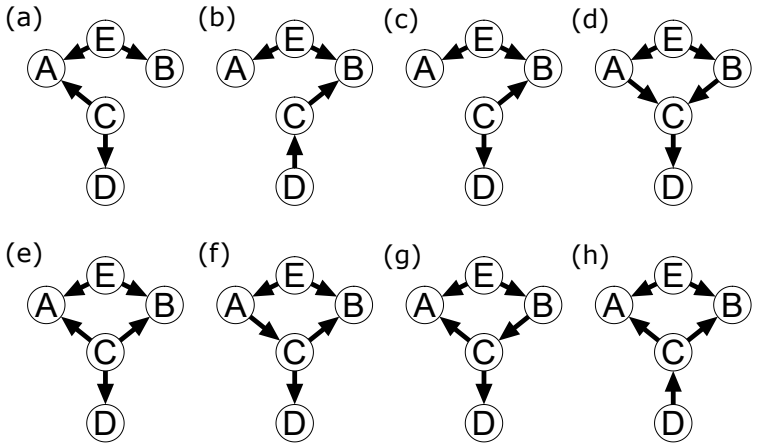


Figure 2.9: Eight additional DAGs which generate the same PMRF between the variables $A - D$ (as shown in Figure 2.3). For each DAG, the set of variables $A - D$ violates the sufficiency assumption, with respect to the variable E, a common cause of both A and B.

If we assume that sufficiency is violated only in that an unobserved variable E is a common cause of A and B, then there are eight more DAGs which generate the same PMRF for the variables $A - D$, and contain one less directed edge between the variables $A - D$ than any of the DAGs in the moral-equivalent set shown in the right-hand column of Figure 2.3. These are depicted in Figure 2.9. Many more moral-equivalent DAGs are possible if the assumption of sufficiency is relaxed further, that is, if different and/or additional connections between E and other variables or allowed, or if more than one unobserved common cause is allowed.

2.A.2 Faithfulness-Violating Weighted DAGs

In the main text, Figure 2.4(d), we depict a weighted DAG which generates the PMRF depicted in the left-hand column of Figure 2.3, but which violates the

faithfulness assumption. The weights matrix of this weighted DAG is given as

$$\boldsymbol{B} = \begin{bmatrix} 0.000 & 0.200 & -0.400 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.485 & 0.000 & 0.152 \\ 0.000 & 0.100 & 0.000 & 0.000 \end{bmatrix} \tag{2.5}$$

Although there is a positive direct effect from B to D, denoted by the parameter $b_{41} = 0.100$, the correlation between B and D when C is conditioned on is zero. This is because conditioning on C induces a slight negative relationship between B and D, which cancels out the positive directed effect from B to D.

It is well known that if two variables X and Y are independent but share a collider Z on which both have a positive effect, there will be a negative conditional dependency between X and Y conditional on Z (cf. VanderWeele & Robins, 2007). Suppose X and Y are standard normal variables, and fully determine the variable Z, both having the same positive effect on Z. Now consider that we condition on Z taking on its average value, $Z = 0$. If we also know that X takes on a moderately low value (say, $X = -0.5$) then it must be the case that Y takes on a high value ($Y = 0.5$) – otherwise, it would not be possible for Z to take on a value of zero. In this way, although both X and Y are causally independent, if we know the value of their common effect Z, then X and Y contain information which can be used to predict one from the other.

Appendix 2.B The SE-set Algorithm: A Tool to Aid Causal Hypothesis Generation

In this appendix we describe a tool which allows us to directly relate a given GGM to a set of equivalent weighted DAGs (i.e., linear path models), based on SEM principles. This tool takes as input an (estimated) precision matrix and outputs a set of weighted DAG structures between the observed variables. While each of these weighted DAGs may lead to very different substantive interpretations, they all lead to the same pattern of partial correlations between the observed variables described by the GGM.

The SE-set algorithm can be used to explore the different weighted DAG structures which may underlie a given GGM, aiding in the generation of causal hypotheses. As the weighted DAG structures in question are based on linear path models, we can say that they are statistically-equivalent in the sense of a structural equation model, in that they produce the same model-implied variance-covariance matrix (Bollen, 1989; MacCallum et al., 1993; Raykov & Marcoulides, 2001; Tomarken & Waller, 2003). This tool is thus called the *statistical-equivalence-set* or *SE-set* algorithm.

Note that we are limited in the current implementation to deriving DAGs which meet the sufficiency assumption (i.e., weighted DAGs between observed variables). However it is not necessary to invoke the faithfulness assumption, and so the SE-set algorithm may produce DAGs which are unfaithful to the set

of conditional (in)dependencies described by the GGM. Furthermore, note that the SE-set algorithm is intended primarily as an illustrative and exploratory tool: The limitations of this algorithm as a stand-alone DAG estimation procedure are addressed in the discussion section of this paper.

2.B.1 Relating the GGM and Weighted DAGs

The basis of the SE-set algorithm is that both the weights matrix of the $p-$variate GGM and the weights matrix of a weighted DAG can be related to each other through the $p \times p$ *inverse variance-covariance* matrix, known as the *precision matrix* $\hat{\boldsymbol{\Omega}}$ (see Equation (2.2)). In most applications of the GGM, for example those using regularized estimation methods, the estimated precision matrix will not be equivalent to the inverse of the observed covariance matrix. Instead, we can describe the inverse of the estimated precision matrix as a *model-implied covariance matrix*, which we will denote $\tilde{\boldsymbol{\Sigma}}$

$$\tilde{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Omega}}^{-1}. \tag{2.6}$$

The weights matrix of a weighted DAG was defined in Equation (2.4) as the matrix of regression parameters $\boldsymbol{B}$ in which child variables are regressed on their parents. In other words, the weighted DAG can be seen as a SEM model with uncorrelated residual terms. From the SEM literature, there is a straightforward expression relating the parameters of such a path model to a model-implied variance covariance matrix. This relationship is given by

$$\tilde{\boldsymbol{\Sigma}} = (\boldsymbol{I} - \boldsymbol{B})^{-1} \boldsymbol{\Psi} (\boldsymbol{I} - \boldsymbol{B})^{-T} \tag{2.7}$$

where $\boldsymbol{\Psi}$ is a matrix of residual variances of $\boldsymbol{X}$, diagonal in the case of uncorrelated residuals (Bollen, 1989). As such, the weights matrix of both the GGM and the weighted DAG can both be seen as (estimated) decompositions of some variance-covariance matrix $\tilde{\boldsymbol{\Sigma}}$. A similar expression is used by Epskamp, Waldorp, et al. (2018) to illustrate the relationship between the precision matrix and an underlying weighted DAG.

Given that we now have expressions relating the weights matrices of a GGM to the weights matrix of a DAG, through the model-implied covariance matrix (Equations (2.6) and (2.7)), we can use this relationship to define the *SE-set* algorithm. If $\boldsymbol{\Psi}$ and $\boldsymbol{B}$ are both known, then they can be combined to find $\tilde{\boldsymbol{\Sigma}}$ using Equation (2.7). However, the inverse operation, solving uniquely for $\boldsymbol{B}$ and $\boldsymbol{\Psi}$ from $\tilde{\boldsymbol{\Sigma}}$ is not typically possible without additional information or assumptions. This is a well-known issue in the SEM literature, which has long identified that many different path models imply the same variance-covariance matrix (MacCallum et al., 1993; Raykov & Marcoulides, 2001).

One circumstance in which it *is* possible to find $\boldsymbol{B}$ directly from the covariance matrix is when the *topological ordering* of the DAG is known, and the residual terms are assumed to be uncorrelated (Levina et al., 2008; Shojaie & Michailidis, 2010). The topological ordering of a DAG is defined as an ordering of nodes such that every parent comes before every child. The graph in Figure 2.1(c) has two valid topological orderings: {Support, Pressure, Stress, Worry} and {Pressure,

Support, Stress, Worry}. If the rows and columns of the covariance matrix $\tilde{\Sigma}$ are sorted according to the topological ordering, then Equation (2.7) gives a unique solution. In that case, $\boldsymbol{B}$ will be a lower triangular matrix with zero's on the diagonal, and Equation (2.7) will be equivalent to an LDL^T matrix decomposition (Abadir & Magnus, 2005).

2.B.2 From GGM to Statistical Equivalent Set

In the settings of interest in the current paper, the topological ordering of the underlying DAG is unknown. In a system of p variables, there are $p!$ possible topological orderings. Thus, every one of the $p!$ possible orderings of the rows and columns of $\tilde{\Sigma}$ produces a (possibly distinct) weights matrix $\boldsymbol{B}$. Typically the number of distinct $\boldsymbol{B}$ matrices will be less than $p!$ as some DAG structures will have more than one equivalent topological ordering. For example, in Figure 2.1(b), the GGM is made up of four variables, Support, Pressure, Stress and Worry. This means there are $4! = 24$ possible topological orderings of an underlying weighted DAG. However at least two of these topological orderings describe the same weighted DAG structure, as described above.

The SE-set algorithm takes as input a $p \times p$ precision matrix and calculates a corresponding $\boldsymbol{B}$ for every $p!$ possible topological ordering of the observed variables. It does this by repeatedly re-ordering the variables in $\tilde{\Sigma}$ and applying the transformation in Equation (2.7). This gives a set of weighted DAG models of size $p!$, the weights matrices of which are collected in the SE-set $\boldsymbol{\mathcal{B}}$. Each weighted DAG in $\boldsymbol{\mathcal{B}}$ leads to the same model-implied variance covariance matrix $\tilde{\Sigma}$. This means that, by construction, each weighted DAG in $\boldsymbol{\mathcal{B}}$ is statistically equivalent.

This definition of statistical equivalence comes from the structural equation modeling literature, where statistically equivalent path models are those which yield the same fit, defined by the distance between the model-implied variance covariance matrix $\tilde{\Sigma}$ and the population covariance matrix $\boldsymbol{S}$, and the restrictiveness of the model, typically defined by the number of freely estimated parameters (Bollen, 1989). Each element of $\boldsymbol{\mathcal{B}}$ is statistically equivalent to each other in this sense. By construction, each matrix $\mathbf{B}$ defines a path model with $p(p-1)/2$ paths and p residual variance terms, which all yield the same model-implied covariance matrix $\tilde{\Sigma}$. That a single variance-covariance matrix can typically be generated by a large number of distinct statistically equivalent path models is a well known issue in the SEM literature (MacCallum et al., 1993; Raykov & Marcoulides, 2001; Tomarken & Waller, 2003).

After deriving all $p!$ possible weighted DAGs, the size of the SE-set $\boldsymbol{\mathcal{B}}$ can be reduced in a second step by rounding or thresholding the values in the weights matrices, and removing duplicates. In the next subsection we discuss in greater detail the mathematical details of the algorithm.

2.B.3 Mathematical Basis of the SE-set Algorithm

Take it that we have some $p-$variate set of variables V, related by a $p \times p$ precision matrix $\boldsymbol{\Omega}$. Each path model in the SE-set of $\boldsymbol{\Omega}$ is defined by a unique ordering of

the variables V, of which there are $p!$. The output of the algorithm is thus a set of matrices

$$\mathcal{B} = \{\boldsymbol{B}_1, \boldsymbol{B}_2, \dots \boldsymbol{B}_{p!}\} \tag{2.8}$$

of size $p!$. To generate $\mathcal{B}$ we first generate $\mathcal{V}$, the set containing each $p!$ topological ordering of the variables V. The elements of $\mathcal{V}$ are thus ordered sequences of p variable names

$$\mathcal{V} = \{\{V_1, V_2, V_3, \dots V_p\}, \{V_2, V_1, V_3, \dots V_p\} \dots\} \tag{2.9}$$

In a second step, the input precision matrix $\boldsymbol{\Omega}$ is transformed to a model-implied variance-covariance matrix $\tilde{\Sigma}$ by taking its inverse (see Equation (2.6))

$$\boldsymbol{\Omega}^{-1} = \tilde{\Sigma}$$

under the restriction that $\tilde{\Sigma}$ is positive definite. Each element of $\mathcal{V}$ defines an ordering of the rows and columns of this matrix $\tilde{\Sigma}$. We will denote by $\tilde{\Sigma}_q$ the model-implied variance covariance matrix with rows and columns ordered according to $\mathcal{V}_q$.

For each $\tilde{\Sigma}_q$, we can calculate a corresponding weights matrix $\boldsymbol{B}_q$ by equating the relationship

$$\tilde{\Sigma_q} = (\boldsymbol{I} - \boldsymbol{B}_q)^{-1} \boldsymbol{\Psi}_q (\boldsymbol{I} - \boldsymbol{B}_q)^{-T} \tag{2.10}$$

where $\boldsymbol{\Psi}_q$ is a p- dimensional diagonal matrix of residual covariances, to an LDL^T matrix composition

$$\tilde{\Sigma_q} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T \tag{2.11}$$

where $\boldsymbol{L} = (\boldsymbol{I} - \boldsymbol{B}_q)^{-1}$ and $\boldsymbol{D} = \boldsymbol{\Psi}$. The $\boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$ decomposition is unique for positive-definite matrices (Abadir & Magnus, 2005), giving us a unique weights matrix $\boldsymbol{B}_q$ for a given topologically ordered covariance matrix $\tilde{\Sigma_q}$.

This equivalence is based on the property that a topologically ordered weights matrix of a DAG $\boldsymbol{B}$ will always be lower-triangular, with zero's on the diagonal. As a result, we can define the inverse $(\boldsymbol{I} - \boldsymbol{B}_q)^{-1}$ as a convergent infinite series

$$\begin{aligned}(\boldsymbol{I} - \boldsymbol{B}_q)^{-1} &= \boldsymbol{I} + \boldsymbol{B} + \boldsymbol{B}^2 + \boldsymbol{B}^3 + \dots \\ &= \sum_{z=0}^{\infty} \boldsymbol{B}^z\end{aligned}$$

where $\boldsymbol{B}^z$ will also be lower-triangular with zero-diagonal for all $z > 0$. This means that $(\boldsymbol{I} - \boldsymbol{B}_q)^{-1}$ will be a lower triangular matrix with diagonal elements equal to one (Abadir & Magnus, 2005).

In practice the $\boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$ decomposition of a matrix can be done by first calculating the cholesky decomposition

$$\tilde{\Sigma_q} = \boldsymbol{G}\boldsymbol{G}^T \tag{2.12}$$

where $\boldsymbol{G}$ is an upper-triangular matrix, and thus

$$\boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T = (\boldsymbol{L}\boldsymbol{D}^{1/2})(\boldsymbol{L}\boldsymbol{D}^{1/2})^T) = \boldsymbol{G}^T\boldsymbol{G} \tag{2.13}$$

Letting $\mathbf{D}^{1/2} = \mathbf{S}$, we can write

$$\mathbf{G}^{\mathbf{T}} = \mathbf{LS} \tag{2.14}$$

and so

$$\mathbf{G}^{\mathbf{T}}\mathbf{S}^{-1} = \mathbf{L} \tag{2.15}$$

As $\mathbf{S}$ is a diagonal matrix, taking it's inverse is the same as taking the scalar inverse of each diagonal element and collecting these in a diagonal matrix of the same dimensions. As $\boldsymbol{L} = (\boldsymbol{I} - \mathbf{B}_q)^{-1}$ has ones on the diagonal, it must be the case that $G_{ii}/s_{ii} = 1$, for all i, which means $s_{ii} = G_{ii}$. To find $\mathbf{L} = (\mathbf{I} - \mathbf{B}_q)^{-1}$ we simply divide each column of $\mathbf{G}$ by the corresponding diagonal element in that column. Then we have

$$\boldsymbol{B}_q = \mathbf{I} - \mathbf{L}^{-1} \tag{2.16}$$

Repeating this operation for all topological orderings in $\mathcal{V}$ defines the SE-set algorithm. The output of this algorithm thus yields a set of matrices $\boldsymbol{B}_q$ for $1 \leq q \leq p!$, which are collected in the SE-set $\boldsymbol{\mathcal{B}}$.

2.B.4 Details of R implementation

The SE-set algorithm is implemented in the statistical programming language R. The current implementation of this algorithm is computationally feasible for systems of $p \leq 12$ variables. This is due to vector size limits in R, which has maximum vector length $2^{31} - 1$. The topological orderings $\mathcal{V}$ are generated all at once and stored in a vector using the *combinat* library in R (Chasalow, 2012). This means that topological orderings cannot be generated for $p > 13$ as this would violate the maximum vector length ($13! > 2^{31} - 1$).

For $p \leq 12$ the current implementation runs extremely quickly on a standard machine, taking an average of .23 seconds for the empirical analysis in the current paper, and yields a matrix as output which is easily manipulated as a standard R object. Limited increases in the maximum number of variables may be possible using recursive programming, and storing objects outside of the R working environment, but this is not implemented currently. The size of the output, and time taken to apply all of the transformations, grows at a greater than exponential rate with p, which means that practically applicable versions of this algorithm may not be feasible for much larger systems.

With the intention of aiding researchers in exploring the SE-set, a number of options are implemented to simplify and reduce the number of matrices in the output $\boldsymbol{\mathcal{B}}$, and to mimic the typical context of researchers using GGM based techniques. First, it is recommended that researchers input a precision matrix which corresponds to a standardized covariance matrix, or which is calculated on the basis of standardized variables, as is typically recommended in GGM applications. Pre-standardizing variables ensures that the sizes of paths are directly comparable. By default thresholding and rounding of matrix elements is applied, up to two decimal places, to eliminate small paths of less interest to researchers.

Second, for ease of comparison, the rows and columns of each matrix are re-ordered to some reference-ordering, typically $\mathcal{V}_1$, the ordering of rows and columns in the input precision matrix. Third, in practice, many elements of $\boldsymbol{\mathcal{B}}$

will describe identical linear path models: for a given weighted DAG, there may be more than one equivalent topological ordering. The intended use of this tool is for researchers interested in the different distinct (causal) structures which may generate $\tilde{\Sigma}$, where we consider a-priori all distinct structures to be equally plausible. As such, the implementation of this algorithm by default removes any duplicate matrices in $\mathcal{B}$.

One disadvantage to rounding and thresholding matrix elements is that elements of $\mathcal{B}$ may no longer be statistically equivalent, that is, they may imply different variance-covariance matrices. However these differences are likely to be relatively small, and this disadvantage is offset by the increased simplification and interpretability of the different path models.

Appendix 2.C Empirical Illustration Details

In this section we will provide further technical details on the empirical illustration shown in the main text. We will first review the methodology used for constructing a GGM based on the results reported by Hoorelbeke et al. (2016), and the necessary differences in methodology used in the current paper. Following this we will validate the SE-set analysis used in the current paper by showing that the weighted DAGs which result from this analysis reproduce the precision matrix which was estimated in the GGM analysis.

2.C.1 GGM Re-analysis

Hoorelbeke et al. used the *parcor* package in their GGM analysis, which takes as input a full dataset, and returns a partial correlation matrix (Kraemer, Schaefer, & Boulesteix, 2009). As the SE-set algorithm requires a precision matrix, and we do not have access to the original dataset, it is not possible to perfectly reproduce their estimated GGM.

Instead, we use the reported sample correlation matrix, shown in the upper triangle of Table 2.1. We estimate the GGM using the *EBICglasso* function from the package *qgraph* (Epskamp et al., 2012). This function estimates a regularized precision matrix, requiring only the zero-order (marginal) correlation matrix and sample size as input, with the selection of the tuning parameter performed automatically by using EBIC fit indices. Internally, this function first estimates the precision matrix, before transforming the estimated precision matrix to a partial correlation matrix: The estimated precision matrix can be returned using the *returnAllResults* option.

The marginal correlations reported by Hoorelbeke et al. (2016), used as input to the *qgraph* package, are shown in the upper-triangle of Table 2.1. The estimated partial correlations, used to create Figure 2.5(a) are shown in the lower triangle of Table 2.1.

	BRIEF_WM	PASAT_ACC	Adapt ER	Maladapt ER	Resilience	Resid Depres
BRIEF_WM	-	.060	-.100	.280	-.660	.500
PASAT_ACC	0	-	-.260	.03	-.21	.090
Adapt ER	0	0	-	.120	.390	-.260
Maladapt ER	0	0	0	-	-.330	.300
Resilience	-.325	0	.092	-.039	-	-.640
Resid Depres	.094	0	0	0	-.301	-

Table 2.1: Reported correlations and estimated partial correlations for the empirical example. The upper triangle shows the marginal correlations reported by Hoorelbeke et al. (2016). The lower triangle shows the partial correlations estimated using the reported correlation matrix as input to the EBICglasso function from the *qgraph* package.

2.C.2 SE-set Input Recovery

To verify that the linear DAG weights matrices resulting from the SE-set algorithm $\boldsymbol{B}_q \in \boldsymbol{\mathcal{B}}$ all approximately reproduce the input precision matrix $\hat{\boldsymbol{\Omega}}$, we re-calculate the precision matrix given by each of the 720 un-rounded matrices in $\boldsymbol{\mathcal{B}}$ using Equation (2.7). For this purpose, the $\boldsymbol{\Psi}$ matrices corresponding to each $\boldsymbol{B}_q$ were saved when the SE-set was generated. In Table 2.2 we show the bias of the re-calculated precision matrices, compared to the input precision matrix. We can see that the weights matrices $\boldsymbol{B}$ reproduce the correct precision matrix, up to the 6th decimal place. The RMSE is not reported as the error for each element of $\boldsymbol{\mathcal{B}}$ was very similar, meaning the RMSE is approximately equal to the absolute value of the bias.

	BRIEF_WM	PASAT_ACC	Adapt ER	Maladapt ER	Resilience	Resid Depress
BRIEF_WM	$2.8 * 10^{-6}$					
PASAT_ACC	0	0				
Adapt ER	$-1.2 * 10^{-8}$	0	$4.3 * 10^{-8}$			
Maladapt ER	$5.3 * 10^{-9}$	0	$-1.4 * 10^{-9}$	7.5 * 10−9		
Resilience	$1.1 * 10^{-6}$	0	$-4.4 * 10^{-8}$	$1.7 * 10^{-8}$	$8.3 * 10^{-7}$	
Resid Depress	$-2.3 * 10^{-7}$	0	$-1.1 * 10^{-8}$	$4.8 * 10^{-9}$	$9.5 * 10^{-8}$	$2.0 * 10^{-7}$

Table 2.2: Average error ("bias") of elements of $\boldsymbol{\mathcal{B}}$ in reproducing the input precision matrix $\hat{\boldsymbol{\Omega}}$. Zero elements of $\hat{\boldsymbol{\Omega}}$ are highlighted in red.

We can see from Table 2.2 that positive elements of $\hat{\boldsymbol{\Omega}}$, for example the diagonal elements as well as ω_{51}, ω_{54} and ω_{65} appear to have a slight positive bias, and negative elements $(\omega_{53}, \omega_{61})$ a slight negative bias. The true zero-elements, highlighted in red in the table below, have a mix of positive, negative and zero bias. All parameters relating to PASAT_ACC, which is unconnected to all other variables, are recovered as exactly zero. However other non-zero elements are recovered as positive or negative non-zero elements, but the size of these elements is very small. This is not necessarily problematic, as the input precision matrix makes use of thresholding in any case, where such parameters would be set exactly to zero.

This performance analysis is conducted on the un-rounded equivalence set, also including duplicates. Removing duplicates does not make a substantial difference to the results due to the small variances in individual errors. Rounding

makes some difference to the performance, but not substantially: the signs of some biases change, and the absolute values of the largest biases are now larger, of the order 10^{-3}.

Chapter 3

A Continuous-Time Approach to Intensive Longitudinal Data: What, Why and How?

Abstract

The aim of this chapter is to: a) provide a broad didactic treatment of the first-order stochastic differential equation model – also known as the continuous-time (CT) first-order vector autoregressive (VAR(1)) model; and b) argue for and illustrate the potential of this model for the study of psychological processes using intensive longitudinal data. We begin by describing what the CT-VAR(1) model is, and how it relates to the more commonly used discrete-time VAR(1) model. Assuming no prior knowledge on the part of the reader, we introduce important concepts for the analysis of dynamic systems, such as stability and fixed points. In addition we examine why applied researchers should take a continuous-time approach to psychological phenomena, focusing on both the practical and conceptual benefits of this approach. Finally, we elucidate how researchers can interpret CT models, describing the direct interpretation of CT model parameters as well as tools such as impulse response functions, vector fields, and lagged-parameter plots. To illustrate this methodology we re-analyze a single-subject experience-sampling dataset with the *R* package *ctsem*; for didactic purposes, *R* code for this analysis is included, and the dataset itself is publicly available.

This chapter has been adapted from: Ryan, O., Kuiper, R. M. & Hamaker, E. L. (2018). A Continuous-Time Approach to Intensive Longitudinal Data: What, Why and How? In K. v. Montfort, J. H. L. Oud, & M. C. Voelkle (Eds.) *Continuous Time Modeling in the Behavioral and Related Sciences*. New York, NY: Springer. Author contributions: OR and ELH conceptualized the initial project. OR and RMK wrote the *R* code and generated figures. OR wrote the paper. RMK and ELH helped further develop the ideas in the project, discussed progress and provided textual feedback.

3.1 Introduction

The increased availability of intensive longitudinal data – such as obtained with ambulatory assessments, experience sampling, ecological momentary assessments and electronic diaries – has opened up new opportunities for researchers to investigate the *dynamics* of psychological *processes*, that is, the way psychological variables evolve, vary and relate to one another over time (cf. Bolger & Laurenceau, 2013; Hamaker et al., 2005; Chow, Ferrer, & Hsieh, 2011). A useful concept in this respect is that of people being dynamic systems whose current state depends on their preceding states. For instance, we may be interested in the relationship between momentary stress and anxiety. We can think of stress and anxiety as each defining an axis in a two-dimensional space, and let the values of stress and anxiety at each moment in time define a position in this space. Over time, the point that represents a person's momentary stress and anxiety moves through this two-dimensional space, and our goal is to understand the lawfulness that underlie these movements.

There are two frameworks that can be used to describe such movements: 1) the *discrete-time* (DT) framework, in which the passage of time is treated in discrete steps; and 2) the *continuous-time* (CT) framework, in which time is viewed as a continuous variable. Most psychological researchers are at least somewhat familiar with the DT approach, as it is the basis of the vast majority of longitudinal models used in the social sciences. In contrast, CT models have gained relatively little attention in fields such as psychology: This is despite the fact that many psychological researchers have been advocating their use for a long time, claiming that the CT approach overcomes practical and conceptual problems associated with the DT approach (e.g., Boker, 2002; Chow et al., 2005; Oud & Delsing, 2010; Voelkle, Oud, Davidov, & Schmidt, 2012). We believe there are two major hurdles that hamper the adoption of the CT approach in psychological research. First, the estimation of CT models typically requires the use of specialized software (cf. Driver, Oud, & Voelkle, 2017; Chow, Ferrer, & Nesselroade, 2007; Oravecz, Tuerlinckx, & Vandekerckhove, 2016) or unconventional use of more common software (cf. Boker, Neale, & Rausch, 2004; Boker, Deboeck, Edler, & Keel, 2010; J. S. Steele & Ferrer, 2011). Second, the results from CT models are not easily understood, and researchers may not know how to interpret and represent their findings.

Our goal in this chapter is twofold. First, we introduce readers to the perspective of psychological processes as CT processes; we focus on the *conceptual reasons* for which the CT perspective is extremely valuable in moving our understanding of processes in the right direction. Second, we provide a didactic description of how to interpret the results of a CT model, based on our analysis of an empirical dataset. We examine the direct interpretation of model parameters, examine different ways in which the dynamics described by the parameters can be understood and visualized, and explain how these are related to one another throughout. We will restrict our primary focus to the simplest DT and CT models, that is, first-order (vector) autoregressive models and first-order differential equations.

The organization of this chapter is as follows. First, we provide an overview of the DT and CT models under consideration. Second, we discuss the practical and conceptual reasons researchers should adopt a CT modeling approach. Third, we illustrate the use and interpretation of the CT model using a bivariate model estimated from empirical data. Fourth, we conclude with a brief discussion of more complex models which may be of interest to substantive researchers.

3.2 Two Frameworks

The relationship between the DT and CT frameworks has been discussed extensively by a variety of authors. Here, we briefly reiterate the main issues, as this is vital to the subsequent discussion. For a more thorough treatment of this topic, the reader is referred to Voelkle et al. (2012). We begin by presenting the first-order vector auto-regressive model in DT, followed by the presentation of the first-order differential equation in CT. Subsequently, we show how these models are connected, and discuss certain properties which can be inferred from the parameters of the model. For simplicity, and without loss of generalization, we describe single-subject DT and CT models, in terms of observed variables. Extensions for multiple-subject data, and extensions for latent variables, in which the researchers can account for measurement error by additionally specifying a measurement model, are readily available (in the case of CT models, see for example Boker et al., 2004; Oravecz & Tuerlinckx, 2011; Driver et al., 2017).

3.2.1 The Discrete-Time Framework

DT models are those models for longitudinal data in which the passage of time is accounted for only with regards to the order of observations. If the true data-generating model for a process is a DT model, then the process only takes on values at discrete moments in time (e.g., hours of sleep per day or monthly salary). Such models are typically applied to data that consist of some set of variables measured repeatedly over time. These measurements typically show auto-correlation, that is, serial dependencies between the observed values of these variables at consecutive measurement occasions. We can model these serial dependencies using (discrete-time) auto-regressive equations, which describe the relationship between the values of variables observed at consecutive measurement occasions.

The specific type of DT model that we will focus on in this chapter is the first-order Vector Auto-Regressive (VAR(1)) model (cf. Hamilton, 1994). Given a set of V variables of interest measured at N different occasions, the VAR(1) describes the relationship between $\boldsymbol{y}_\tau$, a $V \times 1$ column vector of variables measured at occasion τ (for $\tau = 2, \ldots, N$) and the values those same variables took on at the preceding measurement occasion, the vector $\boldsymbol{y}_{\tau-1}$. This model can be expressed as

$$\boldsymbol{y}_\tau = \boldsymbol{c} + \boldsymbol{\Phi}\boldsymbol{y}_{\tau-1} + \boldsymbol{\epsilon}_\tau, \tag{3.1}$$

where $\boldsymbol{\Phi}$ represents a $V \times V$ matrix with autoregressive and cross-lagged coefficients that regresses $\boldsymbol{y}_\tau$ on $\boldsymbol{y}_{\tau-1}$. The $V \times 1$ column vector $\boldsymbol{\epsilon}_\tau$ represents the variable-specific random shocks or innovations at that occasion, which are normally distributed with mean zero and a $V \times V$ variance-covariance matrix $\boldsymbol{\Psi}$. Finally, $\boldsymbol{c}$ represents a $V \times 1$ column vector of intercepts.

In the case of a stationary process, the mean $\boldsymbol{\mu}$, and the variance-covariance matrix of the variables $\boldsymbol{y}_\tau$ (generally denoted $\boldsymbol{\Sigma}$), do not change over time.[1] Then, the vector $\boldsymbol{\mu}$ represents the long-run expected values of the random variables, $E(\boldsymbol{y}_\tau)$, and is a function of the vector of intercepts and the matrix with lagged regression coefficients, that is, $\boldsymbol{\mu} = (\boldsymbol{I} - \boldsymbol{\Phi})^{-1}\boldsymbol{c}$, where $\boldsymbol{I}$ is a $V \times V$ identity matrix (cf. Hamilton, 1994). In terms of a V-dimensional dynamical system of interest, $\boldsymbol{\mu}$ represents the *equilibrium position* of the system. By definition, τ is limited to positive integers; that is, there is no .1th or 1.5th measurement occasion.

Both the single-subject and multilevel versions of the VAR(1) model have frequently been used to analyze intensive longitudinal data of psychological variables, including symptoms of psychopathology, such as mood- and affect-based measures (Browne & Nesselroade, 2005; Moberly & Watkins, 2008; Rovine & Walls, 2006; Bringmann et al., 2016; Bringmann, Lemmens, Huibers, Borsboom, & Tuerlinckx, 2015). In these cases, the auto-regressive parameters ϕ_{ii} are often interpreted as reflecting the *stability*, *inertia* or *carry-over* of a particular affect or behavior (Kuppens, Allen, & Sheeber, 2010; Kuppens et al., 2012; Koval, Kuppens, Allen, & Sheeber, 2012). The cross-lagged effects (i.e., the off-diagonal elements ϕ_{ij} for $i \neq j$) quantify the lagged relationships, sometimes referred to as the *spill-over*, between different variables in the model. These parameters are often interpreted in substantive terms, either as predictive or Granger-causal relationships between different aspects of affect or behavior (Granger, 1969; Ichii, 1991; Watkins, Lei, & Canivez, 2007; Gault-Sherman, 2012; Bringmann et al., 2013). For example, if the standardized cross-lagged effect of $y_{1,\tau-1}$ on $y_{2,\tau}$ is larger than the cross-lagged effect of $y_{2,\tau-1}$ on $y_{1,\tau}$, researchers may draw the conclusion that y_1 is the driving force or dominant variable of that pair (Schuurman, Ferrer, de Boer-Sonnenschein, & Hamaker, 2016). As such, substantive researchers are typically interested in the (relative) magnitudes and signs of these parameters.

3.2.2 The Continuous-Time Framework

In contrast to the DT framework, which treats values of processes indexed by observation τ, the CT framework treats processes as *functions* of the continuous variable time t: The processes being modeled are assumed to vary continuously with respect to time, meaning that these variables may take on values if observed at any imaginable moment. CT processes can be modeled using a broad class of differential equations, allowing for a wide degree of diversity in the types of dynamics that are being modeled. It is important to note that many DT models have

[1] The variance-covariance matrix of the variables $\boldsymbol{\Sigma}$ is a function of both the lagged parameters and the variance-covariance matrix of the innovations, $vec(\boldsymbol{\Sigma}) = (\boldsymbol{I} - \boldsymbol{\Phi} \otimes \boldsymbol{\Phi})^{-1} vec(\boldsymbol{\Psi})$, where $vec(.)$ denotes the operation of putting the elements of an $N \times N$ matrix into an $NN \times 1$ column matrix (p.27 Kim & Nelson, 1999)

a differential equation counterpart. For the VAR(1) model, the CT equivalent is the first-order Stochastic Differential Equation (SDE), where stochastic refers to the presence of random innovations or shocks.

The first-order SDE describes how the position of the V-dimensional system at a certain point in time, $\boldsymbol{y}(t)$, relative to the equilibrium position $\boldsymbol{\mu}$, is related to the *rate of change* of the process with respect to time (i.e., $\frac{d\boldsymbol{y}(t)}{dt}$) in that same instant. The latter can also be thought of as a vector of *velocities*, describing in what direction and with what magnitude the system will move an instant later in time (i.e., the ratio of the change in position over some time-interval, to the length of that time-interval, as the length of the time-interval approaches zero). The first-order SDE can be expressed as

$$\frac{d\boldsymbol{y}(t)}{dt} = \boldsymbol{A}(\boldsymbol{y}(t) - \boldsymbol{\mu}) + \boldsymbol{G}\frac{d\boldsymbol{W}(t)}{dt}. \tag{3.2}$$

The model representation in Equation (3.2) is referred to as the differential form as it includes the derivative $\frac{d\boldsymbol{y}(t)}{dt}$. The same model can be represented in the integral form, in which the derivatives are integrated out, sometimes referred to as the solution of the derivative model. The integral form of this particular first-order differential equation is known as the CT-VAR(1) or Ornstein-Uhlenbeck model (Oravecz, Tuerlinckx, & Vandekerckhove, 2011). In this form, we can describe the same system, but now in terms of the positions of the system (i.e., the values the variables take on) at different points in time. For notational simplicity, we can represent $\boldsymbol{y}(t) - \boldsymbol{\mu}$ as $\boldsymbol{y}^{\boldsymbol{c}}(t)$, denoting the position of the process as a deviation from its equilibrium.

The CT-VAR(1) model can be written as

$$\boldsymbol{y}^{\boldsymbol{c}}(t) = \boldsymbol{e}^{\boldsymbol{A}\Delta t}\boldsymbol{y}^{\boldsymbol{c}}(t - \Delta t) + \boldsymbol{w}(\Delta t) \tag{3.3}$$

where $\boldsymbol{A}$ has the same meaning as above, the $V \times 1$ vector $\boldsymbol{y}^{\boldsymbol{c}}(t - \Delta t)$ represents the position as a deviation from equilibrium some time-interval Δt earlier, $\boldsymbol{e}$ represents the matrix exponential function, and the $V \times 1$ column vector $\boldsymbol{w}(\Delta t)$ represents the stochastic innovations, the integral form of the Wiener process in Equation (3.2). These innovations are normally distributed with a variance-covariance matrix that is a function of the time-interval between measurements Δt, the drift matrix $\boldsymbol{A}$, and the diffusion matrix $\boldsymbol{\Gamma}$ (cf. Voelkle et al., 2012).[2] As the variables in the model have been centered around their equilibrium, we omit any intercept term. The relationship between lagged variables, that is, the relationships between the positions of the centered variables in the multivariate space, separated by some time-interval Δt, is an (exponential) function of the drift matrix $\boldsymbol{A}$ *and* the length of that time-interval.

[2] Readers should note that there are multiple different possible ways to parameterize the CT stochastic process in integral form, and also multiple different notations used (e.g., Oravecz et al., 2011; Voelkle et al., 2012).

3.2.3 Relating DT and CT Models

It is clear from the integral form of the first-order SDE given in Equation (3.3) that the relationship between lagged values of variables is dependent on the length of the time-interval between these lagged values. As such, if the DT-VAR(1) model in Equation 3.1 is fitted to data generated by the CT model considered here, then the auto-regressive and cross lagged effects matrix $\mathbf{\Phi}$ will be a function of the time-interval Δt between the measurements. We denote this dependency by writing $\mathbf{\Phi}(\Delta t)$. This characteristic of the DT model has been referred to as the lag or time-interval problem (Gollob & Reichardt, 1987; Reichardt, 2011).

The precise relationship between the CT-VAR(1) and DT-VAR(1) effects matrices is given by the well-known equality

$$\mathbf{\Phi}(\Delta t) = \boldsymbol{e}^{\boldsymbol{A}\Delta t}. \tag{3.4}$$

Despite this relatively simple relationship, it should be noted that taking the exponential of a matrix is not equivalent to taking the exponential of each of the elements of the matrix. That is, any lagged effect parameter $\phi_{ij}(\Delta t)$, relating variable i and variable j across time-points, is not only dependent on the corresponding CT cross-effect a_{ij}, but is a non-linear function of the interval and *every other element* of the matrix $\boldsymbol{A}$. For example, in the bivariate case the DT cross-lagged effect of $y_1(t-\Delta t)$ on $y_2(t)$, denoted $\phi_{21}(\Delta t)$, is given by

$$\frac{a_{21}(e^{\frac{1}{2}(a_{11}+a_{22}+\sqrt{a_{11}^2+4a_{12}a_{21}-2a_{11}a_{22}+a_{22}^2})\Delta t} - e^{\frac{1}{2}(a_{11}+a_{22}-\sqrt{a_{11}^2+4a_{12}a_{21}-2a_{11}a_{22}+a_{22}^2})\Delta t})}{\sqrt{a_{11}^2+4a_{12}a_{21}-2a_{11}a_{22}+a_{22}^2}} \tag{3.5}$$

where e represents the scalar exponential. In higher dimensional models, these relationships quickly become intractable. For a derivation of Equation 3.5 we refer readers to Appendix 3.A.

This complicated non-linear relationship between the elements of $\mathbf{\Phi}$ and the time-interval has major implications for applied researchers who wish to interpret the parameters of a DT-VAR(1) model in the substantive terms outlined above. In the general multivariate case, the size, sign, and relative strengths of both auto-regressive and cross-lagged effects may differ depending on the value of the time-interval used in data collection (Oud, 2007; Reichardt, 2011; Dormann & Griffin, 2015; Deboeck & Preacher, 2016). As such, conclusions that researchers draw regarding the stability of processes, and the nature of how different processes relate to one another may differ greatly depending on the time-interval used.

While the relationship in Equation 3.4 describes the DT-VAR(1) effects matrix we would find given data generated by a CT-VAR(1) model, the reader should note that not all DT-VAR(1) processes have a straightforward equivalent representation as a CT-VAR(1). For example, a univariate discrete-time AR(1) process with a negative auto-regressive parameter cannot be represented as a CT-AR(1) process; as the exponential function is always positive, there is no A that satisfies

Equation 3.4 for $\Phi < 0$. As such, we can refer to DT-VAR(1) models with a CT-VAR(1) equivalent as those which exhibit 'positive autoregression'. We will focus throughout on the CT-VAR(1) as the data-generating model.[3]

3.2.4 Types of Dynamics: Eigenvalues, Stability and Equilibrium

Both the DT-VAR(1) model and the CT-VAR(1) model can be used to describe a variety of different types of dynamic behavior. As the dynamic behavior of a system is always understood with regards to how the variables in the system move in relation to the equilibrium position, often dynamic behaviors are described by differentiating the type of equilibrium position or *fixed point* in the system (Strogatz, 2015). In the general multivariate case, we can understand these different types of dynamic behavior or fixed points with respect to the *eigenvalues* of the effects matrices $\boldsymbol{A}$ or $\boldsymbol{\Phi}$ (see Appendix 3.A for a more detailed explanation of the relationship between these two matrices and eigenvalues). In this chapter we will focus on *stable* processes, in which, given a perturbation, the system of interest will inevitably return to the equilibrium position. We limit our treatment to these types of processes, because we believe these are most common and most relevant for applied researchers. A brief description of other types of fixed points and how they relate to the eigenvalues of the effects matrix $\boldsymbol{A}$ is given in the discussion section – for a more complete taxonomy we refer readers to Strogatz (2015).

In DT settings, stable processes are those for which the absolute values of the eigenvalues of $\boldsymbol{\Phi}$ are smaller than one. In DT applications researchers also typically discuss the need for *stationarity*, that is, time-invariant mean and variance, as introduced above. Stability of a process ensures that stationarity in relation to the mean and variance hold. For CT-VAR(1) processes, stability is ensured if the real-parts of the eigenvalues of $\boldsymbol{A}$ are negative. It is interesting to note that the equilibrium position of stable processes can be related to our observed data in various ways: In some applications $\boldsymbol{\mu}$ is constrained to be equal to the mean of the observed values (e.g., Hamaker & Grasman, 2015; Hamaker et al., 2005), while in others the equilibrium can be specified a-prior or estimated to be equal to some (asymptotic) value (e.g., Bisconti, Bergeman, & Boker, 2004).

We can further distinguish between dynamic processes that have real eigenvalues, complex eigenvalues, or in the case of systems with more than two variables, a mix of both. In the section "Making Sense of CT Models" we will focus on the interpretation of a CT-VAR(1) model with real, negative, non-equal eigenvalues. We can describe the equilibrium position of this system as a *stable node*. In the discussion section we examine another type of system which has been the focus of psychological research, sometimes described as a damped linear oscilla-

[3] In general, there is no straightforward CT-VAR(1) representation of DT-VAR(1) models with real, negative eigenvalues. However it may be possible to specify more complex continuous-time models which do not exhibit positive autoregression. Notably, M. Fisher (2001) demonstrates how a DT-AR(1) model with negative autoregressive parameter can be modeled with the use of two continuous-time (so-called) Ito processes.

tor (e.g., Boker, Montpetit, Hunter, & Bergeman, 2010), in which the eigenvalues of $\boldsymbol{A}$ are complex, with a negative real part. The fixed point of such a system is described as a *stable spiral*. Further detail on the interpretation of these two types of systems is given in the corresponding sections.

3.3 Why Researchers Should Adopt a CT Perspective

There are both practical and theoretical benefits to CT model estimation over DT model estimation. Here we will discus three of these practical advantages which have received notable attention in the literature. We then discuss the fundamental conceptual benefits of treating psychological processes as continuous-time systems.

The first practical benefit to CT model estimation is that the CT model deals well with observations taken at unequal intervals, often the case in experience sampling and ecological momentary assessment datasets (Oud & Jansen, 2000; Voelkle et al., 2012; Voelkle & Oud, 2013). Many studies use random intervals between measurements, for example to avoid participant-anticipation of measurement occasions, potentially resulting in unequal time-intervals both within and between participants. The DT model, however, is based on the assumption of equally spaced measurements, and as such estimating the DT model from unequally spaced data will result in an estimated $\boldsymbol{\Phi}$ matrix that is a blend of different $\boldsymbol{\Phi}(\Delta t)$ matrices for a range of values of Δt.

The second practical benefit of CT modeling over DT modeling is that, when measurements are equally spaced, the lagged effects estimated by the DT models are not generalizable beyond the time-interval used in data collection. Several different researchers have demonstrated that utilizing different time-intervals of measurement can lead researchers to reach very different conclusions regarding the values of parameters in $\boldsymbol{\Phi}$ (Reichardt, 2011; Oud & Jansen, 2000; Voelkle et al., 2012). The CT model has thus been promoted as facilitating better comparisons of results between studies, as the CT effects matrix $\boldsymbol{A}$ is independent of time-interval (assuming a sufficient frequency of measurement to capture the relevant dynamics).

Third, the application of CT models allows us to explore how cross-lagged effects are expected to change depending on the time-interval between measurements, using the relationship expressed in Equation (3.4). Some authors have used this relationship to identify the time-interval at which cross-lagged effects are expected to reach a maximum (Deboeck & Preacher, 2016; Dormann & Griffin, 2015). Such information could be used to decide upon the 'optimal' time-interval that should be used in gathering data in future research.

While these practical concerns regarding the use of DT models for CT processes are legitimate, there may be instances in which alternative practical solutions can be used, without necessitating the estimation of a CT model. For instance, the problem of unequally spaced measurements in DT modeling can be addressed by defining a time grid and adding missing data to your observations, to make the occasions approximately equally spaced in time. Some simu-

lation studies indicate that this largely reduces the bias that results from using DT estimation of unequally spaced data (De Haan-Rietdijk, Voelkle, Keijsers, & Hamaker, 2017).

Furthermore, the issue of comparability between studies that use different time-intervals can be solved, in certain circumstances, by a simple transformation of the estimated $\boldsymbol{\Phi}$ matrix, described in more detail by Kuiper and Ryan (2018). Given an estimate of $\boldsymbol{\Phi}(\Delta t)$ we can solve for the underlying $\boldsymbol{A}$ using Equation 3.4. This is known as the "indirect method" of CT model estimation (Oud, van Leeuwe, & Jansen, 1993). However this approach cannot be applied in all circumstances, as it involves using the matrix logarithm, the inverse of the matrix exponential function. As the matrix logarithm function in the general case does not give a unique solution, this method is only appropriate if both the estimated $\boldsymbol{\Phi}(\Delta t)$ and true underlying $\boldsymbol{A}$ matrices have real eigenvalues only (for further discussion of this issue see Hamerle, Nagl, & Singer, 1991).

However, the CT perspective has added value above and beyond the potential practical benefits discussed above. Multiple authors have argued that psychological phenomena, such as stress, affect and anxiety, do not vary in discrete steps over time, but likely vary and evolve in a continuous and smooth manner (Boker, 2002; Gollob & Reichardt, 1987). Viewing psychological processes as CT dynamic systems has important implications for the way we conceptualize the influence of psychological variables on each other. Gollob and Reichardt (1987) give the example of a researcher who is interested in the effect of taking aspirin on headaches: This effect may be zero shortly after taking the painkiller, substantial an hour or so later, and near zero again after twenty-four hours. All of these results may be considered as accurately portraying the effect of painkillers on headaches *for a specific time-interval*, although each of these intervals considered separately represent only a snapshot of the process of interest.

It is only through examining the underlying dynamic trajectories, and exploring how the cross-lagged relationships evolve and vary as a function of the time-interval, that we can come to a more complete picture of the dynamic system of study. We believe that – while the practical benefits of CT modeling are substantial – the conceptual framework of viewing psychological variables as CT processes has the potential to transform longitudinal research in this field.

3.4 Making Sense of CT Models

In this section, we illustrate how researchers can evaluate psychological variables as dynamic CT processes by describing the interpretation of the drift matrix parameters $\boldsymbol{A}$. We describe multiple ways in which the dynamic behavior of the model in general, as well as specific model parameters, can be understood. In order to aid researchers who are unfamiliar with this type of analysis, we take a broad approach in which we incorporate the different ways in which researchers interested in dynamical systems and similar models interpret their results. For instance, Boker and colleagues (e.g., Boker, Montpetit, et al., 2010) typically interpret the differential form of the model directly; in the econometrics literature it is

typical to plot specific trajectories using Impulse Response Functions (Johnston & DiNardo, 1972); in the physics tradition, the dynamics of the system are inspected using Vector Fields (e.g., Boker & McArdle, 1995); the work of Voelkle, Oud and others (e.g., Voelkle et al., 2012; Deboeck & Preacher, 2016) typically focuses on the integral form of the equation, and visually inspecting the time-interval dependency of lagged effects.

We will approach the interpretation of a single CT model from these four angles, and show how they each represent complimentary ways to understand the same system. For ease of interpretation we focus here on a bivariate system; the analysis of larger systems is addressed in the discussion section.

3.4.1 Substantive Example from Empirical Data

To illustrate the diverse ways in which the dynamics described by the CT-VAR(1) model can be understood, we make use of a substantive example. This example is based on our analysis of a publicly-available single-subject ESM dataset (Kossakowski, Groot, Haslbeck, Borsboom, & Wichers, 2017). The subject in question is a 57-year old male with a history of major depression. The data consists of momentary, daily and weekly items relating to affective states. The assessment period includes a double-blind phase in which the dosage of the participants anti-depression medication was reduced. We select only those measurements made in the initial phases of the study, before medication reduction; it is only during this period that we would expect the system of interest to be stable. The selected measurements consists of 286 momentary assessments over a period of 42 consecutive days. The modal time-interval between momentary assessments was 1.766 hours (inter-quartile range of 1.250 to 3.323).

For our analysis we selected two momentary assessment items, "I feel down" and "I am tired", which we will name Down (Do) and Tired (Ti), respectively. Feeling down is broadly related to assessments of negative affect (Meier & Robinson, 2004), and numerous cross-sectional analyses have suggested a relationship between negative affect and feelings of physical tiredness or fatigue (e.g., Denollet & De Vries, 2006). This dataset afforded us the opportunity to investigate the links between these two processes from a dynamic perspective. Each variable was standardized before the analysis to facilitate ease of interpretation of the parameter estimates. Positive values of Do indicate that the participant felt down more than average, negative values indicate below-average feelings of being down, and likewise for positive and negative values of Ti.

The analysis was conducted using the *ctsem* package in *R* (Driver et al., 2017). Full details of the analysis can be found in Appendix 3.B. Parameter estimates and standard errors are given in Table 3.1, including estimates of the stochastic part of the CT model, represented by the diffusion matrix $\mathbf{\Gamma}$. The negative value of γ_{21} indicates that there is a negative co-variance between the stochastic input to the rates of change of Do and Ti; in terms of the CT-VAR(1) representation, there is a negative covariance between the residuals of Do and Ti in the same measurement occasion. Further interpretation of the diffusion matrix falls beyond the scope of the current chapter. As the analysis is meant as an illustrative

Parameter	Value	Std. Error
a_{11}	-0.995	0.250
a_{21}	0.375	0.441
a_{12}	0.573	0.595
a_{22}	-2.416	1.132
γ_{11}	1.734	0.612
γ_{21}	-0.016	0.650
γ_{22}	4.606	1.374

Table 3.1: Parameter estimates from substantive example

example only, we will throughout interpret the estimated drift matrix parameter as though they are true population parameters.

3.4.2 Interpreting the Drift Parameters

The drift matrix relating the processes Down ($Do(t)$) and Tired ($Ti(t)$) is given by

$$\boldsymbol{A} = \begin{bmatrix} -0.995 & 0.573 \\ 0.375 & -2.416 \end{bmatrix}. \tag{3.6}$$

As the variables are standardized, the equilibrium position is $\boldsymbol{\mu} = [0, 0]$ (i.e., $E[Do(t)] = E[Ti(t)] = 0$). The drift matrix $\boldsymbol{A}$ describes how the position of the system at any particular time t (i.e., $Do(t)$ and $Ti(t)$) relates to the *velocity* or *rate of change* of the process, that is, how the position of the process is changing. The system of equations which describe the dynamic system made up of Down and Tired is given by

$$\begin{bmatrix} E\left[\frac{dDo(t)}{dt}\right] \\ E\left[\frac{dTi(t)}{dt}\right] \end{bmatrix} = \begin{bmatrix} -0.995 & 0.573 \\ 0.375 & -2.416 \end{bmatrix} \begin{bmatrix} Do(t) \\ Ti(t) \end{bmatrix} \tag{3.7}$$

such that

$$E\left[\frac{dDo(t)}{dt}\right] = -0.995Do(t) + 0.573Ti(t) \tag{3.8}$$

$$E\left[\frac{dTi(t)}{dt}\right] = 0.375Do(t) - 2.416Ti(t) \tag{3.9}$$

where the rates of change of Down and Tired at any point in time are both dependent on the positions of both Down and Tired at that time.

Before interpreting any particular parameter in the drift matrix, we can determine the type of dynamic process under consideration by inspecting the eigenvalues of $\boldsymbol{A}$. The eigenvalues of $\boldsymbol{A}$ are $\lambda_1 = -2.554$ and $\lambda_2 = -0.857$; since both eigenvalues are negative, the process under consideration is stable. This means that if the system takes on a position away from equilibrium (e.g., due to a random shock from the stochastic part of the model on either Down or Tired), the

system will inevitably return to its equilibrium position over time. It is for this reason that the equilibrium position or fixed point in stable systems is also described as the *attractor point*, and stable systems are described as *equilibrium-reverting*. As the eigenvalues of the system are real-valued as well as negative, the system returns to equilibrium with an *exponential decay*; when the process is far away from the equilibrium, it takes on a greater velocity, that is, moves faster towards equilibrium. We can refer to the type of fixed point in this system as a *stable node* (Strogatz, 2015).

Typical of such an equilibrium-reverting process, we see negative CT auto-effects $a_{11} = -0.995$ and $a_{22} = -2.416$. This reflects that, if either variable in the system takes on a position away from the equilibrium, they will take on a velocity of opposite sign to this deviation, that is, a velocity which returns the process to equilibrium. For higher values of $Do(t)$, the rate of change of $Do(t)$ is of greater (negative) magnitude, that is, the velocity towards the equilibrium is higher. In addition, the auto-effect of $Ti(t)$ is more than twice as strong (in an absolute sense) as the auto-effect of $Do(t)$. If there were no cross-effects present, this would imply that $Ti(t)$ returns to equilibrium faster than $Do(t)$; however, as there are cross-effects present, such statements cannot be made in the general case from inspecting the auto-effects alone.

In this case the cross-effects of $Do(t)$ and $Ti(t)$ on each others rates of change are positive rather than negative. Moreover, the cross-effect of $Ti(t)$ on the rate of change of $Do(t)$ ($a_{12} = 0.573$) is slightly stronger than the corresponding cross-effect of $Do(t)$ on the rate of change of $Ti(t)$ ($a_{21} = 0.375$). These cross-effects quantify the force that each component of the system exerts on the other. However, depending on what values each variable takes on at a particular point in time t, that is, the position of the system in each of the $Do(t)$ and $Ti(t)$ dimensions, this may translate to $Do(t)$ pushing $Ti(t)$ to return faster to its equilibrium or to deviate away from its equilibrium position, and vice versa. To better understand both the cross-effects and auto-effects described by $\boldsymbol{A}$, it is helpful to visualize the possible trajectories of our two-dimensional system.

3.4.3 Visualizing Trajectories

We will now describe and apply two related tools which allow us to visualize the trajectories of the variables in our model over time: Impulse Response Functions and Vector Fields. These tools can help us to understand the dynamic system we are studying, by exploring the dynamic behavior which results from the drift matrix parameters.

3.4.3.1 Impulse Response Functions

Impulse Response Functions (IRFs) are typically used in the econometrics literature to aid in making forecasts based on a DT-VAR model. The idea behind this is to allow us to explore how an impulse to one variable in the model at occasion τ will affect the values of both itself and the other variables in the model at occasions $\tau + 1$, $\tau + 2$, $\tau + 3$ and so on. In the stochastic systems we focus on

in this chapter, we can conceptualize these impulses as random perturbations or innovations, or alternatively as external interventions in the system.[4] IRFs thus represent the trajectories of the variables in the model over time, following a particular impulse, assuming no further stochastic innovations (see Chapter 9 of Johnston & DiNardo, 1972, for more detail).

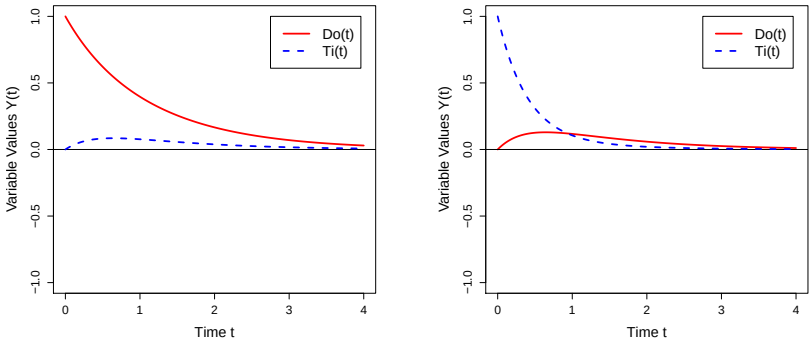
To specify impulses in an IRF, we generally assign a value to a single variable in the system at some initial occasion, $y_{i,\tau}$. The corresponding values of the other variables at the initial occasion $y_{j,\tau}$, $j \neq i$ are usually calculated based on, for instance, the covariance in the stochastic innovations, $\boldsymbol{\Psi}$, or the stable covariance between the processes $\boldsymbol{\Sigma}$. Such an approach is beneficial in at least two ways: first, it allows researchers to specify impulses which are more likely to occur in an observed dataset; second, it aids researchers in making more accurate future predictions or forecasts. For a further discussion of this issue in relation to DT-VAR models, we refer the reader to Johnston and DiNardo (1972) pages 298-300. Below, we will take a simplified approach and specify bi-variate impulses at substantively interesting values.

The IRF can easily be extended for use with the CT-VAR(1) model. We can calculate the impulse response of our system by taking the integral form of the CT-VAR(1) model in Equation (3.3) and a) plugging in the $\boldsymbol{A}$ matrix for our system, b) choosing some substantively interesting set of impulses $\boldsymbol{y}(t = 0)$ and c) calculating $\boldsymbol{y}(t)$ for increasing values of $t > 0$. To illustrate this procedure, we will specify four substantively interesting sets of impulses. The four sets of impulses shown here include $\boldsymbol{y}(0) = [1, 0]$, reflecting what happens when $Do(0)$ takes on a positive value 1 standard deviation above the persons mean, while $Ti(0)$ is at equilibrium; $\boldsymbol{y}(0) = [0, 1]$ reflecting when $Ti(0)$ takes on a positive value of corresponding size while $Do(0)$ is at equilibrium; $\boldsymbol{y}(0) = [1, 1]$ reflecting what happens when $Do(0)$ and $Ti(0)$ both take on values 1 standard deviation above the mean; and $\boldsymbol{y}(0) = [1, -1]$ reflecting what happens when $Do(0)$ and $Ti(0)$ take on values of equal magnitude but opposite valence (1SD more and 1SD less than the mean respectively). Figures 3.1(a) to 3.1(d) contain the IRFs for both processes in each of these four scenarios.

Examining the IRFs shows us the equilibrium-reverting behavior of the system: Given any set of starting values, the process eventually returns, in an exponential fashion, to the bivariate equilibrium position where both processes take on a value of zero. In Figure 3.1(a), we can see that when $Ti(t)$ is at equilibrium and $Do(0)$ takes on a value of plus one, then $Ti(t)$ is pushed away from equilibrium in the same (i.e., positive) direction. In substantive terms, when our participant is feeling down at a particular moment, he begins to feel a little tired. Eventually, both $Do(t)$ and $Ti(t)$ return to equilibrium due to their negative auto-effects. The feelings of being down and tired have returned to normal around $t = 4$, that is, four hours after the initial impulse; stronger impulses ($|Do(0)| > 1$) will result in the system taking longer to return to equilibrium, and weaker impulses ($|Do(0)| < 1$) would dissipate quicker.

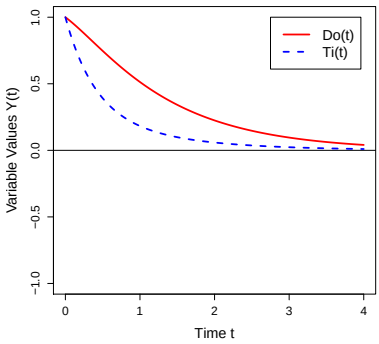
Figure 3.1(b) shows the corresponding reaction of $Do(t)$ at equilibrium to a

[4] Similar functions can be used for deterministic systems (those without a random innovation part), however in these cases the term *initial value* is more typically used.

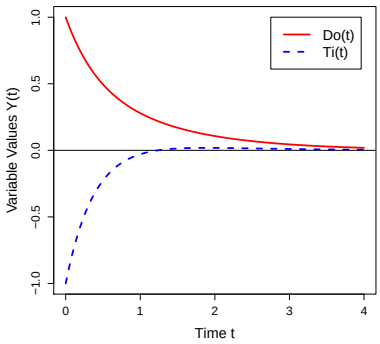


(a) $Do(0) = 1, Ti(0) = 0$

(b) $Do(0) = 0, Ti(0) = 1$



(c) $Do(0) = 1, Ti(0) = 1$



(d) $Do(0) = 1, Ti(0) = -1$

Figure 3.1: Impulse response function for the model in Equation (3.9) for four different sets of impulses; red solid line = $Do(t)$ and blue dashed line = $Ti(t)$.

positive value of $Ti(0)$. We can further see that the deviation of $Do(t)$ in Figure 3.1(b) is greater than the deviation of $Ti(t)$ in Figure 3.1(a): a positive value of $Ti(t)$ exerts a greater push on $Do(t)$ than vice versa, because of the greater cross-effect of $Ti(t)$ on $Do(t)$. In this case this strong cross-effect, combined with the relatively weaker auto-effect of $Do(t)$, results in $Do(t)$ taking on a higher value than $Ti(t)$ at around $t = 1$, one hour after the initial impulse. Substantively, when our participant is feeling physically tired at a particular moment (Figure 3.1(b)), he begins to feel a down over the next couple of hours, before eventually these feelings return to normal (again in this case, around 4 hours later).

Figures 3.1(c) further demonstrates the role of the negative auto-effects and positive cross-effects in different scenarios. In Figure 3.1(c), both processes take on positive values at $t = 0$; the positive cross-effects result in both processes returning to equilibrium at a slower rate than in Figures 3.1(a) and 3.1(b). In substantive terms this means that, when the participant is feeling very down, and very tired, it takes longer for the participant to return to feeling normal. Here

also the stronger auto-effect of $Ti(t)$ than $Do(t)$ is evident: although both processes start at the same value, an hour later $Ti(1)$ is much closer to zero than $Do(1)$, that is, $Ti(t)$ decays faster to equilibrium than $Do(t)$. In substantive terms, this tells us that when the participant is feeling down and physically tired, he recovers much quicker from the physical tiredness than he does from feeling down.

In Figure 3.1(d), we see that $Do(0)$ and $Ti(0)$ taking on values of opposite signs results in a speeding-up of the rate at which each variable decays to equilibrium. The auto-effect of $Do(t)$ is negative, which is added to by the positive cross-effect of $Ti(t)$ multiplied by the negative value of $Ti(0)$. This means that $Do(0)$ in Figure 3.1(d) takes on a stronger negative velocity, in comparison to Figures 3.1(a) or 3.1(c). A positive value for $Do(0)$ has a corresponding effect of making $Ti(0)$ take on an even stronger positive velocity. Substantively, this means that when the participant feels down, but feels less tired (i.e. more energetic) than usual, both of these feelings wear off and return to normal quicker than in the other scenarios we examined. The stronger auto-effect of $Ti(t)$, in combination with the positive cross-effect of $Do(t)$ on $Ti(t)$, actually results in $Ti(t)$ shooting past the equilibrium position in the $Ti(t)$ dimension ($Ti(t) = 0$) and taking on positive values around $t = 1.5$, before the system as a whole returns to equilibrium. Substantively, when the participant is feeling initially down but quite energetic, we expect that he feels a little bit more tired than usual about an hour and half later, before both feelings return to normal.

3.4.3.2 Vector Field

Vector fields are another technique which can be used to visualize the dynamic behavior of the system by showing potential trajectories through a bivariate space. In our case the two axes of this space are $Do(t)$ and $Ti(t)$. The advantage of vector fields over IRFs in this context is that in one plot it shows how, for a range of possible starting positions, the process is expected to move in the (bivariate) space a moment later. For this reason, the vector field is particularly useful in bivariate models with complex dynamics, in which it may be difficult to obtain the full picture of the dynamic system from a few IRFs alone. Furthermore, by showing the dynamics for a grid of values, we can identify areas in which the movement of the process is similar or differs.

To create a vector field, $E\left[\frac{d\boldsymbol{y}(t)}{dt}\right]$ is calculated for a grid of possible values for $y_1(t)$ and $y_2(t)$ covering the full range of the values both variables can take on. The vector field for $Do(t)$ and $Ti(t)$ is shown in Figure 3.2. The base of each arrow represents a potential position of the process $\boldsymbol{y}(t)$. The head of the arrow represents where the process will be if we take one small step in time forward, that is the value of $\boldsymbol{y}(t + \Delta t)$ as Δt approaches zero. In other words, the arrows in this vector field represent the information of two derivatives, $dDo(t)/dt$ and $dTi(t)/dt$. Specifically, the direction the arrow is pointing is a function of the sign (positive or negative) of the derivatives, while the length of the arrow represents the magnitude of this movement, and is a function of the absolute values of the derivative(s).

If an arrow in the vector field is completely vertical, this means that, for that

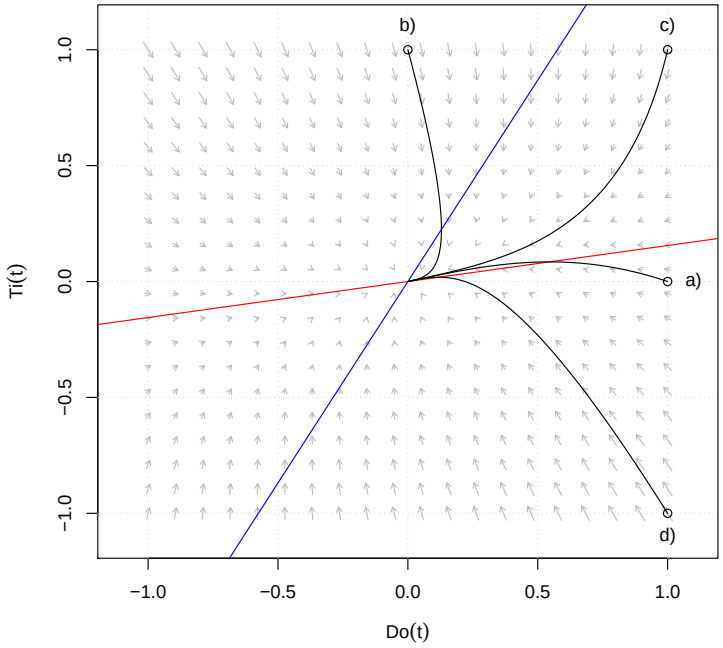


Figure 3.2: Vector field for $Do(t)$ and $Ti(t)$, including blue and red nullclines.

position, taking one small step forward in time would result in a change in the system's position along the $Ti(t)$ axis (i.e., a change in the value of Tired), but not along the $Do(t)$ axis (that is, $dDo(t)/dt = 0$ and $dTi(t)/dt \neq 0$). The converse is true for a horizontal arrow (that is, $dDo(t)/dt \neq 0$ and $dTi(t)/dt = 0$). The two lines in Figure 3.2, blue and red, identify at which positions $dDo(t)/dt = 0$ and $dTi(t)/dt = 0$, respectively; these are often referred to as *nullclines* or equivalently, solution lines. If the nullclines are not perfectly perpendicular to one another, this is due to the presence of at least one cross-effect. The point at which these nullclines cross represents the equilibrium position in this two-dimensional space, here located at $Do(t) = 0, Ti(t) = 0$. The crossing of these nullclines splits the vector field in four quadrants, each of which is characterized by a different combination of negative and positive values for $dDo(t)/dt$ and $dTi(t)/dt$. The top left and bottom right quadrants represent areas in which the derivatives are of opposite sign, $dDo(t)/dt > 0$ & $dTi(t)/dt < 0$ and $dDo(t)/dt < 0$ & $dTi(t)/dt > 0$, respectively. The top right and bottom left quadrants represent areas where the derivatives are of the same sign, $dDo(t)/dt < 0$ & $dTi(t) < 0$ and $dDo(t)/dt > 0$ & $dTi(t) > 0$, respectively.

By tracing a path through the arrows, we can see the trajectory of the system of interest from any point in the possible space of values. In Figure 3.2, we

include the same four bi-variate trajectories as we examined with the IRFs. Instead of the IRF representation of two variables whose values are changing, the vector field represents this as the movement of one process in a two-dimensional space. For instance, the trajectory starting at $Do(t) = 0$ and $Ti(t) = 1$ begins in the top-left quadrant, where $dDo(t)/dt$ is positive and $dTi(t)/dt$ is negative; this implies that the value of Down will increase, and the value of Tired will decrease (as can be seen in Figure 3.1(b)). Instead of moving directly to the equilibrium along the $Ti(t)$ dimension, the system moves away from equilibrium along the $Do(t)$ dimension, due to the cross-effect of $Ti(t)$ on $Do(t)$, until it moves into the top-right quadrant. In this quadrant, $dDo(t)/dt$ and $dTi(t)/dt$ are both negative; once in this quadrant the process moves towards equilibrium, tangent to the $dDo(t)/dt$ nullcline. The other trajectories in Figure 3.2 analogously describe the same trajectories as in Figure 3.1(a), 3.1(c) and 3.1(d).

In general, the trajectories in this vector field first decay quickest along the $Ti(t)$ dimension, and slowest along the $Do(t)$ dimension. This can be clearly seen in trajectories b), c), and d). Each of these trajectories first change steeply in the $Ti(t)$ dimension, before moving to equilibrium at a tangent to the red ($\frac{dDo(t)}{dt}$) nullcline. This general property of the bi-dimensional system is again related to the much stronger auto-effect of $Ti(t)$, and the relatively small cross-effects. In a technical sense we can say that that $Do(t)$ represents the 'slowest eigen-direction' (see Strogatz, 2015, Chapter 5).

3.4.4 Inspecting the Lagged Parameters

Another way to gain insight into the processes of interest is by determining the relationships between lagged positions of the system, according to our drift matrix. To this end, we can use Equation (3.4) to determine $\mathbf{\Phi}(\Delta t)$ for some Δt. For instance, we can see that the auto-regressive and cross-lagged relationships between values of Down and Tired given $\Delta t = 1$ are

$$\mathbf{\Phi}(\Delta t = 1) = \begin{bmatrix} 0.396 & 0.117 \\ 0.077 & 0.106 \end{bmatrix}. \tag{3.10}$$

For this given time-interval, the cross-lagged effect of Down on Tired ($\phi_{21}(\Delta t = 1) = 0.077$) is smaller than the cross-lagged effect of Tired on Down ($\phi_{12}(\Delta t = 1) = 0.117$). However, as shown in Equation (3.5) the value of each of these lagged effects changes in a non-linear way depending on the time-interval chosen. To visualize this, we can calculate $\mathbf{\Phi}(\Delta t)$ for a range of Δt, and represent this information graphically in a lagged parameter plot, as in Figure 3.3.

From Figure 3.3, we can see that both cross-lagged effects reach their maximum (and have their maximum difference) at a time-interval of $\Delta t = 0.65$; furthermore, we can see that the greater cross-effect (a_{12}) results in a stronger cross-lagged effect $\phi_{12}(\Delta t)$ for a range of Δt. Moreover, we can visually inspect how the size of each of the effects of interest, as well as the difference between these effects, varies according to the time-interval. From a substantive viewpoint, we could say that the effect of feeling physically tired has the strongest effect on feelings of being down around 40 minutes later.

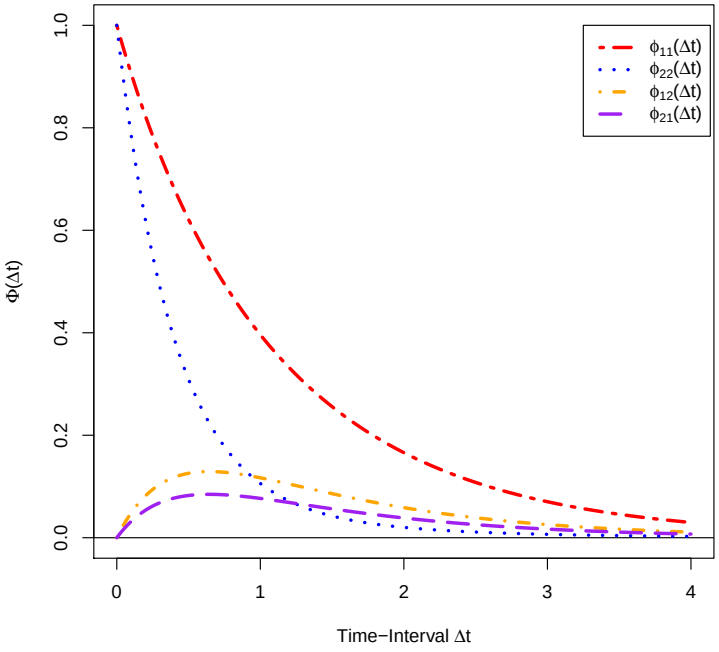


Figure 3.3: The elements of $\mathbf{\Phi}(\Delta t)$ for the bivariate example (i.e., $\phi_{11}(\Delta t)$, $\phi_{12}(\Delta t)$, $\phi_{21}(\Delta t)$, $\phi_{22}(\Delta t)$) plotted for a range of values for Δt.

While the shape of the lagged parameters may appear similar to the shapes of the trajectories plotted in the IRFs, lagged parameter plots and IRFs represent substantively different information. IRFs plot the positions of each variable in the system as they change over time, given some impulse ($\boldsymbol{y}(t)$ vs t given some $\boldsymbol{y}(0)$). In contrast, lagged parameter plots show how the lagged relationships change depending on the length of the time-interval between them, independent of impulse values ($\boldsymbol{e}^{\boldsymbol{A}\Delta t}$ vs Δt). The lagged relationships can be thought of as the components which go into determining any specific trajectories.

3.4.5 Caution With Interpreting Estimated Parameters

It is important to note that in the above interpretation of CT models, we have treated the matrix $\boldsymbol{A}$ as known. In practice of course researchers should take account of the uncertainty in parameter estimates. For example, the *ctsem* package also provides lagged parameter plots with credible intervals to account for this uncertainty.

Furthermore, researchers should be cautious about extrapolating beyond the data. For instance, when we consider a vector field, we should be careful about

interpreting regions in which there is little or no observed data (cf. Boker & McArdle, 1995). The same logic applies for the interpretation of IRFs for impulses that do not match observed values. Moreover, we should also be aware that interpreting lagged parameter plots for time-intervals much shorter than those we observe data at is a form of extrapolation: It relies on strong model-based assumptions, such as ruling out the possibility of a high-frequency higher-order process (Voelkle & Oud, 2013; Voelkle et al., 2012).

3.5 Discussion

In this chapter we have set out to clarify the connection between DT- and CT-VAR(1) models, and how we can interpret and represent the results from these models. So far we have focused on single-subject, two-dimensional, first-order systems with a stable node equilibrium. However, there are many ways in which these models can be extended, to match more complicated data and/or dynamic behavior. Below we consider three such extensions: a) systems with more than two dimensions (i.e., variables); b) different types of fixed points resulting from non-real eigenvalues of the drift matrix; and c) moving from single-subject to multi-level datasets.

3.5.1 Beyond Two-Dimensional Systems

In the empirical illustration, we examined the interpretation of a drift matrix in the context of a bivariate CT-VAR(1) model. Notably, the current trend in applications of DT-VAR(1) models in psychology has been to focus more and more on the analysis of large systems of variables, as typified, for example, by the dynamic network approach of Bringmann et al. (2013, 2016). The complexity of these models grows rapidly as the number of variables is added: To estimate a full drift matrix for a system of three variables, we must estimate nine unique parameters, in contrast to four drift matrix parameters for a bivariate system. In addition, we must estimate a three-by-three covariance matrix for the residuals, rather than a two-by-two matrix.

The relationship between the elements of $\boldsymbol{A}$ and $\boldsymbol{\Phi}(\Delta t)$ becomes even less intuitive once the interest is in a system of three variables, because the lagged parameter values are dependent on the drift matrix as a whole, as explained earlier. This means that both the relative sizes, as well as the signs of the cross-lagged effects may differ depending on the interval: The same lagged parameter may be negative for some time-intervals and positive for others, and zero-elements of $\boldsymbol{A}$ can result in corresponding non-zero elements of $\boldsymbol{\Phi}$ (cf. Deboeck & Preacher, 2016; Aalen, Røysland, Gran, & Ledergerber, 2012; Aalen et al., 2016; Aalen, Gran, Røysland, Stensrud, & Strohmaier, 2018). Therefore, although we saw in our bivariate example that, for instance, negative CT cross-effects resulted in negative DT cross-lagged effects, this does not necessarily hold in the general case (Kuiper & Ryan, 2018).

Additionally, substantive interpretation of the lagged parameters in systems

with more than two variables also becomes less straightforward. For example, Deboeck and Preacher (2016), Aalen et al. (2012, 2016) and Aalen et al. (2018) argue that the interpretation of $\mathbf{\Phi}(\Delta t)$ parameters in mediation models (with three variables and a triangular $\boldsymbol{A}$ matrix) as direct effects may be misleading: Deboeck and Preacher argue that instead they should be interpreted as *total* effects. This has major consequences for the practice of DT analyses and the interpretation of its results.

3.5.2 Complex and Positive Eigenvalues

The empirical illustration is characterized by a system with negative, real, non-equal eigenvalues, which implies that the fixed point in the system is a *stable node*. In theory, however, there is no reason that psychological processes must adhere to this type of dynamic behavior. We can apply the tools we have defined already to understand the types of behavior that might be described by other types of drift matrices. Notably, some systems may have drift matrices with complex eigenvalues, that is, eigenvalues of the form $\alpha \pm \omega i$, where $i = \sqrt{-1}$ is the imaginary number, $\omega \neq 0$, α is referred to as the real part, and ωi as the imaginary part of the eigenvalue. If the real component of these eigenvalues is negative ($\alpha < 0$), then the system is still stable, and given a deviation it will return eventually to a resting state at equilibrium. However, unlike the systems we have described before, these types of systems spiral or oscillate around the equilibrium point, before eventually coming to rest. Such systems have been described as *stable spirals*, or alternatively as *damped* (linear or harmonic) *oscillators* (Voelkle & Oud, 2013; Boker, Montpetit, et al., 2010).

A vector field for a process which exhibits this type of stable spiral behavior is shown in Figure 3.4, with accompanying trajectories. The drift matrix corresponding to this vector field is

$$\boldsymbol{A} = \begin{bmatrix} -0.995 & 0.573 \\ -2.000 & -2.416 \end{bmatrix} \tag{3.11}$$

which is equivalent to our empirical example above, but with the value of a_{21} altered from 0.375 to −2.000. The eigenvalues of this matrix are $\lambda_1 = -1.706 + 0.800i$ and $\lambda_2 = -1.706 - 0.800i$. In contrast to our empirical example, we can see that the trajectories follow a spiral pattern; the trajectory which starts at $y_1(t) = 1, y_2(t) = 1$ actually overshoots the equilibrium in the $Ti(t)$ dimension before spiraling back once in the bottom quadrant. There are numerous examples of psychological systems that are modeled as damped linear oscillators using second-order differential equations, which include the first- and second-order derivatives (cf. Boker & Nesselroade, 2002; Bisconti et al., 2004; Boker, Montpetit, et al., 2010; Horn, Strachan, & Turkheimer, 2015). However, as shown here, such behavior may also result from a first-order model.

Stable nodes and spirals can be considered the two major types of stable fixed points, as they occur whenever the real part of the eigenvalues of $\boldsymbol{A}$ are negative, that is $\alpha < 0$. Many other types of stable fixed points can be considered as special cases: when we have real, negative eigenvalues that are exactly equal, the

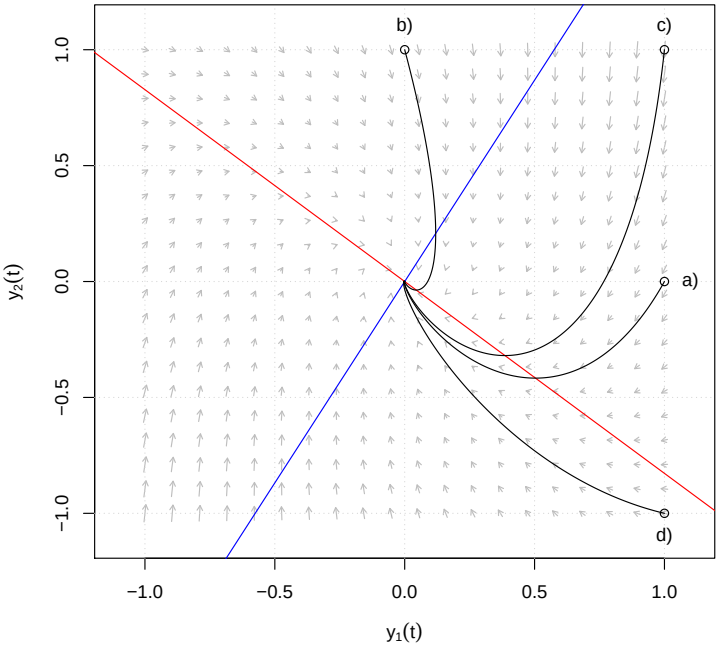


Figure 3.4: Vector field for a stable spiral corresponding to a drift matrix with negative real part complex eigenvalues.

fixed point is called a *stable star node* (if the eigenvectors are distinct), or a *stable degenerate node* (if the eigenvectors are not distinct). In contrast, if the real-part of the eigenvalues of $\boldsymbol{A}$ are positive then the system is unstable, also referred to as non-stationary or a unit-root in the time series literature (Hamilton, 1994). This implies that, given a deviation, the system will not return to equilibrium; in contrast to stable systems, in which trajectories are *attracted* to the fixed point, the trajectories of unstable systems are *repelled* by the fixed point. As such we can also encounter unstable nodes, spirals, star nodes and degenerate nodes. The estimation and interpretation of unstable systems in psychology may be fruitful ground for further research.

Two further types of fixed points may be of interest to researchers; in the special case where the eigenvalues of $\boldsymbol{A}$ have an imaginary part and no real part ($\alpha = 0$), the fixed point is called a *center*. In a system with a center fixed point, trajectories spiral around the fixed point without ever reaching it. Such systems exhibit oscillating behavior, but without any damping of oscillations; certain biological systems, such as the circadian rhythm, can be modeled as a dynamic system with a center fixed point. Such systems are on the borderline between stable and unstable systems, sometimes referred to as *neutrally stable*; trajectories

are neither attracted to or repelled by the fixed point. Finally, a *saddle point* occurs when the eigenvalues of $\boldsymbol{A}$ are real but of opposite sign (one negative, one positive). Saddle points have one stable and one unstable component; only trajectories which start exactly on the stable axis return to equilibrium, and all others do not. Together spirals, nodes and saddle points cover the majority of the space of possible eigenvalues for $\boldsymbol{A}$. Strogatz (2015) describes the different dynamic behavior generated by different combinations of eigenvalues of $\boldsymbol{A}$ in greater detail.

3.5.3 Multilevel Extensions

The time series literature (such as from the field of econometrics) as well as the dynamic systems literature (such as from the field of physics) tends to be concerned with a single dynamic system, either because there is only one case ($N = 1$), or because all cases are exact replicates (e.g., molecules). In psychology however, we typically have data from more than one person, and we also know that people tend to be highly different. Hence, when we are interested in modeling their longitudinal data, we should take their differences into account somehow. The degree to which this can be done, depends on the number of time points we have per person. In traditional panel data, we typically have between two and six waves of data. In this case, we should allow for individual differences in means or intercepts, in order to separate the between-person, stable differences from the within-person dynamic process, while assuming the lagged relationships are the same across individuals (cf. Hamaker et al., 2015).

In contrast, experience sampling data and other forms of intensive longitudinal data consist of many repeated measurement per person, such that we can allow for individual differences in the lagged coefficients. This can be done by either analyzing the data of each person separately, or by using a dynamic multilevel model in which the individuals are allowed to have different parameters (cf. Driver et al., 2017; Boker, Staples, & Hu, 2016). Many recent studies have shown that there are substantial individual differences in the dynamics of psychological phenomena, and that these differences can be meaningfully related to other person characteristics, such as personality traits, gender, age, and depressive symptomatology, but also to later health outcomes and psychological well-being (e.g., Kuppens et al., 2012, 2010; Bringmann et al., 2013).

While the current chapter has focused on elucidating the interpretation of a single-subject CT-VAR(1) model, the substantive interpretations and visualization tools we describe here can be applied in a straightforward manner to, for example, the fixed effects estimated in a multilevel CT-VAR(1) model, or to individual-specific parameters estimated in a multilevel framework. The latter would however, lead to an overwhelming amount of visual information. The development of new ways of summarizing the individual differences in dynamics, based on the current tools, is a promising area for future research.

3.5.4 Conclusion

There is no doubt that the development of dynamical systems modeling in the field of psychology has been hampered by the difficulty in obtaining suitable data to model such systems. However this is a barrier that recent advances in technology will shatter in the coming years. Along with this new source of psychological data, new psychological theories are beginning to emerge, based on the notion of psychological processes as dynamic systems. Although the statistical models needed to investigate these theories may seem exotic or difficult to interpret at first, they reflect the simple intuitive and empirical notions we have about psychological processes: Human behavior, emotion and cognition fluctuate continuously over time, and the models we use should reflect that. We hope that our treatment of CT-VAR(1) models and their interpretation will help researchers to overcome the knowledge-barrier to this approach, and can serve as a stepping stone towards a broader adaptation of the CT dynamical system approach to psychology.

Appendix 3.A Matrix Exponential

Similarly to the scalar exponential, the matrix exponential can be defined as an infinite sum

$$\boldsymbol{e}^{\boldsymbol{A}} = \sum_{k=0}^{\infty} \frac{1}{k!} \boldsymbol{A}^k$$

The exponential of a matrix is not equivalent to taking the scalar exponential of each element of the matrix, unless that matrix is diagonal. The exponential of a matrix can be found using an eigenvalue decomposition

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^{-1}$$

where $\boldsymbol{V}$ is a matrix of eigenvectors of $\boldsymbol{A}$ and $\boldsymbol{D}$ is a diagonal matrix of the eigenvalues of $\boldsymbol{A}$ (cf. Moler & Van Loan, 2003). The matrix exponential of $\boldsymbol{A}$ is given by

$$\boldsymbol{e}^{\boldsymbol{A}} = \boldsymbol{V}\boldsymbol{e}^{\boldsymbol{D}}\boldsymbol{V}^{-1}$$

where $\boldsymbol{e}^{\boldsymbol{D}}$ is the diagonal matrix whose entries are the scalar exponential of the eigenvalues of $\boldsymbol{A}$. When we want to solve for the matrix exponential of a matrix multiplied by some constant Δt we get

$$e^{\boldsymbol{A}\Delta t} = \boldsymbol{V}e^{\boldsymbol{D}\Delta t}\boldsymbol{V}^{-1} \tag{3.12}$$

Take it that we have a 2×2 square matrix given by

$$\boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

and we wish to solve for $e^{\boldsymbol{A}\Delta t}$. The eigenvalues of $\boldsymbol{A}$ are given by

$$\lambda_1 = \frac{1}{2}(a + d - \sqrt{a^2 + 4bc - 2ad + d^2})$$
$$\lambda_2 = \frac{1}{2}(a + d + \sqrt{a^2 + 4bc - 2ad + d^2})$$

where we will from here on denote

$$R = \sqrt{a^2 + 4bc - 2ad + d^2}$$

for notational simplicity. The exponential of the diagonal matrix made up of eigenvalues, multiplied by the constant Δt is given by

$$\boldsymbol{e}^{\boldsymbol{D}\Delta t} = \begin{bmatrix} e^{\frac{1}{2}(a+d-R)\Delta t} & 0 \\ 0 & e^{\frac{1}{2}(a+d+R)\Delta t} \end{bmatrix}$$

The matrix of eigenvectors of $\boldsymbol{A}$ is given by

$$\boldsymbol{V} = \begin{bmatrix} \frac{a-d-R}{2c} & \frac{a-d+R}{2c} \\ 1 & 1 \end{bmatrix}$$

assuming $c \neq 0$, with inverse

$$\boldsymbol{V}^{-1} = \begin{bmatrix} \frac{-c}{R} & \frac{a-d+R}{2R} \\ \frac{c}{R} & \frac{-a+d+R}{2R} \end{bmatrix}.$$

Multiplying $\boldsymbol{V} e^{\boldsymbol{D}} \boldsymbol{V}^{-1}$ gives us

$$e^{\boldsymbol{A}\Delta t} = \begin{bmatrix} \frac{R-a+d}{2R} e^{\lambda_1 \Delta t} + \frac{R+a-d}{2R} e^{\lambda_2 \Delta t} & \frac{b(-e^{\lambda_1 \Delta t} + e^{\lambda_2 \Delta t})}{R} \\ \frac{c(-e^{\lambda_1 \Delta t} + e^{\lambda_2 \Delta t})}{R} & \frac{R+a-d}{2R} e^{\lambda_1 \Delta t} + \frac{R-a+d}{2R} e^{\lambda_2 \Delta t} \end{bmatrix} \tag{3.13}$$

Note that we present here only a worked out example for a 2×2 square matrix. For larger square matrices (representing models with more variables), the eigenvalue decomposition remains the same although the terms for the eigenvalues, eigenvectors and determinants become much less feasible to present.

Appendix 3.B Empirical Example Data Analysis

```
## Data Preparation
# Load the ctsem package
library(ctsem)

# Data is available form https://osf.io/c6xt4/ #
setwd("./ESMdata")

#Load Data#
rawdata <- read.csv("ESMdata.csv",
           header=TRUE,
           stringsAsFactors = FALSE)

# Select only measurements which take place in
# the control and initial (no medication reduction) phase
rawdata <- subset(rawdata,rawdata$phase==1|rawdata$phase==2)

# Select only the variables of interest
sel <- c("mood_down","phy_tired")
data <- rawdata[,(names(rawdata) %in% sel)]

# Standardize the selected variables
for(j in 1:dim(data)[2]){
  data[,j] <- (data[,j]-mean(data[,j]))/sd(data[,j])
}

# Create a time vector representing hours since first measurement
# Required by ctsem function ctIntervalise
t1 <- as.POSIXct(paste(rawdata$date, rawdata$resptime_s),
                                format="%d/%m/%y %H:%M:%S")
```

```
time <- as.numeric(difftime(t1,t1[1], units="hours"))

# Attach this time variable to the selected items
data$time = time

# Create an ID variable
data$id = rep(1,dim(data)[1])

# Rename pat_agitate = Y1 -- event_import = Y2
colnames(data) = c("Y1", "Y2", "time", "id")

# Get data in wide format for ctsem
datawide <- ctLongToWide(datalong = data, id = "id",
                         time = "time",
                         manifestNames = c("Y1","Y2"))

# Create time-interval variable
datawide <- ctIntervalise(datawide = datawide,
                          Tpoints = dim(data)[1],
                          n.manifest = 2,
                          manifestNames = c("Y1","Y2"))


## Data analysis

# First specify the bivariate model, with 2 observed variables
model <- ctModel(n.manifest = 2, n.latent = 2,
                 Tpoints = 286,
                 LAMBDA = diag(nrow = 2),
                 MANIFESTMEANS = matrix(data = 0, nrow = 2, ncol = 1),
                 MANIFESTVAR = matrix(data = 0, nrow = 2, ncol = 2),
                 DRIFT = "auto",
                 CINT = matrix(data = 0, nrow = 2, ncol = 1),
                 DIFFUSION = "auto",
                 TRAITVAR = NULL,
                 MANIFESTTRAITVAR = NULL,
                 startValues = NULL)

# Fit the model to the data
fit <- ctFit(data = datawide,
             ctmodelobj = model,
             objective = "Kalman",
             stationary = c("TOVAR", "TOMEANS"),
             iterationSummary = T,
             carefulFit = T, showInits = F,
             asymptotes = F,
```

```
            meanIntervals = F,
            plotOptimization = F,
            nofit = F, discreteTime = F,
            verbose = 0)

summary(fit)
```

Chapter 4

Time to Intervene: A Continuous-Time Approach to Network Analysis and Centrality

Abstract

Dynamical network analysis of experience sampling data has become increasingly popular in clinical psychology. In this approach, discrete-time (DT) VAR model parameters are interpreted as direct relationships between psychological processes, and centrality measures are used to identify which variable should be targeted for an intervention. There are two methodological problems with this practice. First, VAR models suffer from time-interval dependency, yielding different conclusions based on an often arbitrary choice of sampling scheme. Second, the exact link between centrality measures and intervention effects is unclear in this context and relies on intuitive interpretations of model parameters that may not be valid. In this paper we address both issues by proposing a continuous-time (CT) approach to network analysis. We make use of CT-VAR models, which allow one to explicitly model time-interval dependency. To aid intervention targeting we develop new centrality measures based on CT networks. We formulate these measures within the interventionist causal inference framework, and show how these measures can be used to identify optimal targets for either an acute or continuous intervention.

This chapter has been adapted from: Ryan, O. & Hamaker, E. L. (under review). Time to Intervene: A Continuous-Time Approach to Network Analysis and Centrality. Author contributions: OR and ELH conceptualized the initial project. OR wrote the paper and *R* code. ELH helped further develop the ideas in the project, discussed progress and provided textual feedback.

4.1 Introduction

Dynamical network analysis, based on lagged regression models, has become a popular approach for the analysis of experience sampling data in psychology (Bringmann et al., 2013; Borsboom & Cramer, 2013). In clinical psychology in particular, such analyses have been promoted as an aid in developing personalized treatments for psychopathology. To facilitate this, *centrality measures* calculated from parameter estimates are often used to identify which variable in the network to *target for an intervention* (Bringmann et al., 2013; A. J. Fisher & Boswell, 2016; Kroeze et al., 2017; Epskamp, van Borkulo, et al., 2018; Rubel, Fisher, Husen, & Lutz, 2018; Bak, Drukker, Hasmi, & van Os, 2016; Bringmann et al., 2015; Bastiaansen et al., 2019; A. J. Fisher, Reeves, Lawyer, Medaglia, & Rubel, 2017; Christian et al., 2019).

Two developments in the methodological literature highlight problems with this practice. First, the estimation of dynamical network structures is based on path estimates from a discrete-time (DT) first-order Vector Auto-regressive (VAR) model. The DT-VAR model suffers from the problem of *time-interval dependency*, meaning that the parameters of this model can potentially lead to dramatically different conclusions based on how the observations are spaced in time (Gollob & Reichardt, 1987; Kuiper & Ryan, 2018). This is an issue which numerous authors have argued could be resolved by modeling psychological processes as unfolding *continuously* over time using continuous-time (CT) models (Boker, 2002; van Montfort, Oud, & Voelkle, 2018; Ryan, Kuiper, & Hamaker, 2018). Second, the use of out-of-the-box centrality measures for the identification of intervention targets has been critiqued from various sides, both in terms of their applicability to psychological networks (Bringmann et al., 2019) and in their actual ability to detect optimal intervention targets (Borgatti, 2005; van Elteren & Quax, 2019; Dablander & Hinne, 2019).

Although the consequences of taking a CT approach have been discussed in general elsewhere (cf. Boker, 2002; Voelkle et al., 2012; Aalen et al., 2016, 2012; Deboeck & Preacher, 2016), the specific implications for dynamical network analysis have not yet been investigated. Taking a CT approach yields practical benefits, such as an ability to deal with unequally spaced measurements, but also leads to a new outlook on the meaning and interpretation of lagged regression parameters. As a consequence, taking a CT perspective not only undermines the typical interpretation of DT-VAR networks as well as their accompanying centrality measures, but it also creates opportunities for the development of new ways to understand the underlying dynamic process and how best to intervene on it.

In this paper we introduce a CT approach to dynamical network analysis, and develop new centrality measures which can be used to gain more insight into what variables should be targeted for interventions, and what kinds of interventions can be expected to lead to what kinds of outcome. In so doing we tie together diverse strands of the psychological and methodological literature. Specifically, we show how centrality measures are related to path-specific effects from the SEM literature, which allows us to connect previous research on path-specific effects in CT models (Aalen et al., 2012; Deboeck & Preacher, 2016) and

intervention-based approaches to causal modeling (Eichler & Didelez, 2010; VanderWeele, 2015; Pearl, 2009).

This paper is organized as follows. In the first part of the paper, we provide an overview of the DT-VAR model, how path-specific effects and centrality measures are used to identify intervention targets based on this model in practice, and the problems that arise therein. Second, we present the CT approach to dynamic network analysis based on the CT-VAR model, introducing a new network representation into the psychological canon from the causal inference literature, and exploring the new insights this approach yields. Third, we introduce new fit-for-purpose centrality measures which both reflect the CT nature of the underlying process and have a clear and direct link to interventions and the choice of optimal intervention targets. Throughout, we use a hypothetical example for illustrative purposes. Moreover, for simplicity, the developments on this paper focus on single subject models, though the critiques and measures developed here generalize in a straightforward way to within-subjects parameters of multi-level models.

4.2 Current Practice: DT-VAR Networks

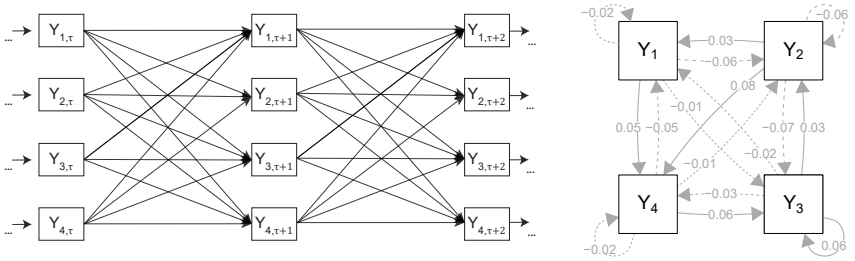
In this section we discuss the first-order discrete-time Vector Auto-Regressive (DT-VAR) model and its representation as a dynamical network. Subsequently, we describe how path-specific effects and centrality measures based on this model are used by applied researchers to identify intervention targets, and establish how these two practices are connected. We end by discussing two problems with current practice, namely the lack of clarity regarding how centrality measures relate to the effects of interventions (i.e., *the intervention problem*) and the time-interval dependency of parameter estimates (i.e., *the time-interval problem*).

4.2.1 The DT-VAR model

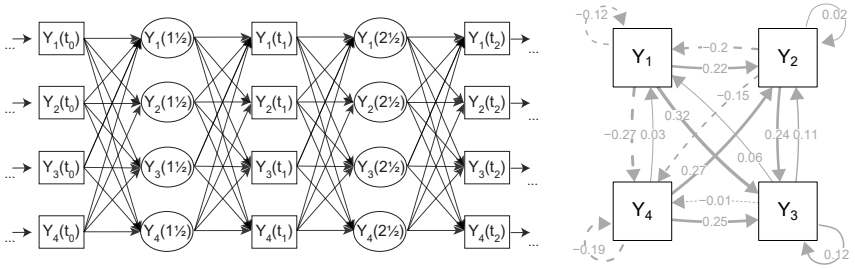
The DT-VAR model is a single-subject time-series model that describes dynamic relationships between variables measured repeatedly over time. Lagged regression parameters encode the effect of a variable on itself (an auto-regressive effect) or another variable (a cross-lagged effect) at the next measurement occasion (i.e., at a lag of one). This model can be written as

$$\boldsymbol{Y}_\tau = \boldsymbol{c} + \boldsymbol{\Phi}\boldsymbol{Y}_{\tau-1} + \boldsymbol{\epsilon}_\tau \tag{4.1}$$

where, given p variables, the $p \times 1$ vector of random variables $\boldsymbol{Y}$ at occasion τ is regressed on the $p \times 1$ vector of those same variables at the previous occasion, $\boldsymbol{Y}_{\tau-1}$. The $p \times p$ matrix of lagged regression parameters is denoted $\boldsymbol{\Phi}$, while the $p \times 1$ vectors $\boldsymbol{c}$ and $\boldsymbol{\epsilon}_\tau$ denote the intercepts and random shocks respectively, the latter being normally distributed with mean zero and variance-covariance matrix $\boldsymbol{\Psi}$ (Hamilton, 1994). Two crucial assumptions of the DT-VAR model are, first, that the same amount of time (denoted Δt) elapses between two subsequent measurement occasions, and second, that the underlying process is stationary,



(a) DT-VAR path-model and dynamical network representing $\mathbf{\Phi}(1)$



(b) DT-VAR model with latent waves and dynamical network representing $\mathbf{\Phi}(\frac{1}{2})$

Figure 4.1: Path-model (left-hand side) and network (right-hand side) representations of two four-variable DT-VAR models. In the path models, the presence of an arrow linking two variables denotes some non-zero dependency between them, conditional on all variables at the previous wave. For the networks solid arrows denote positive parameters and dashed arrows represent negative parameters.

which entails that the means, variance and covariances, and lagged regression parameters remain the same over time.

The DT-VAR model describes a relatively simple dynamic system. Each variable has some baseline level or equilibrium position defined by its mean value. Typically, the variables themselves are centered around this mean value (e.g. Asparouhov, Hamaker, & Muthén, 2018), and so the intercept term is often omitted for notational simplicity ($\boldsymbol{c} = \mathbf{0}$), a convention we will adopt throughout the remainder of the paper. The random shocks $\boldsymbol{\epsilon}_\tau$ push the system away from equilibrium, and the lagged parameters $\mathbf{\Phi}$ determine how the variables react to these shocks, eventually returning to equilibrium over time.

The DT-VAR model can be represented as either a path-model, as shown in the left-hand panel of Figure 4.1(a), or as a dynamical network structure, as shown in the right-hand panel, where the nodes represent the random variables, and the edges represent the values of the lagged parameters $\mathbf{\Phi}$ (Bringmann et al., 2013; Epskamp, van Borkulo, et al., 2018). The DT-VAR model, and its multi-level extension, are popular for the analysis of experience sampling data in clinical psychology (Bringmann et al., 2013; Pe et al., 2015; A. J. Fisher & Boswell, 2016;

Kroeze et al., 2017; Rubel et al., 2018; Bak et al., 2016).[1] In dynamical network approaches, each variable is typically used to represent some psychological construct (such as an emotion or symptom of disorder) that varies over time, which we refer to as a process. The lagged parameters in $\mathbf{\Phi}$ are typically interpreted as *direct effects* of these processes on each other over time.

For example, take it that the four variables in Figure 4.1(a) represent (repeated measurements of) *Stress* (Y_1), *Anxiety* (Y_2), feeling *Self-Conscious* (Y_3) and feelings of *Physical Discomfort* (Y_4). We will refer to this throughout as the *Stress-Discomfort* system. We can see from the parameter values in the dynamical network that all variables share reciprocal cross-lagged relationships with all other variables: The network is fully connected, with a mix of positive and negative cross-lagged and auto-regressive parameters of different sizes. Typically, a cross-lagged parameter such as $\phi_{41} = 0.05$ would be interpreted as the direct effect of current Stress ($Y_{1,\tau}$) on Physical Discomfort at the next measurement occasion ($Y_{4,\tau+1}$), conditional on (i.e., controlling for) current feelings of Anxiety, Self-Conscious, and Physical Discomfort ($Y_{2,\tau}, Y_{3,\tau}, Y_{4,\tau}$). This parameter is weakly positive, leading to the interpretation that a high level of current Stress has a small positive direct effect on feelings of Physical Discomfort at the next occasion.

4.2.2 Intervention Targets from DT-VAR models

To identify which variables should be considered targets for an intervention based on a DT-VAR model, psychology researchers have mainly used two approaches: a) *path-specific effects*, which are inspired by the SEM literature (Bollen, 1987); and b) *centrality measures*, which come from the broader network analysis literature (Freeman, 1978; Opsahl, Agneessens, & Skvoretz, 2010).

4.2.2.1 Path-Specific Effects

Path-specific effects have been used to describe the *total*, *direct* and *indirect* effects of one variable on another, and are calculated using the well-known *path-tracing rules* from the SEM literature (Bollen, 1987). The total effect of Stress levels now ($Y_{1,\tau}$) on Physical Discomfort two measurement occasions later ($Y_{4,\tau+2}$) is the sum of direct effect pathways ($Y_{1,\tau} \rightarrow Y_{4,\tau+1} \rightarrow Y_{4,\tau+2}$ and $Y_{1,\tau} \rightarrow Y_{1,\tau+1} \rightarrow Y_{4,\tau+2}$) and indirect effect pathways through the mediating variables Anxiety ($Y_{1,\tau} \rightarrow Y_{2,\tau+1} \rightarrow Y_{4,\tau+2}$) and Self-Conscious ($Y_{1,\tau} \rightarrow Y_{3,\tau+1} \rightarrow Y_{4,\tau+2}$) (Cole & Maxwell, 2003). Note here that to define a path-tracing effect between variables Y_1 and Y_4 we must also specify the number of measurement occasions that elapse between the relevant instances of those variables, that is, the *lag* of interest.

[1] Note that in many applications, the multi-level extension of the VAR(1) model is used, for the purposes of regularizing the parameter estimates and modeling between-person differences in parameters. The primary interest of researchers using these multi-level models is in the individual within-person parameter estimates (Bringmann et al., 2013; Schuurman et al., 2016; Asparouhov et al., 2018; Suls, Green, & Hillis, 1998; Lodewyckx, Tuerlinckx, Kuppens, Allen, & Sheeber, 2011; Liu, Kuppens, & Bringmann, 2019). For the sake of notational simplicity —and without loss of generality— we focus on the single-subject version of the VAR(1) model throughout the paper.

If we accept the interpretation of $\mathbf{\Phi}$ parameters as direct causal effects, we may suggest that interventions should target variables that have strong direct or total effects on others, and/or those mediators through which the strongest indirect effects pass (for examples in a general SEM context, see Bernat, August, Hektner, & Bloomquist, 2007; Bramsen et al., 2013). For instance, based on the parameters in Figure 4.1(a), we might suggest Anxiety as an intervention target due to the relatively strong *lag-one direct effect* on Physical Discomfort ($\phi_{42} = .08$), or because it is a mediator of the largest *lag-two* indirect effect, from Stress to Physical Discomfort ($Y_{1,\tau} \rightarrow Y_{2,\tau+1} \rightarrow Y_{4,\tau+2} = -.005$).

4.2.2.2 Centrality Measures

Researchers adopting a network approach have used *centrality measures* to determine which variable should be considered the optimal intervention target in a network (e.g., Bringmann et al., 2013; A. J. Fisher & Boswell, 2016; Kroeze et al., 2017; Epskamp, van Borkulo, et al., 2018; Rubel et al., 2018; Bak et al., 2016; Bringmann et al., 2015; Bastiaansen et al., 2019). Centrality measures are used to summarize the relationships a particular variable has with the network as a whole, typically summing over the individual relationships that variable shares with all others in the network. To our knowledge, the precise connection between path-specific effects and centrality measures for DT-VAR models has not yet been described in the literature. However, a close inspection of the computation and interpretation of many popular centrality measures reveals that they are very similar to path-tracing effects: Specifically, many centrality measures are interpreted as capturing either *total*, *direct* and *indirect* effects, and in turn these measures are often closely related to summaries of the corresponding path-specific effects. Here, we will focus on three such measures (with the exact connection between these measures and path-tracing effects given in Appendix 4.A.)

The *Two-Step Expected Influence* measure ($EI_i^{(2)}$; Robinaugh et al., 2016; Kaiser & Laireiter, 2018) is typically interpreted as a summary of *total effects*: In path-tracing terms it is the sum of lag one direct effects and lag two total effects (see Appendix 4.A for details). As such, variables with a high Two-Step Expected Influence could be expected to exert a high overall influence on the system, making it an attractive intervention target. For the Stress-Discomfort system, this measure would suggest Anxiety as an intervention target, $EI_i^{(2)} = 0.035$.

The *One-Step Expected Influence* ($EI_i^{(1)}$; Robinaugh et al., 2016; Kaiser & Laireiter, 2018) and *Out-Strength centrality* (Opsahl et al., 2010) measures are interpreted as summarizing direct effects. They are both sums of lag-one direct effects, with the latter taking the absolute value. Due to their close connection we will focus exclusively on the One-Step Expected Influence measure in this paper. For the Stress-Discomfort System, Anxiety also has the highest One-Step Expected Influence ($EI_2^{(1)} = \phi_{12} + \phi_{32} + \phi_{42} = 0.038$).

Finally, *Betweenness* is interpreted as indicating the degree to which a variable "funnels information flow", similar to how mediating variables funnel *indirect effects* (e.g. Bringmann et al., 2013; Opsahl et al., 2010; Freeman, 1977). This mea-

sure is conceptually similar to determining which variables are strong mediators, although paths are calculated by summing, rather than multiplying parameters, as in path-tracing rules. The Betweenness Centrality of Stress, Anxiety and Physical Discomfort are all equal ($BC_1 = BC_2 = BC_4 = 1$), with Self-Conscious attaining a lower score ($BC_3 = 0$), indicating that the former three are equally attractive intervention targets in terms of the role they play in indirect effects in the network.

4.2.3 Two Problems with Current Practice

From the above we can conclude that there is a clear logical thread that can be followed from the interpretation of DT-VAR parameters as direct effects through to the use of centrality measures for the identification of intervention targets. However, there are two major problems associated with the approach outlined above, both of which relate to the intuitive interpretation of lagged model parameters as direct effects. We detail these problems below, and elaborate on how to overcome these problems in the remainder of the paper.

4.2.3.1 The Intervention Problem

The first problem with current practice is that it is not clear exactly how or why particular parameters or centrality measures should be informative about the effects of interventions. The idea that a cross-lagged parameter ϕ_{ji} represents a direct effect can be defended from two perspectives. First, from a purely path-tracing point of view, if we assume that the path model in Figure 4.1(a) is the true model, then ϕ_{ji} represents the path that we can trace from $Y_{i,\tau}$ to $Y_{j,\tau+1}$ without passing through any values of other variables in the model. Second, we can use the notion of *Granger causality*, which states that, if we assume that all relevant variables have been measured (i.e. no unobserved confounding) then the parameter ϕ_{ji} represents a *direct causal effect* of $Y_{i,\tau}$ on $Y_{j,\tau+1}$ (Granger, 1969). This is a notion of causality which is often invoked in empirical applications of the DT-VAR model (cf. Cole & Maxwell, 2003; Hamaker et al., 2015; Bulteel et al., 2016).

However, what is absent from both conceptualizations of lagged parameters as direct effects is an explicit link between a particular parameter and the effect of a particular intervention. Without such a link, it is impossible to give meaningful answers to some of the most basic questions which arise regarding how path-specific effects and centrality measures should be applied in practice, such as: Should we prefer to intervene on a variable with high direct, total, or indirect relationships? What kind of intervention should we apply, and what kind of effect would we expect to see? And under what conditions do these measures inform us of this? Establishing an unambiguous link between interventions on the one hand, and parameters and centrality measures on the other, is a fundamental prerequisite if we are to use these measures to inform intervention targets in a principled way. Such a link is necessary as, despite their intuitive appeal, any given centrality measure cannot be simply assumed to be informative about intervention targets. This issue has been highlighted by, amongst others, van El-

teren and Quax (2019) and Dablander and Hinne (2019), who show that central nodes are often not the optimal target for an intervention, as well as Borgatti (2005) and more recently Bringmann et al. (2019), who detail at length how the utility of common centrality measures is highly dependent on the particular type of network under consideration.

To solve this problem, we will make use of the interventionist causal inference framework (Hernan & Robins, 2019; Pearl, 2009; Rubin, 1974). In this framework, a causal relationship between X and Y is defined as the change we would expect to see in Y if we were to intervene and change the value of X. The key appeals of this approach are first, that it places an emphasis on an intervention-effect as the target of inference, and second that it allows one to derive the conditions under which the effects of such an intervention can be identified from relationships between variables in observational data. Crucially, intervention-based definitions have also been extended to path-specific effects: VanderWeele and Tchetgen Tchetgen (2017) have shown that—under certain circumstances—path-tracing total, direct and indirect effects in a lagged regression model (similar to the DT-VAR model discussed above) can be interpreted as the *expected change* in Y, given interventions to change the values of the cause variable X and the mediating variable(s) M. Although, to our knowledge, intervention-based definitions have not been extended to centrality measures, the connection between centrality and path-specific effects established above opens up a potential avenue by which such a definition could be established. In Section 4.4.1 we will make use of these connections to establish new centrality measures which are inspired by this framework, ensuring an unambiguous connection between the value of those measures and the choice of intervention target.

Of course, the identification of these intervention effects in practice will still rely on a number of idealized assumptions. We must assume that there are no unmeasured confounder variables, that we can intervene in the system without changing how the variables in the system relate to one another (an assumption known as modularity; Pearl, 2009), and that the first-order time series model we fit to the data correctly characterizes the process. Assumptions regarding confounders and modularity are fundamental to the identification of intervention-effects from data in any context (e.g. from experiments, cross-sectional data or time-series). However, even if we are willing to make these strong and optimistic assumptions, there is still a fundamental issue specific to lagged regression models which we must account for in order to come to sensible conclusions about intervention targets. That fundamental issue is the *time-interval problem*.

4.2.3.2 The Time-Interval Problem

As pointed out by Gollob and Reichardt (1987), and since then by many others (Oud & Jansen, 2000; Reichardt, 2011; Voelkle et al., 2012; Deboeck & Preacher, 2016; Kuiper & Ryan, 2018), the effect of one variable on another depends critically on the time-interval between them. This is referred to as the *time-interval*

problem.[2] This problem arises when the processes under investigation vary at a higher frequency than we observe them, which means that the processes take on values at points in time in-between those particular occasions at which we measure the process. In path-modeling terms, this means that there are unmeasured and un-modeled values of the observed variables $\boldsymbol{Y}$ in-between measurement occasions, depicted as latent variables in Figure 4.1(b). The time-interval problem has two consequences. First, the value of the regression parameters $\boldsymbol{\Phi}$ are a function of the time-interval between observations (denoted $\boldsymbol{\Phi}(\Delta t)$), and so, except under restrictive circumstances, may change their sign, size and relative ordering if a different time-interval is used (Kuiper & Ryan, 2018). Second, the lagged regression parameters at *any* interval should be interpreted as *total*, rather than direct effects from a path-tracing perspective. The reason for this is that cross-lagged parameters constitute both direct pathways and indirect pathways through latent values of the other processes in the model (Cole & Maxwell, 2003; Reichardt, 2011; Deboeck & Preacher, 2016).

To illustrate both issues, let us take it that the parameters that we introduced in Figure 4.1(a) represent the lagged relationships at *one-hour intervals*; we denote these parameters as $\boldsymbol{\Phi}(\Delta t = 1)$. In theory, we could also have obtained observations of the Stress-Discomfort system at *half-hour intervals*: The path model representing this shorter-interval system is depicted in the left panel of Figure 4.1(b), where the potential half-hour measurements are depicted as *latent* variables $\boldsymbol{Y}(t = 1\frac{1}{2})$ and $\boldsymbol{Y}(t = 2\frac{1}{2})$. The effects matrix relating the half hour realizations of the process is denoted $\boldsymbol{\Phi}(\Delta t = \frac{1}{2})$. A well-known result from the time-series literature allows us to relate the parameters of these two models through the matrix exponent (Hamilton, 1994). In this case, the matrix of lagged parameters of the half-hour system is related to the matrix of lagged parameters at the longer one-hour time-interval by the expression

$$\boldsymbol{\Phi}(\tfrac{1}{2})^2 = \boldsymbol{\Phi}(1) \tag{4.2}$$

that is, by squaring the matrix of parameters at the shorter interval, we obtain the parameters at twice that interval. It is important to note here that squaring a matrix is not equivalent to squaring the parameters of that matrix: Instead, any given parameter in $\boldsymbol{\Phi}(1)$ is a function of multiple parameters in $\boldsymbol{\Phi}(\frac{1}{2})$. For instance, the cross-lagged parameter which regresses $Y_{4,\tau+1}$ on $Y_{1,\tau}$ can be rewritten in terms of the shorter-interval parameters as $\phi_{41}(1) = \phi_{11}(\frac{1}{2})\phi_{41}(\frac{1}{2}) + \phi_{21}(\frac{1}{2})\phi_{42}(\frac{1}{2}) + \phi_{31}(\frac{1}{2})\phi_{43}(\frac{1}{2}) + \phi_{41}(\frac{1}{2})\phi_{44}(\frac{1}{2})$. Clearly, this has major consequences for the interpretation of lagged regression parameters.

First, we can see that the types of conclusions we would intuitively make based on the parameter values in the half-hour network are entirely different than those we made based on the parameters in the one-hour network. To be-

[2] This is also sometimes referred to as the *lag* problem (e.g., Hamaker & Wichers, 2017). The *lag* of a model also describes a notion of time-spacing, but more generally refers to the order of a lagged regression model: Lag-one models relate current values to the exactly preceding measurement occasion, and lag-two models relate current values to the variables two measurement occasions previously, and so forth. Thus, the lag of a model and is distinct from the time-interval between observations, and so for clarity we refer to this as the time-interval problem throughout.

gin with, the magnitude of each individual parameter, as well as their signs and relative ordering, are different in the two networks. For example, in the one-hour network, Stress and Anxiety both have positive lagged relationships with Physical Discomfort ($\phi_{41}(1) = 0.047$ and $\phi_{42}(1) = 0.077$), with the effect of Anxiety being slightly larger. In the half-hour network, the corresponding lagged relationships are both strongly negative, with the effect of Stress now the larger ($\phi_{41}(\frac{1}{2}) = -0.275$ and $\phi_{42}(\frac{1}{2}) = -0.151$). If we calculate centrality measures for the half-hour network, we may conclude that either Physical Discomfort is a promising intervention target, as it has the highest scores on the total and direct influence measures ($EI_4^{(2)} = 0.555$ and $EI_4^{(1)} = 0.557$), or that Anxiety is a promising target as it has the highest Betweenness Centrality ($BC_2 = 3$; See Appendix 4.B for a full table of centrality values).

Second, on a more fundamental level, the presence of the time-interval problem alters how we should *interpret* a set of lagged regression parameters. Specifically, based on the relationship in Equation (4.2), the lagged parameters of the one-hour network $\mathbf{\Phi}(1)$ should be interpreted as *total* rather than *direct effects* (Deboeck & Preacher, 2016; Aalen et al., 2016). Indeed, taking the power of a matrix of direct effects, as in Equation (4.2), is suggested by Bollen (1987) exactly as a method of calculating total effects in the SEM literature. This interpretation also helps clarify how we come to seemingly contradictory conclusions based on the time-interval used. For example, the parameter $\phi_{42}(1) = 0.077$ would typically be interpreted as a direct effect of Anxiety now ($Y_{2,\tau}$) on Physical Discomfort an hour from now ($Y_{4,\tau+1}$). However, when we examine how these variables are related to one another in Figure 4.1(b), we can see that the relationship between them is made up of a number of different pathways, including direct paths ($Y_{2,\tau} \rightarrow Y_2(1\frac{1}{2}) \rightarrow Y_{4,\tau+1}$ and $Y_{2,\tau} \rightarrow Y_4(1\frac{1}{2}) \rightarrow Y_{4,\tau+1}$) as well as paths that pass through latent values of Stress ($Y_{2,\tau} \rightarrow Y_1(1\frac{1}{2}) \rightarrow Y_{4,\tau+1}$) and Self-Conscious ($Y_{2,\tau} \rightarrow Y_3(1\frac{1}{2}) \rightarrow Y_{4,\tau+1}$). In this instance, a combination of strong positive indirect paths and slightly weaker negative direct paths add up to the weak positive total effect in the one-hour network: A high feeling of Anxiety has the effect of lowering Physical Discomfort a half hour later, but also sets off a chain reaction of Stress and Self-Conscious feelings that ultimately leads to an increase in Physical Discomfort an hour later.

In summary, the time-interval problem is present whenever there are latent, unobserved values of our processes of interest in-between measurement occasions. The consequences of the time-interval problem for the identification of intervention targets, using current practice, is twofold. First, any parameter, path-specific effect, or centrality measure may lead to conclusions which are specific to the time-interval used in data collection: *The effect* of X on Y should in fact be interpreted as *the effect at one specific time-interval* (Gollob & Reichardt, 1987). Second, the intuitive interpretation given to path-specific effects and centrality measures based on a path-tracing logic are misleading: Direct effects are better conceptualized as *total effects* (Deboeck & Preacher, 2016; Reichardt, 2011). This leaves the usual path-tracing based notion of "direct", "indirect" and "total" effects, and therefore also of centrality measures which summarize these effects,

open to question. However, our presentation here also highlighted one potential solution to the time-interval problem: Decomposing lagged relationships between observations into truly direct effects operating over a shorter time-interval. This decomposition opens up a new perspective on how lagged relationships should be interpreted, a perspective which we can use to explore time-interval dependency, and avoid coming to misleading or contradictory choices regarding intervention targets.

4.2.4 Conclusion

We have now identified two problems with current DT-VAR network approaches, in particular with the use of these models and associated centrality measures for the identification of intervention targets. The first is that the link from interventions to model parameters and centrality measures is unclear. The second problem is that the parameters and centrality measures themselves suffer from time-interval dependency. In the next section we will present one potential solution to the latter problem: the use of continuous-time models. We will then use this CT approach in the fourth section to address the other problem, describing how CT models can in principle be used to identify intervention targets.

4.3 CT Network Analysis: Accounting for Continuity

In the example given in the previous section we focused on the problems that arise when there is a single un-modeled latent wave of measurements when studying processes that vary over time. In psychological research, there are likely to be many more of these potentially observable but unmeasured process values between measurement occasions. While we are limited to observing psychological processes such as momentary stress and anxiety at discrete measurement occasions, they are likely to vary, evolve and exert influence on one another *continuously over time*: One's feelings of anxiety influences one's feelings of stress a second, thirty seconds, a minute from now and so forth, not only stress levels an hour from now. This means that there will be *infinitely many* latent values of these variables in-between measurement occasions, and moreover, that there are different lagged relationships between variables over a range of different time-intervals.

This perspective is consistent with viewing psychological phenomena as *continuous-time* processes, a perspective described in detail by Boker (2002) and promoted by proponents of CT statistical models in psychology (e.g. Coleman, 1968; Oud & Jansen, 2000; Oravecz et al., 2011; Deboeck & Preacher, 2016; Ryan et al., 2018; van Montfort et al., 2018; Ou, Hunter, & Chow, 2019; Driver et al., 2017; Voelkle et al., 2012). In SEM terms, we can represent a CT process as a path model in which there are infinitely many latent unobserved variable values in-between any two measurement occasions, spaced an infinitesimally small time-interval apart, as depicted in the left-hand panel of Figure 4.2 (see also Deboeck & Preacher, 2016).

Modeling CT processes is based on breaking down the relationship between observed measurement waves into their fundamental building blocks: direct lagged relationships operating over an infinitesimally small, hereby referred to as moment-to-moment, time-interval. These continuous moment-to-moment dynamics are captured by *differential equation* models, and the parameters of these models can be obtained from ESM-type data observed at longer intervals by fitting their *integral form* (Strogatz, 2015; Voelkle et al., 2012). Fitting this type of CT model allows us to overcome the time-interval problem by explicitly accounting for how lagged effects depend on the time-interval: The same moment-to-moment dynamics produce different lagged relationships depending on how far apart in time those variables are spaced.

In this section we will describe how CT models can be applied in the context of dynamical network analysis using a type of network representation known as a *local dependence graph*. First, we describe a simple differential equation and how its parameters can be interpreted. Second, we describe the CT-VAR model, which is the integral form of that simple differential equation and is the CT equivalent of the DT-VAR model. Third, we describe how using the CT-VAR model for network analysis affects our perspective on path-specific effects (based on previous work by Deboeck & Preacher, 2016), and what the consequences are for existing centrality measures.

4.3.1 Differential Equations: Moment-to-Moment Dynamics

The continuous-time equivalent of the VAR model is the first-order stochastic differential equation (SDE), which can be written as

$$\frac{d\boldsymbol{Y}(t)}{dt} = \boldsymbol{A}\boldsymbol{Y}(t) + \boldsymbol{W}(t) \tag{4.3}$$

where $\frac{d\boldsymbol{Y}(t)}{dt}$ on the right is the first derivative or the *rate of change* of the variables $\boldsymbol{Y}$ at time t (denoted $\boldsymbol{Y}(t)$). We can think of this derivative as being equivalent to a (scaled) *change score* $\boldsymbol{Y}(t+s) - \boldsymbol{Y}(t)$ for a very short time-interval ($\lim s \to 0$). This derivative is dependent on the current value of the *mean-centered* variables $\boldsymbol{Y}(t)$, and the $p \times p$ matrix of regression parameters which relates these two is called the *drift matrix* $\boldsymbol{A}$. The $\boldsymbol{W}(t)$ term represents a Wiener process, special kind of mean-zero white noise residual term (described in greater detail by, among others, Oud & Jansen, 2000; Voelkle et al., 2012; Voelkle & Oud, 2013). The type of dynamic system described by this differential equation is the same as that of the DT-VAR model: Each variable $\boldsymbol{Y}(t)$ has an equilibrium value, here defined by its mean; the Wiener process pushes the system away from this equilibrium; and the drift parameters $\boldsymbol{A}$ determine how the variables react to these shocks, eventually returning the system back to equilibrium over time (Strogatz, 2015; Ryan et al., 2018).

The drift parameters for the Stress-Discomfort system are plotted in network form on the right-hand side of Figure 4.2. Here, we use the hexagonal nodes to denote that the weights of the network are drift matrix parameters rather than lagged regression parameters, relating the value of one variable to the *rate of*

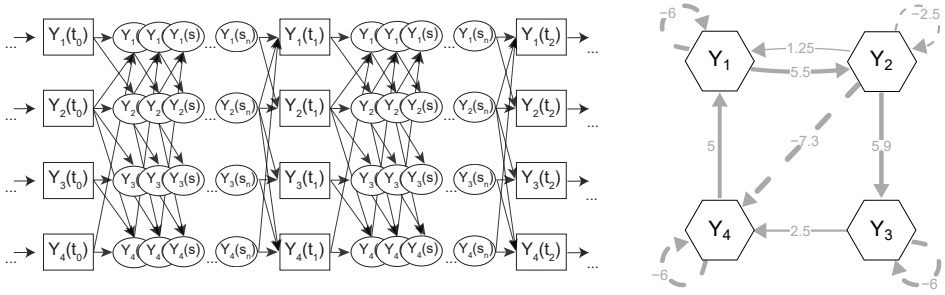


Figure 4.2: CT-VAR path-model (left-hand side) and CT network in the form of a weighted local independence graph representing $\boldsymbol{A}$ (right-hand side). In the path model, the latent variables and ellipses represent an infinite number of latent unobserved process values in-between measurement occasions, spaced an infinitesimally small time-interval apart; the presence of an arrow linking two variables in the path model denotes some non-zero dependency between them, conditional on all variables at the previous "wave", that is, a local dependency. For the network, solid arrows denote positive parameters and dashed arrows represent negative parameters.

change of the other, instead of *the value* of the other at the next occasion. This type of network representation is known as a *local dependence graph* (Schweder, 1970; Aalen, Borgan, Keiding, & Thormann, 1980; Didelez, 2008), of which the right-hand side of Figure 4.2 is a weighted variant. We will refer to this as a CT network, where the edges depict *direct moment-to-moment* dependencies between the time-varying processes.

The interpretation of drift matrix parameters is similar to that of a change-score model from the time-series literature. The diagonal parameters are known as *auto-effects* and are typically negative: If Stress takes on a high positive value, its negative auto-effect ($a_{11} = -6$) ensures it will move back towards equilibrium in response. The higher the absolute value of the auto-effect, the quicker the process will move back towards equilibrium (ignoring all other parameters), similar to an auto-regressive parameter close to zero. As such, the diagonal parameters are also referred to as reflecting the "centralizing tendency" of a variable, for instance by Oravecz et al. (2011). Off-diagonal elements known as *cross-effects* have a similar interpretation to cross-lagged parameters: The negative cross-effect of Anxiety on the rate of change of Physical Discomfort ($a_{42} = -7.3$) means that if Anxiety takes on a positive value, this will result in a decrease in the value of Physical Discomfort. The higher the absolute value of the parameter, the greater the magnitude of the effect (for more details on the interpretation of these parameters, see also Ryan et al., 2018; Voelkle et al., 2012; Oravecz et al., 2011).

From the CT network, we can see that all of the Stress-Discomfort variables have a negative auto-effect on themselves ($a_{11} = a_{33} = a_{44} = -6$); the auto-effect of Anxiety is closer to zero, which roughly corresponds to a stronger auto-regressive effect than the other variables ($a_{22} = -2.5$). Notably there are much fewer direct dependencies (i.e. fewer connections in the network) between processes on a moment-to-moment basis than we saw in the lagged DT-VAR networks in Figures 4.1(a) and (b): For instance, Stress has no direct moment-to-moment ef-

fect on Physical Discomfort ($a_{41} = 0$). The dependencies which are present are mostly positive: Stress and Anxiety exert reciprocal positive effects on one another ($a_{12} = 1.25$ and $a_{21} = 5.5$); Anxiety has a direct positive effect on the rate of change of Self-Conscious; Anxiety has a strong direct negative effect ($a_{42} = -7.3$), and Self-Conscious a positive direct effect ($a_{43} = 2.5$) on the rate of change of Physical Discomfort; finally, Physical Discomfort has a positive direct effect on the rate of change of Stress ($a_{14} = 5$). We will see why there is a discrepancy between the moment-to-moment CT network and the lagged relationships in the DT-VAR network by connecting these two models through the integral form of the differential equation.

4.3.2 The CT-VAR model

We can estimate the parameters of the differential equation model from ESM data by making use of its integral form known as the CT-VAR or Ornstein-Uhlenbeck process (Oud & Jansen, 2000; Oravecz, Tuerlinckx, & Vandekerckhove, 2009; Voelkle et al., 2012; Driver et al., 2017).[3] This allows us to express the relationship between measurement waves observed with any spacing Δt in terms of the moment-to-moment relationships of the first-order SDE:

$$\boldsymbol{Y}(t_\tau) = \boldsymbol{c} + \boldsymbol{e}^{\boldsymbol{A}\Delta t_\tau}\boldsymbol{Y}(t_{\tau-1}) + \boldsymbol{\epsilon}(\Delta t_\tau) \tag{4.4}$$

where variables at the current measurement occasion $\boldsymbol{Y}(t_\tau)$ are regressed on variables at the previous measurement occasion $\boldsymbol{Y}(t_{\tau-1})$. Note that τ refers to the measurement occasion, whereas t refers to the actual time when this measurement took place. Hence, Δt_τ indicates the time interval between two consecutive measurement occasions, which may differ across pairs of observations. The current expression of the CT-VAR model is very similar to the DT-VAR model that was presented in Equation (4.1). The $\boldsymbol{c}$ term represents a vector of intercepts, and $\boldsymbol{\epsilon}(\Delta t_\tau)$ represents the residual vector, normally distributed with mean zero and variance-covariance matrix that is also a function of the time-interval (for more details see Oud & Jansen, 2000; Voelkle et al., 2012; Voelkle & Oud, 2013). In place of the $\boldsymbol{\Phi}$ matrix in Equation (4.1), these lagged variables are related by $\boldsymbol{e}^{\boldsymbol{A}\Delta t_\tau}$, the matrix exponential of the drift matrix $\boldsymbol{A}$, multiplied by the time-interval between those measurement occasions.

It follows that the CT- and DT-VAR lagged regression matrices are related to each other by the expression

$$\boldsymbol{e}^{\boldsymbol{A}\Delta t} = (\boldsymbol{e}^{\boldsymbol{A}})^{\Delta t} = \boldsymbol{\Phi}(\Delta t) \tag{4.5}$$

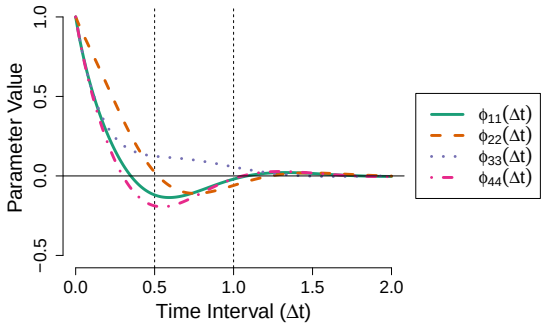
[3] With very high frequency data, this model can be estimated by first calculating the derivative and then fitting the first-order SDE directly to data (see for example Boker, Deboeck, et al., 2010; Chow, 2019). However, when observations are collected with a longer time-interval between measurements (as is typical in ESM settings), we can estimate these same model parameters by fitting the integral form of the differential equation model. Yet another approach involves numerically integrating the differential equation during estimation (e.g., Ou et al., 2019). The current approach, using the analytic integral form allows us to demonstrate the connection between current DT models and their CT equivalent most clearly.

which states that the lagged parameters for any particular time-interval $\mathbf{\Phi}(\Delta t)$ can be found by taking the matrix exponential of the moment-to-moment drift matrix $\boldsymbol{A}$ to the power of the length of that time-interval Δt (cf. Oud & Jansen, 2000; Voelkle et al., 2012). Notice the similarity between this relationship, and the expression used to relate the half-hour and one-hour parameter matrices in Equation (4.2): To find the lagged relationships at a longer time-interval, we again take the appropriate matrix exponent of the lagged relationships at the shorter interval.

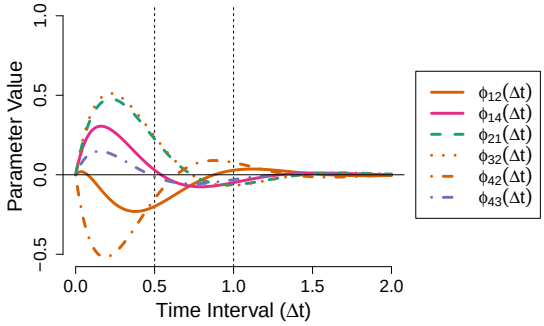
Since there is a clear connection between the CT- and DT-VAR models, we can now consider the two consequences of the time-interval problem from a CT perspective. First, the conclusions we would intuitively make based on a DT-VAR model using *any one particular interval* may differ from those we would make at any other interval. Figure 4.3 depicts the lagged relationships of the Stress-Discomfort system over a range of time-intervals, from zero to two hours, showing clearly that observing this process at say, 15 minute intervals ($\Delta t = 0.25$) instead of half-hour or one-hour intervals leads to a quantitatively and qualitatively different set of conclusions about the underlying process. In Figure 4.4 we show how the DT network centrality measures also change according to the time-interval, leading to different conclusions regarding optimal intervention targets when applying current standard practice (an effect which is particularly pronounced for the betweenness measure in panel (c)).

Second, the CT perspective shows that lagged regression parameters $\mathbf{\Phi}(\Delta t)$ at any interval should be interpreted as total rather than direct effects from a path-tracing perspective (Aalen et al., 2016; Deboeck & Preacher, 2016). This follows directly from the observation that the matrix exponential relationship in Equation (4.5) is simply a generalization of the path-tracing operation described above when relating the half-hour and one-hour networks (Equation (4.2)): The lagged relationships at a very short time-interval are described by (a function of) the $\boldsymbol{A}$ matrix, and to find the relationships between variables spaced further (i.e., Δt) apart, we apply a path-tracing operation through the latent values of $\boldsymbol{Y}(t)$ in-between those occasions. See Appendix 4.C for an accessible derivation of this term as a path-tracing operation.

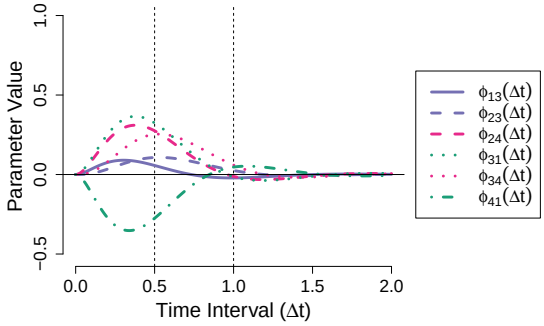
This interpretation also allows us to explain the changing patterns of lagged relationships we observe in Figure 4.3. For example, we know that Stress has no direct moment-to-moment effect on the rate of change of Physical Discomfort ($a_{41} = 0$), but we see that the corresponding lagged relationship in the DT model ($\phi_{41}(\Delta t)$ in Figure 4.3(c)) is strongly negative at short intervals and weakly positive at longer intervals: This is because it is a total effect, made up of one negative indirect pathway (through Anxiety) and one positive indirect pathway (through Anxiety and Self-Conscious). In order to calculate lagged direct and indirect effects from the CT perspective, we can use path-tracing methods defined by Deboeck and Preacher (2016) and Aalen et al. (2016) and described in detail in Appendix 4.D. The direct lagged effect of Stress on Physical Discomfort is found by first omitting the indirect pathways from $\boldsymbol{A}$ (in this case, a_{31}, a_{42}, a_{32} and a_{43}) before applying the matrix exponential function in Equation (4.5). In so doing we trace a path from Stress to Physical Discomfort which does not pass through any



(a) Auto-regressive parameters $\phi_{ii}(\Delta t)$

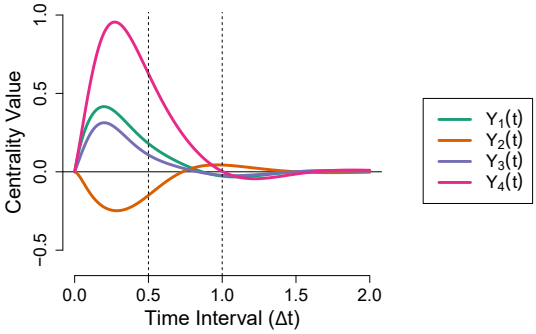


(b) Cross-lagged parameters $\phi_{ij}(\Delta t)$ for which $a_{ij} \neq 0$

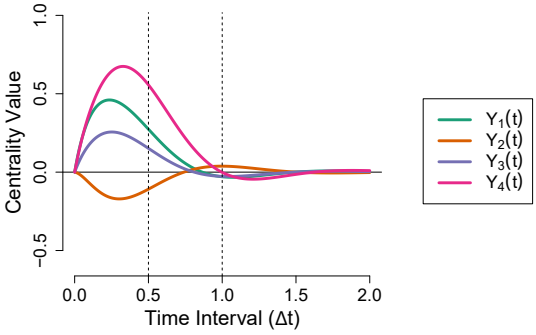


(c) Cross-lagged parameters $\phi_{ij}(\Delta t)$ for which $a_{ij} = 0$

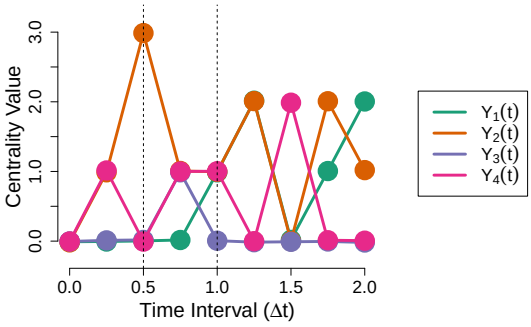
Figure 4.3: Lagged regression parameters as a function of the time-interval for the Stress-Discomfort system. Black dotted lines indicate the parameter values of the half-hour and one-hour networks in Figure 4.1(a) and (b).



(a) Two-Step Expected Influence $EI^{(2)}$



(b) One-Step Expected Influence $EI^{(1)}$



(c) Betweenness Centrality *BC*

Figure 4.4: Centrality measures as a function of the time-interval for the Stress-Discomfort system. Black dotted lines indicate the centrality values of the half-hour and one-hour networks in Figure 4.1(a) and (b).

unobserved, intermediate values of Anxiety or Physical Discomfort. To find the indirect effect, we can either omit the direct dependency between the two target variables, or alternatively we can take the difference between the direct and total effects (see Appendix 4.D for details).

4.3.3 Conclusion

To summarize, the CT-VAR model offers a promising alternative to DT-VAR network approaches, allowing us to directly address the time interval problem. First, it allows for an elegant treatment of unequal time-intervals between observations, which is a typical characteristic of ESM data. Second, the CT model offers an alternative conceptualization of the dynamic network structure by breaking down lagged relationships between measurements into their most fundamental building blocks, that is, the moment-to-moment relationships encoded by the drift matrix $\boldsymbol{A}$. Given an estimate of this drift matrix (which can be obtained with software packages such as *ctsem* (Driver et al., 2017) or *dynr* (Ou et al., 2019)), the relationship in Equation (4.5) allows us to explore how the dynamic relations depend on the time-interval, as we illustrated with the Stress-Discomfort system. Third, we can use path-tracing logic suggested by previous authors to calculate direct, indirect and total lagged relationships in a CT network.

The CT perspective also has consequences for the interpretation and practical utility of DT centrality measures. Specifically, existing measures fail to account for the CT nature of the underlying process, and —as a consequence— even their intuitive interpretation as a summary of direct, indirect and total relationships is not valid from a CT perspective. This implies that using DT centrality measures to decide which variable should be intervened on is simply unfounded. Hence, we need to define new centrality measures that adhere to the CT perspective, but that also have a clear link with interventions.

4.4 Interventions and Centrality for CT Networks

In this section we address the intervention problem by developing new fit-for-purpose centrality measures that both account for the CT nature of the underlying process, and that have an explicit link to the effects of interventions as defined in the interventionist causal inference framework. This allows us to use the CT approach to decide which variable in the network to intervene on. To achieve this goal, we need to take a number of steps.

As we have detailed above, centrality measures tend to serve as summaries of total, direct or indirect effects: For instance, they are intended to reflect the degree of influence a particular variable has on all others in the network, or they are supposed to capture the amount of information flow which passes through a particular variable. As such, to develop new centrality measures that can be interpreted in such terms, we first need to link total, direct and indirect effects in a CT model to the effects of variable interventions.

Once we know how to interpret path-specific effects in terms of particular interventions on one or more variables in the system, then we can define three new centrality measures as summaries of these path-specific effects. By doing this, the new centrality measures we develop can be interpreted in terms of the expected change over time in the network given a *specific type of intervention* on a *specific variable*. Hence these measures can be used to inform researchers about which variable to target, and what kind of intervention should be applied to achieve the desired effect. At the end of this section we apply these CT centrality measures to our running example and show that they lead to different and novel insights into the underlying dynamic system than what would be concluded using current DT measures.

Throughout, we take a simplified approach, based on a number of assumptions that were described briefly in Section 4.2.3.1. That is, we assume that interventions can be made in the system without changing how the variables in the system relate to one another (i.e. modularity), and that all relevant variables are measured and included in the model (i.e. no unmeasured confounders). In addition to assuming a first-order process, we also assume that the CT-VAR model fully describes the dynamics of this first-order process. These simplifying assumptions allow us to provide a straightforward link between CT-VAR model parameters and the effects of variable interventions under ideal conditions, which we will illustrate using the Stress-Discomfort example.

4.4.1 Path-Specific Effects and Interventions

Here we generalize intervention-based definitions of path-specific effects in DT lagged regression models (cf. Eichler & Didelez, 2010; VanderWeele, 2015) to a CT-VAR setting. In doing so, we provide a clear conceptual link between a path-tracing effect in the CT model and an actual intervention.

4.4.1.1 Total Effects and Interventions

In terms of the Stress-Discomfort system, let's imagine that we are able to induce a momentarily high experience of Anxiety in our participant, for instance by making the participant view a negative photograph, a manipulation which has been shown to increase state anxiety in lab studies (Richards & Whittaker, 1990; Richards, French, Johnson, Naparstek, & Williams, 1992). We refer to such an intervention at one moment in time as an *acute intervention* (also referred to as an atomic intervention by Eichler & Didelez, 2010). In the time-series literature this is sometimes referred to as an *impulse*, and in the causal inference literature, they denote such an intervention using the do operator, with $do(Y_2(t) = 1)$ meaning we intervene to set Anxiety to a value of one at time t.

We can visualize this intervention by plotting the expected trajectories of the four different variables in our Stress-Discomfort system following this intervention: This is shown in Figure 4.5(a). Following the initial intervention on Anxiety, the other three variables leave their equilibrium. Eventually, the effect of the intervention fades, and all variables return to their resting state. This shows that

an intervention on Anxiety *changes* the value of all variables in our model. If we want to describe the effect of this intervention on a particular variable, say Physical Discomfort, in formal terms we would express this as a *difference in expected value* of Physical Discomfort when comparing two different interventions on Anxiety.[4] For example, if we instead set Anxiety to its equilibrium value, $do(Y_2(t) = 0)$, none of the variables would move away from equilibrium. By considering the contrast between the two expected values of Physical Discomfort (some time Δt after these interventions), we get the *Total Effect* of Anxiety now on Physical Discomfort. This is written formally as

$$TE_{24}(\Delta t) = \mathrm{E}\Big[Y_4(t+\Delta t) \;\Big|\; do(Y_2(t) = 1)\Big] - \mathrm{E}\Big[Y_4(t+\Delta t) \;\Big|\; do(Y_2(t) = 0)\Big] \tag{4.6}$$

Since we are dealing with a dynamic system, the value of this total effect depends on the time-interval that elapses after the intervention is applied: In Figure 4.5(a) we can see that initially the intervention resulted in Physical Discomfort taking on a negative value, but at longer intervals the effect was to make Physical Discomfort take on a weak positive value.

In order to calculate the value of this total effect, we need to plug in an expression or model for these expected values. Using the simplifying assumptions described above, we plug in the CT-VAR model for each expected value, and we can express this total effect as

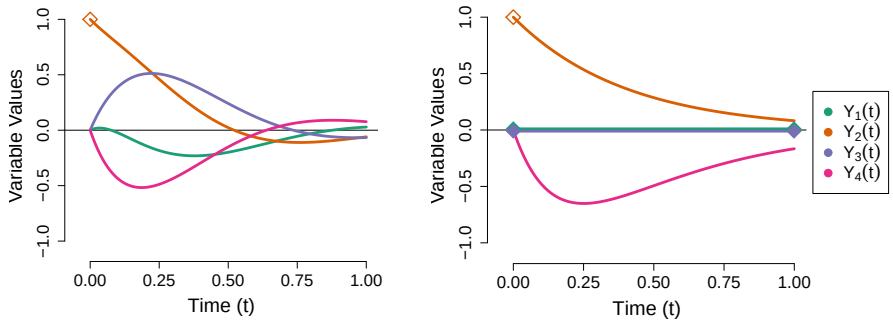
$$TE_{24}(\Delta t) = \boldsymbol{e}^{\boldsymbol{A}\Delta t}{}_{[42]} \tag{4.7}$$

where $\boldsymbol{e}^{\boldsymbol{A}\Delta t}{}_{[42]}$ denotes the parameter in the fourth row, second column of $\boldsymbol{e}^{\boldsymbol{A}\Delta t}$. Critically, the expression for the causal effect presented above is identical to the path-tracing definition of a total effect given by Deboeck and Preacher (2016) and Aalen et al. (2016) and described in Appendix 4.D. We show the equivalence of these two different definitions of the total effect in Appendix 4.E. This implies that the CT total effect, under idealized conditions, reflects a very specific type of variable intervention: An intervention to increase the value of the *cause* variables (Anxiety) at a single point in time. The value of the total effect reflects the change we would expect to see in the *effect* variable (Physical Discomfort) some time later. These definitions may seem intuitive, obvious or overly simplistic, matching the usual interpretation given to regression parameters: If we increase X, how do we expect Y to change? However, this conceptual link will prove helpful in defining other path-specific effects which may be less intuitive, and in defining centrality measures later.

4.4.1.2 Direct Effects and Interventions

The interventionist definition of a direct effect typically consists of two interventions: One to *change* the cause variable, just as we did for the total effect, and one

[4] Here since we are dealing with a single-subject dynamic process, the expectation is defined with respect to the stochastic input, that is, the normally distributed noise process, as is standard in time-series approaches (Hamilton, 1994). Given that we do not change the properties of the system (by the modularity assumption) and that time points t are interchangeable (by the stationarity assumption), we can interpret this as an expectation for an individual over a population of time points t. This is analogous to a causal effect for that individual for an unspecified point in time t.



(a) *TE* Intervention

(b) *DE* Intervention

Figure 4.5: Illustration of each total and direct effect of Anxiety ($Y_2(t)$) on Physical Discomfort ($Y_4(t+\Delta t)$) in terms of the interventions they describe in the system over time. Acute interventions are indicated by empty diamonds, and continuous interventions by filled diamonds.

to *keep* one or more mediating variables *fixed* (cf. Robins, 2003; Robins & Richardson, 2010; VanderWeele, 2015). For example, suppose that we can intervene in our system to keep the variables Stress and Self-Conscious fixed to a normal value *at every moment in time over an interval*. We will refer to this type of intervention as a *continuous intervention*, denoting it $do(\overline{Y_1(t+\Delta t)} = 0)$ and $do(\overline{Y_3(t+\Delta t)} = 0)$ for Stress and Self-Conscious respectively.[5] For example, we can imagine that we can prompt the participant to engage in a long-lasting mindfulness meditation focused on reducing stress reactivity (Hoge et al., 2013), or, alternatively, that we can administer a drug which keeps stress levels fixed to a low or normal level.

Now suppose that we are interested in evaluating the effect of momentarily increasing Anxiety, on the value of Physical Discomfort, if we intervene to keep Stress and Self-Conscious fixed. This is visualized by the trajectories in Figure 4.5(b). Just like the total effect intervention, Anxiety starts at a high level and dissipates back to equilibrium. Here however, we keep Stress and Self-Conscious fixed at all moments in time, so they stay at their equilibrium value. We can see that the effect of increasing Anxiety on Physical Discomfort has changed: Physical Discomfort is pushed even further from equilibrium, taking on a stronger negative value at short intervals. Since we keep the mediating variables fixed, we no longer activate the positive feedback loop present in the total effect. Instead, Physical Discomfort takes longer to return to baseline, still taking on a negative value at $t = 2$.

Just as we did for the total effect, we can define the effect of this intervention formally as the difference between two conditional expectations. Here we change

[5] We can consider this to be a special case of an acute intervention defined by Eichler and Didelez (2010), that is repeated continuously over time. See Appendix 4.E for details. Note that this type of direct would be referred to as a *controlled direct effect* rather than a natural direct effect in the causal inference literature (VanderWeele & Tchetgen Tchetgen, 2017). Since we assume a linear model, the controlled and natural direct effect formulations lead to equivalent results.

Anxiety, but keep Stress and Self-Conscious fixed in both cases, and we can define the effect of this combination of interventions as the *direct effect*. It can be written formally as

$$\begin{aligned} DE_{24\cdot\overline{13}}(\Delta t) = &\mathrm{E}\Big[Y_4(t+\Delta t) \;\Big|\; do(Y_2(t)=1), do(\overline{Y_1(t+\Delta t)}, \overline{Y_3(t+\Delta t)}=0)\Big] \\ &-\mathrm{E}\Big[Y_4(t+\Delta t) \;\Big|\; do(Y_2(t)=0), do(\overline{Y_1(t+\Delta t)}, \overline{Y_3(t+\Delta t)}=0)\Big] \end{aligned} \tag{4.8}$$

and by plugging in the CT-VAR parameters for each conditional expectation, we can express the effect of this intervention as

$$DE_{24\cdot\overline{13}}(\Delta t) = \boldsymbol{e}^{\boldsymbol{A}^{(D[-1,-3])}\Delta t}{}_{[42]} \tag{4.9}$$

where $\boldsymbol{A}^{(D[-1,-3])}$ denotes the drift matrix in which the indirect pathway parameters linking Anxiety to Physical Discomfort through the mediating variables (that is, a_{12}, a_{32}, a_{43}) are set to zero; hence, in this drift matrix only the *direct links* between Anxiety and Physical Discomfort are retained. The proof that the effect of this intervention can be expressed in this way is given in Appendix 4.E.

Again, we can see that this expression is exactly equivalent to the path-tracing definition of a direct effect in a CT model given by Deboeck and Preacher (2016). This means that the CT direct effect under ideal conditions describes the effect of a very specific set of interventions: An acute intervention to change the value of the cause variable, combined with a whole set of continuous interventions to keep each mediating variable fixed in value. Note that we can also define direct effects in which only one of the mediating variables is kept fixed with a continuous intervention in an analogous way: For example, we could describe the direct effect of Anxiety on Physical Discomfort *relative to* Self-Conscious as $DE_{24\cdot\overline{3}}(\Delta t)$, in which only Self-Conscious is kept fixed by a continuous intervention. Then we would need to omit only the parameters a_{32} and a_{43} from the altered drift matrix $\boldsymbol{A}^{(D[-3])}$.

Since direct effects consist of combinations of at least two or more interventions (depending on the number of mediators considered), they may not be the most useful measure to inform interventions in practice. However, it is necessary to define the direct effect in this way in order to define indirect effects, which we discuss next, and which form the basis of the indirect effect centrality measure that we present later.

4.4.1.3 Indirect Effects as Intervention Contrasts

Defining the indirect effect as a combination of unique interventions is somewhat more challenging than doing so for the total and direct effect. This has been discussed in detail both for general mediation models, and more recently for CT models (Robins, 2003; Robins & Richardson, 2010; Didelez, 2019). Here we will define the indirect effect as a *contrast* between a total effect intervention and a direct effect intervention.

Consider that the total effect and the direct effect both describe aspects of the relationship between a cause variable $Y_i(t)$ and an effect variable $Y_j(t+\Delta t)$: They

both describe what happens to $Y_j(t+\Delta t)$ if we intervene in an acute way on $Y_i(t)$. The only difference between the two is that the influence of this acute intervention may differ if we *also* intervene to keep one or more mediator variables fixed, that is, the total effect may not be the same as the direct effect. Hence, for each mediator or set of mediators, we can quantify exactly what difference it makes if we keep it or them fixed. We do this by looking at the *difference* between the total effect and the direct effect.

In other words, the indirect effect $IE(\Delta t)$ describes how the effect of acutely intervening on $Y_i(t)$ *changes* when we also intervene to keep the mediator(s) Y_k fixed from t to $t+\Delta t$ (i.e., with a continuous intervention) . For instance, we may be interested in the mediating roles that both Stress and Self-Conscious play in the relationship between Anxiety on Physical Discomfort. To quantify this, we would define the indirect effect of Anxiety on Physical discomfort (relative to Stress and Self-Conscious) as

$$IE_{24\cdot\overline{13}}(\Delta t) = TE_{24}(\Delta t) - DE_{24\cdot\overline{13}}(\Delta t). \tag{4.10}$$

In terms of the trajectories in Figure 4.5 the indirect effect is the difference in value of Physical Discomfort at any point in time in panel (a)(representing the total effect $TE(\Delta t)$), and the corresponding point in time in panel (b) (representing the direct effect $DE(\Delta t)$), and so this particular indirect effect describes the mediating role of the variables Stress and Self-Conscious combined. We can express this indirect effect in terms of the CT-VAR parameters as

$$IE_{24\cdot\overline{13}}(\Delta t) = \boldsymbol{e}^{\boldsymbol{A}\Delta t}{}_{[42]} - \boldsymbol{e}^{\boldsymbol{A}^{(D[-1,-3])}\Delta t}{}_{[42]} \tag{4.11}$$

that is, as a difference between the total and direct effect calculations given above. It follows that the effect of this intervention is equivalent to the path-tracing definition of the indirect effect (given in Appendix 4.D).

We may be interested in the indirect effect if we wish to answer questions like: How does keeping Stress and Self-Conscious fixed alter the effect of a momentary increase in Anxiety on Physical Discomfort? We can get an idea of the role of these mediators by comparing the total effect and the direct effect trajectories in Figure 4.5(a) and (b) respectively. In this instance we see that keeping both mediators fixed changes how Stress effects Self-Conscious, meaning Self-Conscious reacts in a more strongly negative way to changes in Stress than if both mediators are not kept fixed. We may also be interested in assessing indirect effects one mediator at a time. For instance, if we want to decrease the effect that experiencing an acutely high level of Anxiety has on feelings of Physical Discomfort, we may choose to either apply a continuous intervention on Stress ($IE_{24\cdot\overline{1}}(\Delta t)$) or Self-Conscious ($IE_{24\cdot\overline{3}}(\Delta t)$), depending on whichever particular indirect effect is largest. It is this kind of indirect effect that we will use later in defining a centrality measure that describes the flow of information in the network through a particular variable.

4.4.2 Centrality Measures to Identify Intervention Targets

Having established the link between CT path-specific effects as described by previous authors (Deboeck & Preacher, 2016; Aalen et al., 2018) and variable interventions from the causal inference framework (Eichler & Didelez, 2010; Pearl, 2009; Robins, 2003), we now propose three new centrality measures for CT networks. Each centrality measure is explicitly defined as a summary of one of the intervention-based path-specific effects defined above. This means that these centrality measures are functions of the time-interval and have a clear link to a particular type of variable intervention. The three measures we introduce each capture a summary of the total, direct, and indirect effects of one variable on all others. The first and last measures are likely to be most interesting and useful for researchers in practice, as they allow to choose the optimal target for an acute and continuous intervention respectively. The second measure, summarizing direct effects, is less straightforward to use to inform interventions, but is included for the sake of completeness.

4.4.2.1 CT Total Effect Centrality

We define our first new centrality measure as the *Total Effect Centrality* (TEC) of a variable, which can be calculated by summing the total effect of $Y_i(t)$ on all other variables, at a particular time-interval

$$TEC_i(\Delta t) = \sum_{j \neq i}^{p} TE_{ij}(\Delta t). \tag{4.12}$$

hence, we sum over all the total effects of Y_i on other variables in the network (excluding Y_i itself). The TEC thus summarizes the effect of a *single intervention*, an acute intervention to change $Y_i(t)$, on the system as a whole, that is, the cumulative effect on the network, some time-interval Δt later. Furthermore, since we explicitly make this centrality measure a function of the time-interval, we can examine how the cumulative effect of this intervention varies and evolves following the intervention moment. By calculating this measure for each variable, we can directly inform the choice of optimal intervention target: Which variable should I change the acute value of to achieve the biggest change in the cumulative activation levels in the rest of the system later?

Figure 4.6(a) shows the TEC of each variable in the Stress-Discomfort system over a range of intervals, from $\Delta t = 0$ to $\Delta t = 1.5$. From this we can see that at short intervals, an acute intervention to increase Physical Discomfort has the biggest cumulative effect on the network: Overall, this intervention on Physical Discomfort results in the other variables increasing in value over the next half an hour or so, before eventually the effect of this intervention fades away. Interventions to increase Stress and Self-Conscious respectively have similar but weaker effects. Notably, an intervention to increase Anxiety has a weak net negative effect on the system at shorter intervals, and a weak net positive effect at longer intervals: We would expect this based on our visualization of that intervention in

Figure 4.5(a), where a pulse to Anxiety resulted in Stress and Physical Discomfort taking on negative values at short intervals.

Based on this, if we want to pick the optimal intervention target for an acute intervention, the TEC measure allows us to see that Physical Discomfort is the optimal target for this type of intervention, assuming we can set Physical Discomfort to a low or negative value (e.g. $do(Y_4(t)) = -1$).

4.4.2.2 Direct Effect Centrality

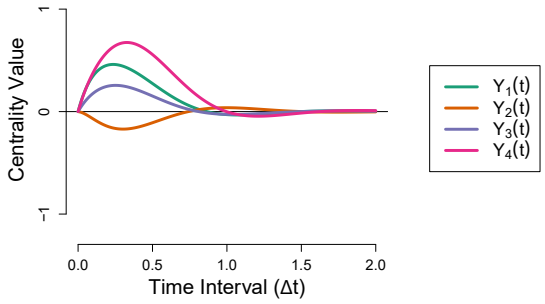
To summarize the *direct influence* of one variable on all others, we propose to take a summary of different *direct effects*, where the latter is defined in Equation (4.8). For instance, we defined the direct effect of Anxiety above as the change in an outcome variable, Physical Discomfort, following the combination of an acute intervention to change Anxiety, and continuous interventions to keep all other mediating variables (Stress and Self-Conscious) fixed. The measure we introduce here, *Direct Effect Centrality* (DEC) takes the sum of all such direct effects from $Y_i(t)$ to all other possible outcome variables, if all remaining variables in the model are kept fixed

$$DEC_i(\Delta t) = \sum_{i \neq j}^{p} DE_{ij \cdot \overline{k}}(\Delta t) \tag{4.13}$$

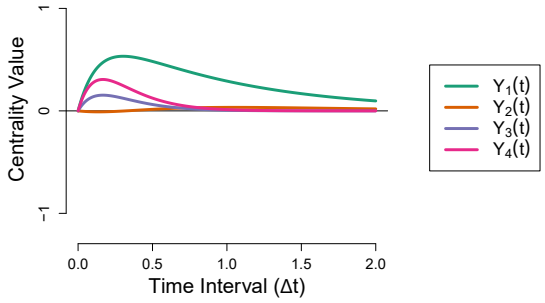
where for each pair $Y_i(t)$ and $Y_j(t + \Delta t)$ all possible mediating variables $k \in p \backslash (i, j)$ are kept fixed. Note that unlike the TEC measure, DEC reflects a summary measure of *different interventions*, as for each pair $Y_i(t)$ and $Y_j(t + \Delta t)$, there is a different set of mediators that must be intervened on to establish the direct effect. For instance, in Figure 4.5(c) we showed the direct effect of Anxiety on Physical Discomfort when intervening to keep Stress and Self-Conscious constant ($DE_{24 \cdot \overline{13}}(\Delta t)$). This represents only one of the three components which make up the DEC of Anxiety: The other two terms consider interventions where either Self-Conscious and Physical Discomfort are kept constant ($DE_{21 \cdot \overline{34}}(\Delta t)$) or Physical Discomfort and Stress are kept constant ($DE_{23 \cdot \overline{14}}(\Delta t)$). Although it is not straightforward to use this measure to choose a specific intervention target, this measure does reflect that part of one variables relationship with the network as a whole which is truly a result of only direct relationships (from a path-tracing perspective).

Figure 4.6(b) shows the direct effect centrality of each variable over a range of intervals. From this we see that Stress has the highest positive DEC, Self-Conscious and Physical Discomfort have weaker overall positive direct influences, that fade to zero at a quicker rate, and Anxiety has a net direct influence closest to zero across a range of intervals.[6] Although Stress has the strongest truly direct influence on the network, the practical value of the DEC —or any measure that is based on direct effects— for choosing intervention targets is limited, as it is

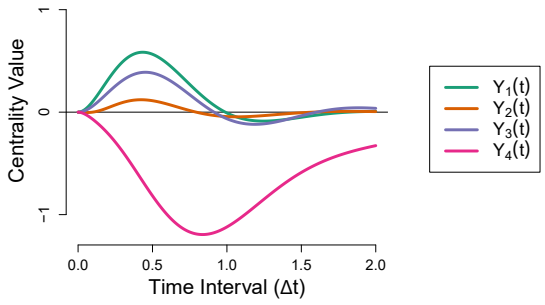
[6]The high DEC of Stress is due to the positive feedback loop between Stress and Anxiety in the underlying drift matrix ($a_{21} = 5.5$, $a_{12} = 1.25$). Since the direct effects involve interventions to control the rest of the variables in the system, and no other pair of variables has this direct feedback loop relationship, all other values of DEC are lower.



(a) Total Effect Centrality $TEC(\Delta t)$



(b) Direct Effect Centrality $DEC(\Delta t)$



(c) Indirect Effect Centrality $IEC(\Delta t)$

Figure 4.6: Illustration of the new total, direct and indirect centrality measures for CT networks, applied to the Stress-Discomfort system.

based on a complex combination of different interventions, that cannot actually take place all at the same time.

4.4.2.3 Indirect Effect Centrality

A much more appealing way to identify an intervention target is by quantifying the role that variable plays *as a mediator* of other relationships between variables in the network. To do so, we can use the indirect effect measure described in Equation (4.10) above. Recall that the CT indirect effect captures the *change* in the effect of $Y_j(t)$ on $Y_k(t + \Delta t)$, if we intervene to keep the mediator Y_i fixed at every moment in time ($TEC_{jk}(\Delta t) - DEC_{jk\cdot i}(\Delta t)$). Hence, we define the *Indirect Effect Centrality* (IEC) of a *mediator variable* Y_i as

$$IEC_i(\Delta t) = \sum_{(j,k):j\neq k\neq i} IE_{jk\cdot i}(\Delta t) \tag{4.14}$$

that is, the sum of all possible indirect effects between other pairs of variables $Y_j(t)$ and $Y_k(t + \Delta t)$, in which Y_i serves as the only mediator. Note here that in comparison to how we described the indirect effect above, we have switched the i, j and k notation to reflect that the $IEC(\Delta t)$ is defined as a property of a mediator, instead of a property of one particular cause-effect relationship. The summation denotes that we omit auto-regressive relationships ($j \neq k$) and pairs of variables where the mediator is either the cause or effect variable ($j \neq i$ and $k \neq i$). The IEC therefore can be understood as quantifying how a *continuous intervention* on Y_i changes the effects of other variables on each other.

As such, this measure may be especially of interest in the case of networks of psychopathology variables. For instance, in the Stress-Discomfort system, we would like to avoid a high value on all four variables as much as possible. The current measure can be used to determine which of these variables is most important in terms of mediating the effects of one variable on another in the system, such that by intervening on this variable, these indirect paths become blocked and the flow of activation from one variable to another is interrupted.

Figure 4.6(c) shows the IEC of each variable over a range of intervals. From this we can see that Physical Discomfort has the strongest indirect effect centrality in absolute terms. A strong negative value of IEC means that keeping Physical Discomfort fixed actually *increases* the size the effects of other variables on each other, since the component direct effects are greater than the corresponding total effects. This happens because Physical Discomfort plays a key role in the only *negative* feedback loop in the network: Since an increase in Anxiety actually *decreases* Physical Discomfort ($a_{42} = -7.3$), the total effect of Anxiety on Stress is less strong than its direct effect. If, however, we intervene to keep Physical Discomfort fixed, then this negative compensating effect is not activated, meaning an increase to Anxiety in fact has a *greater* effect on the network as a whole. In contrast, Stress has the largest positive IEC, meaning that keeping Stress fixed *decreases* the effects of other variables on one another.

From this we would conclude that we should choose *Stress* as a target for a continuous intervention, as it decreases the short-term impact of other variables

in the network on each other. However, we should avoid at all costs applying a continuous intervention on Physical Discomfort: Such an intervention would in fact increase the strength of positive relationships between the other variables.

4.4.2.4 Comparing CT and Existing Centrality Measures

The new CT centrality measures introduced here have a clear link to the choice of intervention target: The TEC and IEC measures can be easily used to select the best target for an acute and continuous intervention respectively. Furthermore, it is clear from comparing the CT centrality measures in Figure 4.6 with the existing DT measures in Figure 4.4 that the new measures lead to different choices in intervention targets.

The TEC closely matches the substantive interpretation typically given to the $EI^{(2)}$ measure (Figure 4.4(a)); however, its value is actually equivalent to calculating the $EI^{(1)}$ measure across a range of time-intervals (Figure 4.4(b)). This implies that, unwittingly, the measure that is typically used to summarize direct influence in a DT network in fact describes the total influence of an acute intervention at a specific time-interval. In contrast to the $EI^{(1)}$ measure, the DEC provides a rather different picture of which variable has a strong direct influence on the network. However, we would argue that truly direct influence is less important when choosing intervention targets in the sense described above.

Finally, we believe that the IEC is a useful way to summarize the mediating role a variable plays in the network, allowing us to choose an optimal target for a continuous intervention. Comparing this measure to Betweenness Centrality in Figure 4.4(c), which is the measure used to quantify indirect influence in DT networks, we see again that we would reach entirely different conclusions. Specifically, there are few intervals at which calculating Betweenness would correctly lead us to choose Stress as an intervention target. Perhaps even more crucially, there are numerous intervals at which we would choose Physical Discomfort as an intervention target, although we know from the IEC measure that we should avoid applying a continuous intervention to this variable.

4.5 Discussion

In this paper we have critically appraised current best practice in dynamical network analysis. We have especially focused on the interpretation of DT-VAR models as representing direct causal dependencies, and on the use of centrality measures predicated on that interpretation to identify optimal intervention targets. We identified two major problems with current best practice and addressed each in turn. First, we addressed the time-interval problem by introducing a continuous-time approach to network analysis. CT models aim to capture the moment-to-moment dynamics operating between processes, which can be represented in network form as a local dependence graph. The estimation of a CT model overcomes both practical and conceptual shortcomings of the DT approach, providing a new outlook on how parameters should be interpreted, and the ability to explore how lagged relations vary and evolve as a function of

the time-interval. Second, we address the lack of clarity regarding how current centrality measures should be used to inform intervention targets. We do this by introducing new centrality measures for CT models which have a clear link to interventionist concepts of causal relationships and can be used to identify the optimal target for either an acute or continuous intervention.

These developments represent a promising new approach to dynamical network analysis. From a practical point of view, recent R packages *ctsem* (Driver et al., 2017) and *dynr* (Ou et al., 2019), have made it relatively easy to apply the CT modeling approach in empirical research (e.g., Voelkle et al., 2012; Voelkle & Oud, 2013; Oud, Voelkle, & Driver, 2018; Ryan et al., 2018), and moreover, in the supplementary materials of this paper we provide extensive *R* functions which can be used to calculate CT path-specific effects and centrality measures introduced in the current paper based on a given CT-VAR drift matrix.[7]

As this was a first step towards linking CT approaches, network analyses and interventionist concepts of causal relationships, our approach was necessarily simplified, both with respect to the interventions we consider, and our treatment of the assumptions necessary for this type of causal inference. Eichler and Didelez (2010) describe a variety of different interventions that could be applied to dynamic systems, and a variety of assumptions that must be met for the effects of these interventions to be identified from data. Driver and Voelkle (2018) describe how various hypothetical interventions can be simulated from a CT model, however, they do not do so within an interventionist framework, and so the assumptions needed to identify the effects of those interventions are unclear. More research is needed to investigate and evaluate identifiability assumptions in a psychological context, and to operationalize actual psychological treatments as dynamic systems interventions. For instance, cognitive-behavioral interventions may be better defined as interventions on moderators of symptom-symptom relationships (i.e., lagged parameters of a network), rather than interventions on the symptoms themselves.

Furthermore, the dynamic models we focused on here, that is the DT-VAR and its CT-equivalent, are very simple models, describing a system that fluctuates around a single fixed point. Although this reflects the majority of dynamical network analyses in empirical research, it may be more beneficial —from a theoretical point of view— to investigate dynamic system models which can show a qualitative change in behavior, for example bi-stable systems which can transition between different equilibria (Haslbeck & Ryan, 2019). For those systems we may be interested in identifying which intervention should be applied in order to move the system from one equilibrium position to another. We hope that the current paper will serve as the groundwork for such future developments.

Finally, a further simplification we made in the current paper was to ignore the role of uncertainty in parameter estimates. Previous research has shown that existing centrality measure estimates can be somewhat unstable as they reflect the sums of numerous parameter estimates (Epskamp, Borsboom, & Fried, 2018), an issue which is likely shared by the CT centrality measures estimated here. In

[7] `https://osf.io/9sgdn/`

practice the uncertainty around those parameters can be obtained in a relatively straightforward way using standard Bayesian approaches (as implemented in *ctsem*) or by bootstrapping. In line with this concern regarding parameter uncertainty, researchers should exercise caution when extrapolating from drift matrix parameters to lagged effects at different intervals. In the main text we assumed a true underlying CT-VAR process, and so could derive how lagged parameters vary and evolve over a range of time-intervals. In practice, the CT-VAR must be estimated from data, which likely contains a relatively limited range of time-intervals between observations. This means that any inference to lagged relationships outside of the observed intervals is a form of model extrapolation and is prone to inaccuracies if the model is misspecified Although the CT model at least allows one to explore what these lagged relationships are expected to look like at any interval, more research is needed to assess the robustness of inferences to intervals outside the data in practice.

In conclusion, while network analysis is a promising conceptual framework for psychopathology, we believe that to move forward in this field we need to become more sophisticated in how we think about the dynamics of the underlying system and how we conceptualize and learn about interventions in those systems. In the current paper we have established the first steps in using differential equation models of the underlying dynamic system to conceptualize network structure and established how centrality and intervention targets can be approached from an interventionist perspective. The approach in the current paper represents a first principled step to moving current practice forward, by placing the emphasis on the dynamic part of dynamic network analysis and placing interventions at the heart of centrality-based metrics.

Appendix 4.A Centrality Measures as Summaries of Path-specific Effects

In this appendix we show how path-specific effects in DT-VAR models are related to three popular centrality measures calculated from DT-VAR networks. For the measures typically interpreted as quantifying the total and direct influence of a variable (i.e. both Expected Influence measures), this relationship is quite straightforward, while for the popular indirect influence measure Betweenness Centrality, the relationship with path-tracing quantities is much farther removed.

In Table 4.1 we provide the formula and description of the three centrality measures we consider in the main text. These are expressed in terms of lagged regression parameters ϕ_{ji}, which represent the lagged effect from process i to process j (i.e., it is the element on the jth row and ith column of the matrix $\mathbf{\Phi}$). The-right hand column of Table 4.1 describes how these calculations relate to path-tracing quantities from the SEM literature. Note that the Expected Influence measures were originally developed for undirected networks (Robinaugh et al., 2016), and so, despite the active applications of those measures for directed networks (e.g. Kaiser & Laireiter, 2018) their precise definition for direct networks is left somewhat ambiguous. For instance, the popular packages *qgraph* (Epskamp et al., 2012) and *networktools* (Jones, 2018) differ slightly in how One-Step Expected Influence is calculated, with the former excluding diagonal elements (i.e. auto-regressive effects) as is common for DT-VAR centrality measures, while the latter includes those elements. The definitions we give here to the One-Step and Two-step Expected Influence measures ($EI_i^{(1)}$ and $EI_i^{(2)}$) omit relationships a variable has with itself either one or two occasions later respectively. We believe this is in keeping with the spirit of how these measures are defined for undirected networks, and allows us to maintain the standard interpretation of centrality measures as reflecting a type of relationship the target variable shares with all *other* variables in the model.

From Table 4.1 we can see that $EI_i^{(1)}$, which is typically interpreted as a summary of direct effects, is in fact the sum of lag-one direct effects of $Y_{i,\tau}$ on all other variables at the next occasion (that is, excluding the auto-regressive direct effect of Y_i on itself at the next occasion). The $EI_i^{(2)}$ measure, which is typically interpreted as reflecting the total influence of a variable, comprises two separate parts. The first part is the sum of lag-two total effects, following standard path-tracing rules, and excluding the total effect of a variable on itself two occasions later. The second part is the $EI_i^{(1)}$ measure for that variable. As such, $EI_i^{(2)}$ measure is a mix of total and direct effects at both lags.

Finally, the Betweenness Centrality measure BC_i, typically interpreted in terms of indirect effects, is only tenuously related to path-tracing quantities. In SEM approaches researchers are typically interested in mediators of indirect effects, where the size of an indirect effects is defined by the product of the component pathways (i.e., path-tracing rules). If we have many indirect pathways, and many potential mediators, we may wish to know which specific indirect ef-

Network Measure	Formula	Description
$EI_i^{(1)}$ *One-Step Expected Influence*	$\sum_{j \neq i}^{p} \phi_{ji}$	Sum of lag 1 *direct effects* $Y_{i,\tau} \rightarrow Y_{j,\tau+1}$ $\forall j \neq i$
$EI_i^{(2)}$ *Two-Step Expected Influence*	$\sum_{j}^{p} \phi_{ji} \sum_{k \neq i}^{p} \phi_{kj} + EI_{1i}$	Sum of *total effects* at lag 2 $Y_{i,\tau} \rightarrow Y_{j,\tau+1} \rightarrow Y_{k,\tau+2}$ $\forall k \neq i$ plus *direct effects* at lag 1
BC_i *Betweenness Centrality*	$M_{jk}(i) = 1$ iff $Y_i \in d(jk)$ $\sum \sum_{k \neq j \neq i}^{p} M_{jk}(i)$ where $d(jk)$ is $max\{\phi_{hj} + \cdots + \phi_{kh}\}$	Counts how often Y_i is a mediator on the shortest *network-path* $Y_{j,\tau} \rightarrow \ldots Y_{i,\cdot} \cdots \rightarrow Y_{k,\tau+q}$

Table 4.1: Relationship between different network metrics and path-tracing quantities, in the context of a VAR(1) model with p variables, regression coefficient matrix $\mathbf{\Phi}$, and corresponding dynamical network with weights matrix $\mathbf{\Phi}^T$.

fect is strongest, and in turn, how often a specific variable acts as a mediator of these strongest indirect effects. It seems that this is how psychological researchers using the BC_i measure typically interpret it (e.g., Bringmann et al., 2013, 2015; David, Marshall, Evanovich, & Mumma, 2018). However, the actual calculation of this measure differs greatly from the mediator-based metric described above. Specifically, instead of identifying the largest indirect effect, Betweenness is based on the identification of the *shortest network-path* between two variables ($d(jk)$). The length of this network-path is based on the inverse of the *sum* rather than the *product* of the individual pathways: While large SEM paths are those where multiplying each individual part leads to a high number, we say that short network-paths are those where the sum of each individual part leads to a small number. Similar to standard path-tracing nomenclature, these network-paths can be either direct (e.g. $Y_{j,\tau} \rightarrow Y_{k,\tau+1} = \phi_{kj}$) or indirect (e.g. $Y_{j,\tau} \rightarrow Y_{i,\tau+1} \rightarrow Y_{k,\tau+2} = \phi_{ij} + \phi_{ki}$) and each path may span a different number of measurement occasions. The Betweenness Centrality of Y_i is found by first calculating all the shortest paths between all pairs of variables, and then counting how often Y_i lies on that shortest path. It is clear then that, despite how this measure is interpreted the relationship of Betweenness Centrality with path-specific effects is much less direct than for the other measures considered above.

Appendix 4.B Centrality Values DT Stress-Discomfort System

In this appendix we present the centrality metrics for the half-hour network ($\boldsymbol{\Phi}(\Delta t = 0.5)$) and the one-hour network ($\boldsymbol{\Phi}(\Delta t = 1)$) as discussed in section 4.2. These are shown in Table 4.2.

	$\Delta t = 0.5$			$\Delta t = 1$		
	$EI^{(2)}$	$EI^{(1)}$	BC	$EI^{(2)}$	$EI^{(1)}$	BC
Stress	-0.025	-0.029	**1**	0.245	0.274	0
Anxiety	**0.035**	**0.038**	**1**	-0.071	-0.109	**3**
Self-Conscious	-0.024	-0.027	0	0.125	0.152	0
Physical Discomfort	0.008	-0.002	**1**	**0.555**	**0.557**	0

Table 4.2: Two-Step Expected Influence ($EI^{(2)}$), One-Step Expected Influence ($EI^{(1)}$) and Betweenness Centrality (BC) for each of the four variables in the half-hour ($\boldsymbol{\Phi}(\Delta t = 0.5)$) and one-hour ($\boldsymbol{\Phi}(\Delta t = 1)$) networks. These measures are based on the lagged parameter matrices displayed in Figure 4.1. In each column, the largest centrality values are highlighted in bold.

Appendix 4.C The Matrix Exponential as Path-Tracing

In this appendix we describe in more detail the relationship between the CT-VAR or Ornstein-Uhlenbeck model, and the notion of path-tracing effects. The CT-VAR model is the integral form of the first-order stochastic differential equation (SDE) model, defined as

$$\frac{d\boldsymbol{Y}(t)}{dt} = \boldsymbol{A}\boldsymbol{Y}(t) + \boldsymbol{W}(t) \tag{4.15}$$

where $\boldsymbol{A}$ is the drift matrix which regresses the derivative on the value of the process at that moment in time, and $\boldsymbol{W}(t)$ represents the stochastic innovation part of the system, also referred to as a Wiener process (which is often denoted $\boldsymbol{G}\frac{d\boldsymbol{W}(t)}{dt}$, cf. Oud & Jansen, 2000; Voelkle et al., 2012; Voelkle & Oud, 2013). The elements of the drift matrix encode direct dependencies between time-varying processes, with a_{ij} representing the direct effect of $Y_j(t)$ on the rate of change of $Y_i(t)$.

The first-derivative $\frac{d\boldsymbol{Y}(t)}{dt}$ is defined as the change in value of $\boldsymbol{Y}(t)$ over the time-interval $t+s$, as the value of s approaches zero

$$\frac{d\boldsymbol{Y}(t)}{dt} = \lim_{s \to 0} \frac{\boldsymbol{Y}(t+s) - \boldsymbol{Y}(t)}{s}$$

which means that the deterministic part of the first-order differential equation

(i.e., ignoring the stochastic innovation part) can be re-written as

$$\lim_{s\to 0} \frac{\boldsymbol{Y}(t+s) - \boldsymbol{Y}(t)}{s} = \boldsymbol{A}\boldsymbol{Y}(t)$$

Re-arranging, we can come to an expression for the relationship between $\boldsymbol{Y}(t)$ and $\boldsymbol{Y}(t+s)$, as $s \to 0$

$$\begin{aligned}\boldsymbol{Y}(t + \lim_{s\to 0} s) &= \lim_{s\to 0} s \times (\boldsymbol{A}\boldsymbol{Y}(t)) + \boldsymbol{Y}(t) \\ &= (\boldsymbol{I} + \boldsymbol{A}\lim_{s\to 0} s)\boldsymbol{Y}(t)\end{aligned}$$

that, is, an expression of the differential equation model as an auto-regressive model of measurements spaced very closely in time. Thus, the auto-regressive and cross-lagged relationships between waves spaced an infinitesimally small time-interval apart (i.e. the moment-to-moment lagged relationships) are given by $\boldsymbol{I} + \boldsymbol{A}\lim_{s\to 0} s$.

Now take it that we are interested in finding an expression relating two observed waves of variables $\boldsymbol{Y}(t)$ and $\boldsymbol{Y}(t+\Delta t)$. We can think of s as a very small fraction of Δt, that is,

$$s = \frac{\Delta t}{n}$$

such that as $n \to \infty$ we get $s \to 0$. This means that we can re-express the relationship between waves spaced an infinitely small time-interval apart as

$$\boldsymbol{Y}(t + \lim_{n\to\infty} \tfrac{\Delta t}{n}) = (\boldsymbol{I} + \boldsymbol{A} \lim_{n\to\infty} \tfrac{\Delta t}{n})\boldsymbol{Y}(t).$$

Now, if we conceptualize the CT-VAR as a path model, as depicted in Figure 4.2 in the main text, then we can find an expression to relate $\boldsymbol{Y}(t)$ and $\boldsymbol{Y}(t+\Delta t)$ by a simple application of path-tracing rules (Bollen, 1987) . That is, we can trace through the $lim_{n\to\infty} n$ latent waves in-between those two occasion, by taking the appropriate power of the moment-to-moment lagged-effects matrix $\boldsymbol{I} + \lim_{n\to\infty} \boldsymbol{A}\frac{\Delta t}{n}$. This path-tracing operation gives us

$$\begin{aligned}\boldsymbol{Y}(t+\Delta t) &= lim_{n\to\infty}(\boldsymbol{I} + \lim_{n\to\infty} \boldsymbol{A}\tfrac{\Delta t}{n})^n \boldsymbol{Y}(t) \\ &= lim_{n\to\infty}\Big\{(\boldsymbol{I} + \tfrac{1}{n}\boldsymbol{A}\Delta t)^n\Big\}\boldsymbol{Y}(t)\end{aligned}$$

By definition, the first term on the right-hand side is exactly the matrix exponential (cf. Abadir & Magnus, 2005, p.250)

$$\boldsymbol{e}^{\boldsymbol{A}\Delta t} = lim_{n\to\infty}\Big\{(\boldsymbol{I} + \tfrac{1}{n}\boldsymbol{A}\Delta t)^n\Big\}, \tag{4.16}$$

giving us

$$\boldsymbol{Y}(t+\Delta t) = \boldsymbol{e}^{\boldsymbol{A}\Delta t}\boldsymbol{Y}(t), \tag{4.17}$$

which gives us the deterministic part of the CT-VAR(1) model.

This derivation shows that the CT-VAR model can be seen as a path model, where the lagged relationships are defined as total effects resulting from path-tracing through an $n \to \infty$ latent waves. Thus, any DT cross-lagged parameter matrix $\boldsymbol{\Phi}(\Delta t) = \boldsymbol{e}^{\boldsymbol{A}\Delta t}$ should be interpreted as reflecting *total effects* relative to the CT-VAR model.

Appendix 4.D Path-Tracing in CT models

In this appendix we describe the calculation of path-specific effects for the CT-VAR model based on path-tracing rules. Both Deboeck and Preacher (2016) and Aalen et al. (2016) describe a method for calculating direct, indirect and total effects in a CT-VAR model, which follow the path-tracing rules laid out by, among others, Bollen (1987). However, these authors only discuss path-tracing with respect to a lower-triangular tri-variate drift matrix, that is, a drift matrix with only three variables and without reciprocal lagged relationships. Here we generalize these path-tracing definitions to drift matrices of arbitrary structure and number, following path-tracing principles. In large part we follow the methods described by the original authors, except in the case of the indirect effect, where non-triangular drift matrices must be approached differently than was done in the simpler scenario of a lower triangular matrix.

To find the path-tracing total effect of $Y_i(t)$ on $Y_j(t+\Delta t)$, which we will here denote $TE_{ij}(\Delta t)$, we simply take the element in the jth row, ith column of the matrix exponential of the drift matrix:

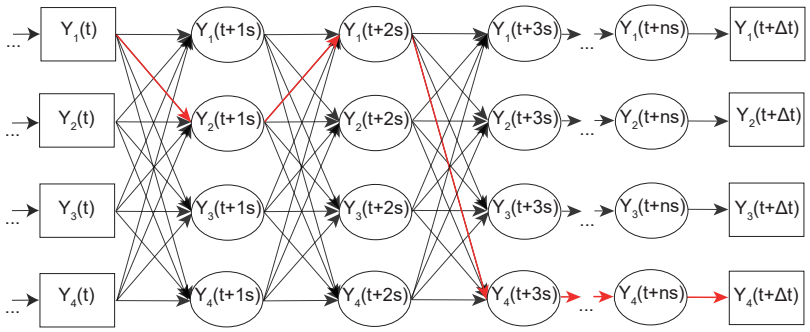
$$TE_{ij}(\Delta t) = \boldsymbol{e}^{\boldsymbol{A}\Delta t}{}_{[ji]} \tag{4.18}$$

This follows from the interpretation of the matrix exponential term $\boldsymbol{e}^{\boldsymbol{A}\Delta t}$ as a path-tracing operation, relative to the moment-to-moment auto-regressive effects matrix ($\boldsymbol{I} + \boldsymbol{A}\lim_{n\to\infty}\frac{\Delta t}{n}$) (described in Appendix 4.C). In Figure 4.7(a) we show a four-variable CT-VAR model with a full $\boldsymbol{A}$ matrix in path-model form, with $n \to \infty$ latent values of the processes in between measurement occasions, spaced at intervals of $s \to 0$. From this it is clear that tracing a path from, for instance, $Y_1(t)$ to $Y_4(t+\Delta t)$ includes paths through latent values of Y_1 and Y_4, (e.g. $Y_1(t) \to Y_1(t{+}1s) \to Y_1(t{+}2s) \to Y_4(t{+}3s) \to \cdots \to Y_4(t{+}\Delta t)$) as well as paths through latent values of Y_2 and Y_3 ($Y_1(t) \to Y_2(t+1s) \to Y_3(t+2s) \to Y_4(t+3s) \to \cdots \to Y_4(t+\Delta t)$). As such, we can interpret this total effect as constituted of all possible pathways linking $Y_1(t)$ and $Y_4(t+\Delta t)$, as is the standard interpretation of a total effect.

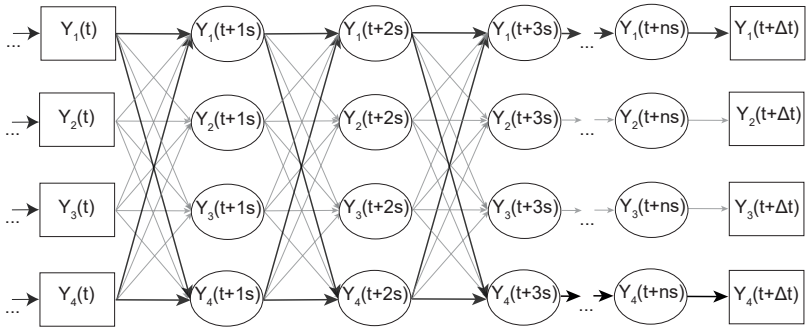
In order to find the path-tracing *direct* effect from $Y_i(t)$ to $Y_j(t+\Delta t)$ relative to some mediator variable(s) Y_k, Deboeck and Preacher (2016) state that the drift matrix should first be altered so that the parameters which make up the indirect pathways are omitted. We can alter the drift matrix to achieve this by setting the kth row and column elements of $\boldsymbol{A}$ to zero, yielding a drift matrix containing only direct relationships between Y_i and Y_j, which we will denote $\boldsymbol{A}^{(D[-k])}$. The path-tracing direct effect is then found by applying the matrix exponential function to the altered drift matrix.

$$DE_{ij\cdot k}(\Delta t) = \boldsymbol{e}^{\boldsymbol{A}^{(D[-k])}\Delta t}{}_{[ji]}. \tag{4.19}$$

For example, for a four-variable system, to define the path-tracing direct effect of $Y_1(t)$ to $Y_4(t+\Delta t)$ relative to the mediators Y_2 and Y_3 we would need to alter the



(a) CT-VAR path model representing total effect pathways



(b) CT-VAR path model representing direct effect pathways

Figure 4.7: Path-model representation of the four-variable CT-VAR model with full drift matrix $\boldsymbol{A}$. In the top panel the red pathway highlights a path which is included in the difference-method calculation of the indirect effect but omitted from the path-method calculation. The bottom panel shows the pathways which make up the direct effect, with indirect paths (removed from the altered drift matrix $\boldsymbol{A}^{(D[-k])}$) shaded in gray

drift matrix as follows

$$\boldsymbol{A} = \begin{array}{c} \\ Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{array} \begin{array}{c} \begin{array}{cccc} Y_1 & Y_2 & Y_3 & Y_4 \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} \end{array}, \quad \boldsymbol{A}^{(D[-2,-3])} = \begin{array}{c} \\ Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{array} \begin{array}{c} \begin{array}{cccc} Y_1 & Y_2 & Y_3 & Y_4 \end{array} \\ \begin{pmatrix} a_{11} & 0 & 0 & a_{14} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ a_{41} & 0 & 0 & a_{44} \end{pmatrix} \end{array}$$

This altered drift matrix defines a new path model, absent of any lagged relationships linking Y_1 to Y_4 through the mediators Y_2 and Y_3. This is displayed in Figure 4.7(b). Applying the matrix exponential function to this new drift matrix, it is clear that we only trace through direct pathways linking $Y_1(t)$ to $Y_4(t+\Delta t)$ (e.g. $Y_1(t) \to Y_1(t+1s) \to Y_1(t+2s) \to Y_4(t+3s) \to \cdots \to Y_4(t+\Delta t)$). This process is exactly equivalent to how Bollen (1987) describes the calculation of a direct effect

using matrix algebra.

To calculate the indirect effect for a lower-triangular drift matrix, both Deboeck and Preacher (2016) and Aalen et al. (2016) describe an operation by which the direct links are omitted from the drift matrix (in the four variable example, this would be a_{14} and a_{41}) before applying the matrix exponential term. We will refer to this as the *trace-method* of calculating an indirect effect. Alternatively, following path-tracing rules in linear models, we could define the indirect effect as the difference between the total and direct effect, which we will refer to as the *difference-method*. For a lower triangular drift matrix, both methods yield the same indirect effect (Deboeck & Preacher, 2016).

However, for non-triangular drift matrices, these definitions will not be equivalent. The reason again follows simple path-tracing rules. The difference method in this scenario quantifies all paths from $Y_i(t)$ to $Y_j(t+\Delta t)$ that pass through some latent value of Y_k. In contrast, the trace-method quantifies fewer paths, that is, all paths that pass through Y_k, but *do not* pass along any direct paths linking $Y_i(t)$ to $Y_j(t+s)$. In Figure 4.7(a) we have highlighted in red a pathway which is included as part of the difference-method indirect effect, but which is not included in the trace-method indirect effect.

In order to maintain the property that the total and direct effects sum to one another, and to allow an easier link to intervention-based definitions of indirect effects in Section 4.4.1 of the main text, we recommend the use of the difference method of calculating indirect effects. As such, we define the path-based indirect effect as

$$IEP_{ij}(\Delta t) = \boldsymbol{e}^{\boldsymbol{A}\Delta t}{}_{[ji]} - \boldsymbol{e}^{\boldsymbol{A}^{(D[-k])}\Delta t}{}_{[ji]}, \tag{4.20}$$

which is equivalent to the difference between the path-tracing total effect and the path-tracing direct effect described above.

Appendix 4.E Interventions and Path-Tracing in CT models

In this appendix we prove the equivalence between the intervention-based definitions of total, direct and indirect effects, and the path-tracing definitions of these quantities (described in Appendix 4.D), given the simplifying assumptions described in the main text.

4.E.1 Total Effect

We define the total effect of $Y_i(t)$ on $Y_j(t+\Delta t)$ as the expected change in value of $Y_j(t+\Delta t)$ given an acute intervention to set the value of $Y_i(t)$ from a constant, y_i^* to a new value y_i. We denote such a variable-setting operation using the do operator (Pearl, 2009), and so can express this total effect as

$$TE_{ij}(\Delta t) = \mathrm{E}\Big[Y_j(t+\Delta t) \;\Big|\; do(Y_i(t) = y_i)\Big] - \mathrm{E}\Big[Y_j(t+\Delta t) \;\Big|\; do(Y_i(t) = y_i^*)\Big] \tag{4.21}$$

By assuming that the system is fully observed (i.e., there are no unobserved confounders), and by assuming modularity, we can substitute the expected value of $Y_j(t+\Delta t)$ following an intervention $do(Y_i(t) = y_i)$ with the expected value given we observe $Y_i(t) = y_i$. This yields the expression

$$TE_{ij}(\Delta t) = \mathrm{E}\Big[Y_j(t+\Delta t) \;\Big|\; Y_i(t) = y_i\Big] - \mathrm{E}\Big[Y_j(t+\Delta t) \;\Big|\; Y_i(t) = y_i^*\Big]$$

Now we plug in the CT-VAR model for those expected values. Take it that $\boldsymbol{Y}(t)$ represents a column vector of variable values with ith element $Y_i(t) = y_i$. Using this, we can express the first expected value as

$$\mathrm{E}\Big[Y_j(t+\Delta t) \;\Big|\; Y_i(t) = y_i\Big] = \{e^{\boldsymbol{A}\Delta t}\boldsymbol{Y}(t)\}_{[j]}$$

that is, the jth element of the column vector obtained by multiplying the square matrix $e^{\boldsymbol{A}\Delta t}$ with the column vector $\boldsymbol{Y}(t)$. To obtain the second expectation, we take it that $\boldsymbol{Y}^*(t)$ represents a column vector of variable values with ith element $Y_i(t) = y_i^*$ but which is otherwise identical to $\boldsymbol{Y}(t)$.

$$\mathrm{E}\Big[Y_j(t+\Delta t) \;\Big|\; Y_i(t) = y_i^*\Big] = \{e^{\boldsymbol{A}\Delta t}\boldsymbol{Y}^*(t)\}_{[j]}$$

Taking the difference between these two expected values we obtain

$$\begin{aligned} TE_{ij}(\Delta t) &= \{e^{\boldsymbol{A}\Delta t}\boldsymbol{Y}(t)\}_{[j]} - \{e^{\boldsymbol{A}\Delta t}\boldsymbol{Y}^*(t)\}_{[j]} \\ &= e^{\boldsymbol{A}\Delta t} \cdot \Big(\boldsymbol{Y}(t) - \boldsymbol{Y}^*(t)\Big)_{[j]} \end{aligned}$$

Since the vectors differ only with respect to their ith element, we obtain

$$TE_{ij}(\Delta t) = {e^{\boldsymbol{A}\Delta t}}_{[ji]} \times (y_i - y_i^*) \tag{4.22}$$

where ${e^{\boldsymbol{A}\Delta t}}_{[ji]}$ is the element in the jth row and ith column of the matrix $e^{\boldsymbol{A}\Delta t}$. If we define the intervention as increasing the value of $Y_i(t)$ by one unit ($y_i - y_i^* = 1$), this yields an expression exactly equivalent to the path-tracing definition of a total effect given in Appendix 4.D.

4.E.2 Direct Effect

We define the direct effect of $Y_i(t)$ on $Y_j(t+\Delta t)$ as the expected change in value of $Y_j(t+\Delta t)$ given an acute intervention to set the value of $Y_i(t)$ from y_i^* to a new value y_i, while also intervening to keep the value of the mediator(s) Y_k fixed to a constant y_k at every moment in time in that interval. We denote this latter continuous intervention using the do operator over an interval of time as $do(\overline{Y_k(t+\Delta t)} = y_k$, and so express the direct effect as

$$\begin{aligned} DE_{ij\cdot\bar{k}}(\Delta t) = \mathrm{E}\Big[&Y_j(t+\Delta t) \;\Big|\; do(Y_i(t) = y_i), do(\overline{Y_k(t+\Delta t)} = y_k)\Big] \\ -\mathrm{E}\Big[&Y_j(t+\Delta t) \;\Big|\; do(Y_i(t) = y_i^*), do(\overline{Y_k(t+\Delta t)} = y_k)\Big] \end{aligned} \tag{4.23}$$

for some mediator(s) $k \in p$. Intuitively, if we want to block the indirect effect that acts through a mediator, we would need to ensure that either the mediator does not react to changes in the cause variable, or that it does not transmit information to the effect variable, or both. If we wish to achieve this by intervening on a variable, it is straightforward to see that we must do so by intervening to set the value of the mediator to a constant at every point in time between t and $t + \Delta t$.

As with the total effect derived above, the next step consists of plugging in the CT-VAR model for the expected values in this expression. However, note that due to the need to define the continuous intervention on the mediator $do(\overline{Y_k(t + \Delta t)} = y_k$ this proof is a little more involved than that of the total effect above (and relies on the strong assumption that applying such an intervention does not change the dynamics of the underlying process). To derive an expression for the direct effect, we first begin with the expression for the expected value of $\boldsymbol{Y}(t + \Delta t)$ given an acute intervention on the cause variable $Y_i(t)$, that is,

$$\mathrm{E}\Big[\boldsymbol{Y}(t + \Delta t) \;\Big|\; Y_i(t) = y_i\Big] = \boldsymbol{e}^{\boldsymbol{A}\Delta t}\boldsymbol{Y}(t)$$

one of the components of the *total effect* given above. Recall from the derivation in Appendix 4.C that we can write the CT-VAR model as describing lagged relationships over an infinitesimally small time interval $lim_{n\to\infty}\frac{\Delta t}{n}$, hereby referred to as the moment-to-moment relationship. This gives us

$$\mathrm{E}\Big[\boldsymbol{Y}(t + \lim_{n\to\infty}\tfrac{\Delta t}{n})\Big] = (\boldsymbol{I} + \boldsymbol{A}\lim_{n\to\infty}\tfrac{\Delta t}{n})\boldsymbol{Y}(t).$$

From now we will treat this expression as defining a moment-to-moment path model, as depicted in Figure 4.2 in the main text, and with a slight abuse of notation we will substitute $\lim_{n\to\infty}\frac{\Delta t}{n}$ for s, which we will define as a "moment" in time. We can express the expected value two "moments" after t as

$$\begin{aligned}\mathrm{E}\Big[\boldsymbol{Y}(t + 2s)\Big] &= (\boldsymbol{I} + \boldsymbol{A}s)\boldsymbol{Y}(t + s).\\ &= (\boldsymbol{I} + \boldsymbol{A}s)(\boldsymbol{I} + \boldsymbol{A}s)\boldsymbol{Y}(t)\\ &= (\boldsymbol{I} + \boldsymbol{A}s)^2\boldsymbol{Y}(t)\end{aligned}$$

where the second and third line follow by substituting in the expression for $\mathrm{E}\Big[\boldsymbol{Y}(t + s)\Big]$ given above.

Now, to define the direct effect we need to express the expected value of $\boldsymbol{Y}(t + 2s)$ given that we have intervened to set the current value of the mediator ($Y_k(t + 2s)$), the value of the mediator one "moment" previously ($Y_k(t + s)$), *and* the initial value of the mediator $Y_k(t)$ to zero. In order to derive such an expression, we introduce two simplifications here. First, since we are focusing on a linear model, and we are interested in the difference between two expected values in which in both cases the mediator Y_k is set to the same value y_k, the specific value we choose for y_k is irrelevant. For ease of notation we will therefore consider only an intervention by which y_k is equal to zero (i.e., the equilibrium position of Y_k). Second, to aid in our derivation, we will express the do operator in matrix algebraic terms. That is, we will represent the operation $do(Y_k(t) = 0)$ using a

transformation matrix $\boldsymbol{D}_{[-k]}$, a $p \times p$ matrix with zeros as off-diagonal elements, a zero on the kth diagonal element, and ones as the other diagonal elements. For instance, a 3×3 matrix $\boldsymbol{D}_{[-2]}$ would be given as

$$\boldsymbol{D}_{[-2]} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Pre-multiplying a column vector by the matrix $\boldsymbol{D}_{[-k]}$ reproduces the original column vector but with a zero as the kth element. That means that $\boldsymbol{D}_{[-k]}\boldsymbol{Y}(t)$ denotes the acute intervention $do(Y_k(t) = 0)$. Again, for ease of notation we will drop the $[-k]$ notation and leave it implied, that is, in the proof below, $\boldsymbol{D} = \boldsymbol{D}_{[-k]}$ unless otherwise specified.

Using this matrix-representation of the do operator, we can express the expected value of $\boldsymbol{Y}(t+2s)$ given that we have intervened to set the current value of the mediator ($do(Y_k(t+2s) = 0)$), the value of the mediator one "moment" previously ($do(Y_k(t+s) = 0)$), *and* the initial value of the mediator ($do(Y_k(t) = 0)$) to zero. Subsequently, since we repeat this acute intervention at every "moment" in time in an interval, we can describe it as a continuous intervention $do(\overline{Y_k(t+s)} = 0)$ that is, an intervention that is present for all possible time points in an interval. The expected value of $\boldsymbol{Y}(t+2s)$ given this continuous intervention can be written as

$$\begin{aligned} \mathrm{E}\Big[\boldsymbol{Y}(t+2s) \;\Big|\; do(\overline{Y_k(t+s)}) = 0\Big] &= \boldsymbol{D}(\boldsymbol{I} + \boldsymbol{A}s)\boldsymbol{D}(\boldsymbol{I} + \boldsymbol{A}s)\boldsymbol{D}\boldsymbol{Y}(t) \\ &= (\boldsymbol{DID} + \boldsymbol{DAD}s)^2\boldsymbol{Y}(t) \end{aligned}$$

Now, using the same substitutions as described in Appendix 4.C, we can express the expected value an arbitrary time interval Δt later, given that we intervene to set Y_k to zero at each of the $\lim_{n\to\infty}$ time points in that interval. This is given by

$$\mathrm{E}\Big[\boldsymbol{Y}(t+\Delta t) \;\Big|\; do(\overline{Y_k(t+\Delta t)}) = 0\Big] = lim_{n\to\infty}\Big\{(\boldsymbol{DID} + \boldsymbol{DAD}\tfrac{\Delta t}{n})^n\Big\}\boldsymbol{Y}(t)$$

Noting that $\boldsymbol{D}$ is an idempotent matrix, and that $\boldsymbol{DID} = \boldsymbol{DI} = \boldsymbol{ID}$, we can simplify this expression to

$$\mathrm{E}\Big[\boldsymbol{Y}(t+\Delta t) \;\Big|\; do(\overline{Y_k(t+\Delta t)}) = 0\Big] = lim_{n\to\infty}\Big\{(\boldsymbol{I} + \boldsymbol{DAD}\tfrac{\Delta t}{n})^n\Big\}\boldsymbol{D}\boldsymbol{Y}(t)$$

which, by the definition of the matrix exponential function simplifies to

$$\mathrm{E}\Big[\boldsymbol{Y}(t+\Delta t) \;\Big|\; do(\overline{Y_k(t+\Delta t)}) = 0\Big] = \boldsymbol{e}^{\boldsymbol{DAD}\Delta t}\boldsymbol{D}\boldsymbol{Y}(t) \tag{4.24}$$

where $\boldsymbol{DY}(t)$ ensures that the initial value of $Y_k(t)$ is set to zero.

Pre- and post-multiplying $\boldsymbol{A}$ by $\boldsymbol{D}_{[-k]}$ has the effect of setting the kth row and column of $\boldsymbol{A}$ to zero. Hence, the expression $\boldsymbol{e}^{\boldsymbol{DAD}\Delta t}$ is exactly equivalent to the path-tracing definition of the direct effect given in Appendix 4.D, that is,

$\boldsymbol{DAD} = \boldsymbol{A}^{(D[-k])}$. This implies that by plugging the above expression in for the expected values in the direct effect definition, we obtain

$$\begin{aligned} DE_{ij\cdot\bar{k}}(\Delta t) &= \boldsymbol{e}^{\boldsymbol{DAD}\Delta t}{}_{[ji]} \times (y_i - y_i^*) \\ &= \boldsymbol{e}^{\boldsymbol{A}^{(D[-k])}\Delta t}{}_{[ji]} \times (y_i - y_i^*) \end{aligned} \tag{4.25}$$

which shows that the effect on $Y_j(t + \Delta t)$ of an acute intervention to change $Y_i(t)$ combined with a continuous intervention to keep the mediator Y_k fixed is identical to the path-tracing direct effect.

4.E.3 Indirect Effect

It follows from the equivalence between path-tracing and intervention-based direct and total effects that the indirect effect, defined as a contrast between those two, can be calculated by taking the different in path-tracing definitions of each component effect, described in Appendix 4.D.

Chapter 5

Recovering Bistable Systems from Psychological Time Series

Abstract

Conceptualizing psychopathologies as complex dynamical systems has become a popular framework to study mental disorders. Especially bistable dynamical systems have received much attention, because their properties map well onto many characteristics of mental disorders. While these models were so far mostly used as stylized toy models, the recent surge in psychological time series data promises the ability to recover such models from data. In this paper we investigate how well popular (e.g., the Vector Autoregressive model) and more advanced (e.g., differential equation estimation) data analytic tools are suited to recover bistable dynamical systems from time series. Using a simulated high-frequency time series (measurement every six seconds) as an ideal case we show that while it is possible to recover global dynamics (e.g., position of fixed points, transition probabilities) it is difficult to recover the microdynamics (i.e., moment to moment interactions) of a bistable system. Repeating all analyses with a sampling frequency typical for Experience Sampling Method studies (measurement every 90 minutes) showed that the recovery of the global dynamics was still successful, but no microdynamics could be recovered. These results raise two fundamental issues involved in studying mental disorders from a complex systems perspective: first, it is generally unclear what to conclude from a statistical model about an underlying complex systems model; and second, if the sampling frequency is too low, it is impossible to recover microdynamics. In response to these results we propose a new modeling strategy based on substantively plausible dynamical systems models.

This chapter has been adapted from: Haslbeck, J. M. B.* & Ryan, O.* (under review). Recovering Bistable Systems from Psychological Time Series. Pre-print: https://psyarxiv.com/kcv3s/. Both JMBH and OR are considered joint first authors, with both contributing equally to this project.

5.1 Introduction

Conceptualizing psychopathologies as complex dynamical systems has become a popular framework to study mental disorders (e.g., Wichers, Wigman, & Myin-Germeys, 2015; Cramer et al., 2016; Borsboom, 2017). This framework is attractive because it acknowledges the fact that many mental disorders are massively multifactorial (e.g., Kendler, 2019), and because it allows one to specify powerful within-person dynamical systems models that capture many of the characteristics hypothesized for mental disorders. The central goal of this framework is to obtain such models to further our understanding of mental disorders, and allow us to develop and test more successful interventions.

The class of dynamic systems that has received most attention in this emerging literature is the class of *bistable systems* (e.g., Wichers et al., 2015; Cramer et al., 2016; Borsboom, 2017; Wichers, Schreuder, Goekoop, & Groen, 2019; van de Leemput et al., 2014; Nelson et al., 2017; R. Kalisch et al., 2019). The reason is that its behavior maps well on many phenomena observed in mental disorders: Bistable systems describe variables that have two stable states, which can be interpreted as different psychological states such as "healthy" or "unhealthy" (e.g., depressed). The stability landscape reflecting the dynamics of the system determines how easy it is to transition from one state to the other, and thereby offers a possible formalization of properties of the mental disorder, such as vulnerability or resilience to developing it (Scheffer et al., 2018). Bistable systems can also show sudden transitions from one state to another, thereby mapping well on, for example, bipolar disorder or the phenomenon of sudden gains and losses in psychotherapy (Stiles et al., 2003; Lutz et al., 2013).

In parallel, the realization that inferences from between-subjects data to within-person data are only possible under stringent assumptions (Molenaar, 2004; Hamaker, 2012) together with the increasing availability of psychological time series collected from mobile devices has led to a surge in studies aiming at recovering the within-person dynamics associated with mental disorders (e.g., Bringmann et al., 2013; Pe et al., 2015; A. J. Fisher et al., 2017; Snippe et al., 2017; Groen et al., 2019). This is an exciting development, because within-person time series potentially allow one to recover bistable systems and other dynamical systems from empirical data. This would be a major step forward for studying mental disorders as complex systems, because so far these models were only used as stylized toy models.

However, so far there has been no systematic investigation into to what extent dynamical systems models can in fact be recovered from psychological time series. To investigate this question for a given dynamical system and data analytic method, it has to be broken down into two parts. The first question is whether the method at hand can recover (some aspect of) a dynamical system *in principle*, that is, with "ideal" data (long time series, extremely high sampling frequency). If this is the case, the second question is whether the method also works with realistic data (shorter time series, much lower sampling frequency). In the present paper, we investigate both questions for a bistable dynamical system and a selection of the most popular (e.g., the Vector Autoregressive (VAR) model; Hamilton,

1994) and some more advanced (e.g., differential equation estimation; Boker, Deboeck, et al., 2010) methods. Specifically, we use a basic bistable dynamical system for emotion dynamics to simulate both an ideal time series with extremely high sampling frequency (measurement every six seconds) and a more realistic time series with a sampling frequency common for Experience Sampling Method (ESM) studies (measurement every 90 minutes). Using these time series, we evaluate how useful each method is for recovering bistable dynamical systems *in principle*, and how useful it can be *in practice* when analyzing realistic ESM time series.

We will show that the popular VAR model (and the Gaussian Graphical Model fitted on its residuals; Epskamp, Waldorp, et al., 2018) is in principle unable to recover the global dynamics (e.g., location and variance of stable states, frequency of transitions) and succeeds only in recovering some of the microdynamics (moment-to-moment interactions) of the true bistable system. However, descriptive statistics, data visualization and more flexible statistical models are able to capture the global dynamics. The only method that recovered the complete bistable system is an iterative model building procedure that directly estimates the system of differential equations (DEs). Reducing the sampling frequency from every six seconds to every 90 minutes affects the considered methods differently. The VAR model and its extensions no longer recover any microdynamics, and the DE-estimation procedure fails. However, descriptives, data visualization and appropriate statistical models still recover the global dynamics. These results raise two fundamental issues involved in studying mental disorders from a complex systems perspective. First, it is generally unclear what to conclude from a statistical model about an underlying complex systems model. Second, if the sampling frequency is too low, it is impossible to recover microdynamics. In response to these findings, we outline a different research strategy to arrive at dynamical systems models for mental disorders: Proposing initial formal models which can subsequently be scrutinized and developed by deriving data implications that can be tested empirically. We will show that in this process many of the presented methods are instrumental to testing predictions of the formal model and thereby triangulating the formal model that captures the true dynamical system best.

Our paper is structured as follows. In Section 5.2 we introduce a simple bistable dynamical system for emotion dynamics, discuss its dynamics and characteristics, and describe how we generate the ideal and the more realistic time series from it. We use the ideal data (measurement every six seconds) in Section 5.3 to evaluate for each method to which extent it can recover a bistable dynamical system. Next, in Section 5.4 we evaluate the same methods but using the time series with a sampling frequency that matches typical ESM studies (measurement every 90min). Finally, in Section 5.5 we discuss the implications of our results for the framework of empirically studying mental disorders from a complex systems perspective, and outline a new research strategy based on formal modeling, which avoids shortcomings of a purely data analytic approach.

5.2 Bistable Emotion System as Data-Generating Model

In this section we present a bistable dynamical system and describe its dynamics (Section 5.2.1), show how we generate data from this system (Section 5.2.2) and discuss its qualitative characteristics (Section 5.2.3).

5.2.1 Model Specification

Bistable dynamical systems are typically formalized within the framework of differential equations (e.g., Hirsch, Smale, & Devaney, 2012; Strogatz, 2015) and so we too adopt this framework to describe the data-generating mechanism. Our goal is to provide an accessible first investigation of how well bistable systems can be recovered from psychological time series and therefore we use one of the simplest multivariate bistable systems as a case study. Specifically, we choose a system with four variables that is a generalization of the classic Lotka Volterra model for competing species (e.g., H. I. Freedman, 1980) to four variables; a similar model was used by van de Leemput et al. (2014) who interpreted the four variables as positive and negative emotion variables, an interpretation we also adopt here. This model is capable of exhibiting two stable states: one in which positive emotions are high and negative emotions are low (the "healthy" state); and one in which the positive emotions are low and the negative emotions are high (the "unhealthy" state).

Note that different types of (bistable) dynamical systems will differ in how difficult they are to recover with a given method and type of time series, and much research is needed to map out the space of dynamical systems model classes, data analytic methods and types of time series. However, the intuition we rely on in the present paper is that if there are fundamental problems in recovering one of the simplest multivariate bistable systems, then these problems will be at least equally severe when recovering a more complicated bistable dynamical system.

The bistable system we use throughout this paper consists of two emotions with positive valence (Cheerful (x_1) and Content (x_2)) and two emotions with negative valence (Anxious (x_3) and Sad (x_4)). The dynamics of the system is defined by the stochastic differential equations

$$\frac{dx_i}{dt} = r_i x_i + \sum_{j=1}^{4} C_{ij} x_j x_i + a_i + \epsilon_i \ , \tag{5.1}$$

where r_i can be thought of as the main effect of an emotion on itself over time, that is, the effect of x_i on its own rate of change. This parameter is set to 1 for positive emotions, and will be varied between $r_3, r_4 \in [0.9, 1.1]$ for negative emotions. We interpret the variations in r_3, r_4 as being related to stress: Higher stress means that the effects of a high degree of negative emotion stays in the system longer. The matrix C represents the dependencies between emotions in the form of *interaction effects*

$$C = \begin{bmatrix} -0.2 & 0.04 & -0.2 & -0.2 \\ 0.04 & -0.2 & -0.2 & -0.2 \\ -0.2 & -0.2 & -0.2 & 0.04 \\ -0.2 & -0.2 & 0.04 & -0.2 \end{bmatrix}.$$

The interactions in the matrix C show that emotions of the same valence reinforce each other, while emotions of different valence suppress each other. For example, $C_{12} = 0.04$ indicates that the rate of change of x_1 (Cheerful) depends on the product of x_1 and x_2 (Content) weighted by 0.04. Similarly, $C_{13} = -0.2$ indicates that the rate of change of x_1 depends on the product of x_1 and x_3 (Anxious) weighted by -0.2. The diagonal elements are quadratic effects: For example, $C_{11} = -0.2$ indicates that the rate of change of x_1 depends on the product $x_1 x_1 = x_1^2$. Note that we choose the matrix C to be symmetric purely for the sake of simplicity. Since we aim to specify the simplest possible bistable system, we specify that all within-valence effects (e.g., C_{12} and C_{34}) are equal to 0.04 and all between-valence interaction effects (e.g., C_{13} and C_{24}) and quadratic effects (C_{ii}) are equal to -0.2.

We interpret $x_i = 0$ as the absence of positive/negative emotion, and therefore do no not allow emotions to become negative. We ensure this with high probability by setting the constant $a_i = 1.6$ for all i. The Gaussian noise term ϵ_i has a mean of zero and a fixed standard deviation σ and represents short-term fluctuations in emotions due to the environment the system interacts with. Note that we used the same parameterization as van de Leemput et al. (2014), except that in our model we use an additive noise term instead of a multiplicative noise term for simplicity and set all $a_i = 1.6$.

Due to the symmetries in C, r and a, emotions with the same valence are exchangeable. We can therefore describe the dynamics of the 4-dimensional system using a 2-dimensional system consisting of one dimension for positive emotions and one dimension for negative emotions (for details see Appendix 5.A). Figure 5.1 illustrates the dynamics of the deterministic part (i.e., with $\epsilon_i = 0$) of this model: Panel (a) displays the stable (solid lines) and unstable (dashed lines) fixed points for positive (green) and negative (red) emotions, as a function of stress. For example, for a low stress level of 0.9 there is only a single fixed point: the positive emotions (PE) have the value 5.28 and the negative emotions (NE) have the value 1.15. We therefore also refer to this fixed point as the healthy state. If the stress level remains unchanged, the system will always end up at this fixed point, no matter how one chooses the starting values. This dynamic is illustrated in the corresponding vector field in panel (b). The arrows depict the partial derivatives with respect to the two emotions and therefore describe the linearized dynamics at a given point in the 2-d space. The vector field shows us that whichever initial values we choose, the system will always end up at the fixed point at (PE = 5.28, NE = 1.15). Thus, the system with stress = 0.9 describes a person whose emotions can be changed by external influences, but eventually always returns to the healthy state of having strong positive emotions and weak negative emotions. The solid lines in panel (b) indicate the values of positive and negative emotions for which the two differential equations are zero. At the intersections of those

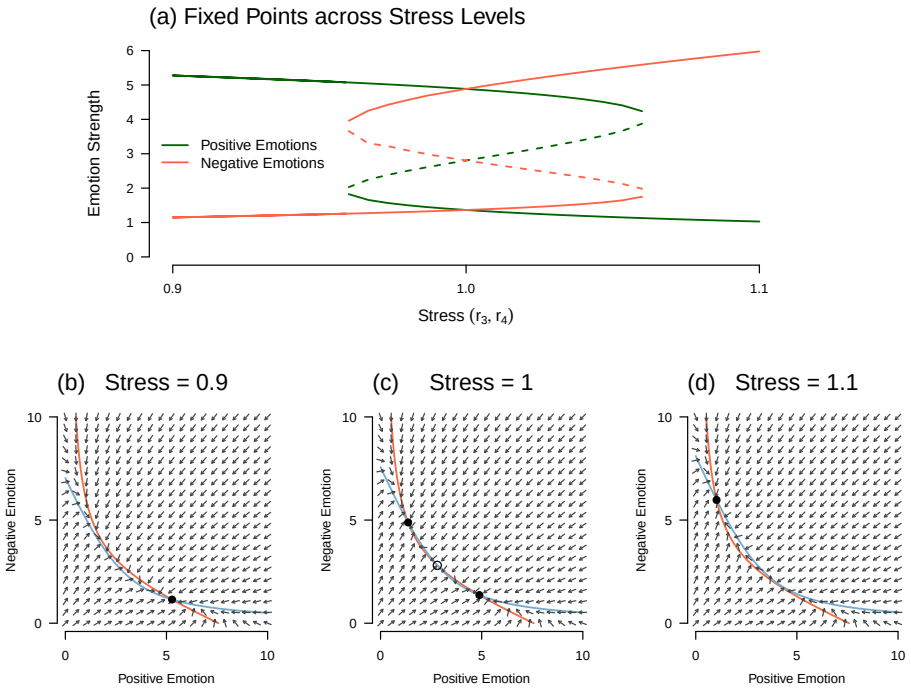


Figure 5.1: The dynamics of the bistable system we will use as the data-generating model throughout the paper. Panel (a) shows the fixed points of the deterministic part of the model as a function of stress, operationalized by the rate of change of the negative emotions. Solid lines indicate stable fixed points and dashed lines indicate unstable fixed points. Panels (b), (c) and (d) show the vector fields of the system for the stress values $r_3, r_4 = 0.9, 1$ and 1.1. Solid points indicate stable fixed points and empty points indicate unstable fixed points. The solid lines indicate the values at which derivative of positive emotion (orange) and negative emotion (light blue) is equal to zero. At the points at which the two lines meet, both derivatives are equal to zero and the system remains in this (stable) state.

lines both differential equations are equal to zero, which means that the system does not change anymore, which is the definition of a fixed point.

Panel (a) of Figure 5.1 shows that when increasing stress from 0.9 until around 0.95, the stable fixed point changes *quantitatively*: The value of positive emotion value decreases, and the value of negative emotion value increases. However, from around 0.95 on the dynamics of the system change *qualitatively*: the system now has three fixed points. For example, at stress = 1, the fixed points are (PE = 4.89, NE = 1.36), (PE = 2.80, NE = 2.80), and (PE = 1.36, NE = 4.89). The first fixed point is the stable healthy fixed point we also observed for values smaller than 0.9. The second fixed point is an unstable fixed point. Specifically, it is a saddle point, because the arrows in the vector field flow towards this fixed point in one direction, but flow away in the other direction (Strogatz, 2015). The third fixed point is again stable, however, now negative emotions have a high value and positive emotions have a low value. We could call this fixed point the unhealthy

fixed point.

The presence of these three fixed points means that, if the system is initialized anywhere except on the diagonal, the system will end up at one of the two stable fixed points. This behavior is illustrated in panel (c), which shows the vector field of the system for stress = 1. We see that eventually all arrows point away from the unstable fixed point (PE = 2.80, NE = 2.80) and towards one of the two stable fixed points. Thus, the system will never converge to this point except if it is initialized exactly on the diagonal. For all other starting values, the system will converge to one of the two stable fixed points. For the particular case of stress = 1, starting values above the diagonal line will converge to the unhealthy fixed point (PE = 1.36, NE = 4.89), whereas starting values below the diagonal line will converge to the healthy fixed point (PE = 4.89, NE = 1.36). This system describes a person that starts out in the healthy (unhealthy) state, and always returns to the healthy (unhealthy) state after small outside influences. However, a large influence can push the person into the unhealthy (healthy) state, and now the person remains there until a large enough influence pushes her back into the healthy (unhealthy) state.

When increasing stress further until around 1.06, we observe again a quantitative change of the three fixed points: the negative emotions go up, and the positive emotions go down. However, from around 1.06 on the system changes again qualitatively. It now again exhibits only one fixed point, which is now the unhealthy fixed point. Thus, when stress is larger than around 1.06, the system will always converge to the unhealthy fixed point. This behavior is illustrated in panel (d), which depicts the vector field for the system with stress = 1.1. We see that there is only a single fixed point at (PE = 1.03, NE = 5.98) and the arrows show that the system will always converge to this point. This system describes a person that will always return to the unhealthy state, no matter how large of an outside influence is applied.

So far, we only discussed the deterministic part of the model, that is, our model with noise set to zero (i.e., with $\epsilon_i = 0$). Introducing noise changes the dynamics of the system, and how exactly it changes depends on the stress level. For low stress (below 0.95), the system will fluctuate around the healthy fixed point. For high stress (above 1.06), the system will fluctuate around the unhealthy fixed point. The interesting behavior is observed for stress values between 0.95 and 1.06: then, the system will fluctuate around one of the two fixed points, but occasionally the noise will be large enough to push the system to the other fixed point. The frequency of switching is a function of the distance between the two fixed points, the vector field between the two fixed points, and the variance of the Gaussian noise process ϵ_i. If the variance is low, the probability of a noise draw that is large enough to "push" the system to the other fixed point is small, and consequently the frequency of switching is low. In contrast, if the variance is high, the probability of a large enough noise draw to switch to the other fixed point is high, and consequently the switching frequency is high.

5.2.2 Generating Time Series from Bistable System

In the previous section we have shown that our dynamical system is bistable for stress values $(r_3, r_4) \in [0.96, 1.06]$. In the remainder of this manuscript we keep stress constant at stress = 1, and therefore study the bistable dynamical system with the dynamics displayed in panel (c) of Figure 5.1 and the fixed points described above. Apart from stress we chose all parameters as indicated in the previous section.

To obtain a plausible switching frequency for emotion dynamics we set the standard deviation of the Gaussian noise term $\sigma = 4.5$. Note that a system can be bistable, but the outside influences (the noise term) are so weak that the system switches very infrequently or not at all. In such cases the bistable system is more difficult (infrequent) or impossible (no switches) to recover. Thus, our choice of σ represents an ideal situation, and all presented methods will perform worse with a lower switching frequency.

In the remainder of this section we describe how we generated the two time series that we will use throughout the paper. First, we generate an "ideal" time series with an extremely high sampling frequency of 1 measurement every six seconds (Section 5.2.2.1). Second, we generate a more realistic time series with measurements every 90 minutes, a sampling frequency typical for ESM studies (Section 5.2.2.2).

5.2.2.1 Ideal Time Series

We generated data by computing the numerical solution to the model in Section 5.2.1 with stress = 1 on the interval $[0, 20160]$, using Euler's method (e.g., Atkinson, 1989) with a step size of 0.01. We interpret a time step of 1 as one minute, and therefore the time series spans two weeks ($60 \times 24 \times 14 = 20160$). We obtain a time series by sampling the numerical solution obtained via Euler's method 10 times per minute (or every six seconds). We therefore obtain the ideal time series with $20160 \times 10 = 201600$ measurements, which appears to switch between fixed points around 17 times.[1] Figure 5.2 displays this time series.

We choose this unrealistically ideal time series (two weeks, one measurement every six seconds, continuous response scale, no measurement error or missing values) with 201600 measurements to be able to study the usefulness of different data analytic methods *in principle*. That is, we study the usefulness of all methods on the *population level*, which is the situation in which we have infinitely many observations and sampling variation does not exist. With 201600 observations, in this setting we approximate "infinitely many" for all practical purposes.

Note that we would not be able to investigate how well different methods perform in principle, if we made the time series more realistic by choosing a shorter time interval or sampling it with a lower sampling frequency. In such a case we would not know whether a method cannot recover (an aspect of) the bistable system for fundamental reasons, or because the time series is too short

[1] The code to generate data and reproduce all analyses and results shown in this paper can be found at `https://github.com/jmbh/RecoveringBistableSystems/`.

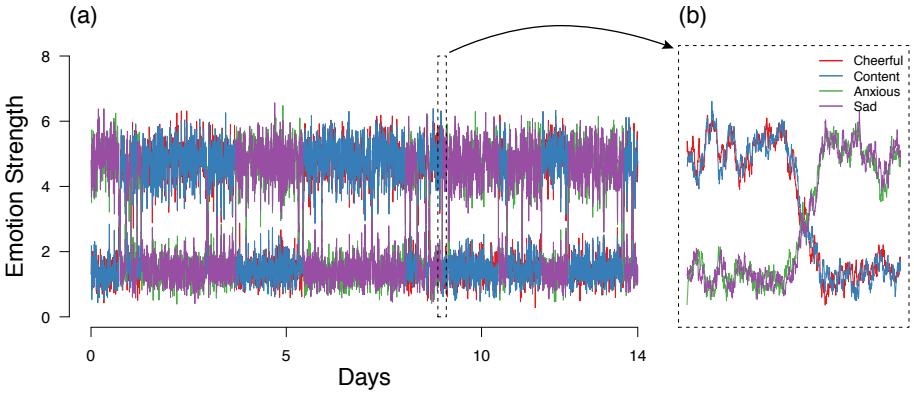


Figure 5.2: Panel (a) shows the ideal time series of the four emotion variables Cheerful, Content, Anxious and Sad. We see that the system switches 17 times between healthy and unhealthy state. Panel (b) displays the twelfth switch, which is a transition from the unhealthy to the healthy state, which occurs on day 9.

or the sampling frequency too low. We therefore first study all methods with the ideal time series in order to identify their fundamental limitations. In the second part of the paper, we make the time series more realistic by taking measurements at a sampling frequency that is typical for ESM studies. This will allow us to investigate the impact of sampling frequency on all methods. In the following section we describe how we generate this ESM time series.

5.2.2.2 Experience Sampling (ESM) Time Series

Clearly, the ideal time series is very different from time series data sets obtained from typical ESM studies. The perhaps two most important differences between the ideal time series and realistic time series are the measurement scale and the sampling frequency. With respect to the measurement scale, most ESM studies do not use a continuous response but, for example, a 7-point Likert scale. Regarding the sampling frequency, ESM studies investigating psychological variables typically do not measure more frequently than every 90 minutes (e.g., Bringmann et al., 2013; Pe et al., 2015; A. J. Fisher et al., 2017; Snippe et al., 2017; Groen et al., 2019). Thus, ESM time series have a much lower sampling frequency (every 90 minutes) than the ideal time series used above (every six seconds). To be able to explain possible drops in performance of certain methods when making the time series more realistic, we must only change one aspect of the time series. While the measurement scale can possibly be made near-continuous, there are certainly hard limits on how many times one can notify a person each day with an ESM questionnaire. We therefore consider the sampling frequency the more fundamental constraint in realistic data, and thus make it the focus of our investigation in Section 5.4.

Taking a measurement every 90 minutes in the two weeks of the original data

yields 224 measurements. This would mean that we would compare the "ideal" time series with 201600 measurements which essentially implies the absence of sampling variation to an ESM time series with 224 measurement which implies quite a considerable degree of sampling variation. Thus, any comparison would be confounded by the difference in the number of measurement points (i.e. sample size). To avoid this confound, we increase the measurement interval of the ESM time series to 1800 weeks, which ensures that the new ESM time series has exactly the same sample size as the ideal data ($\frac{224}{2} \times 1800 = 201600$). Thereby, we ensure that any drop in performance is a function of the lowered sampling frequency and cannot be explained by lower sample size (and higher sampling variation). Note that studying the performance of methods as a function of sample size (sampling variation) is of paramount importance to evaluate how useful a given method is in a realistic application. However, here we study the more fundamental question of the impact of reducing the sampling frequency to a level that is typical for ESM studies. We do this because if we find that a method is ill-suited to recover (some aspect of) a bistable system with a realistic sampling frequency on the population level (i.e., with infinite sample size), then it does not make sense to investigate the performance of the method in the less ideal scenario with realistic (small) sample sizes.

So far, we only discussed that we sample every 90 minutes. However, to emulate ESM measurements, we also need to formalize how exactly ESM questions measure the four emotion variables. This is far from trivial: questions in some ESM studies refer to the very moment of measurement and are phrased along the lines of "How cheerful do you feel right now?". Such measurements could be formalized by defining the measurement as the set of current values of the system (a "snapshot" of the system) at the measurement time. In contrast, other ESM studies refer to the time period since the last measurement. A question of this type could be phrased "How cheerful did you feel in the time since the last notification?". Such measurements could be formalized by defining the measurement as the average values of the system since the last measurement. However, many other measurement functions are also possible. In this paper we analyze the first type of ESM question, because its measurement function is the simplest. However, we also performed all analyses with the second kind of ESM question, and all our main conclusions also hold in this situation.

Figure 5.3 displays the two week long original time series (see also Figure 5.2 panel (a)) next to the ESM time series which was obtained by taking "snapshots" of the process at 90 min intervals. The ESM time series in panel (b) appears less dense, which is what we would expect since it contains only 1/900 of the time points of the ideal time series in panel (a). However, we see that the system is still bistable and that the location of and variance around the fixed points is largely the same. In Section 5.4 we will use this emulated "snapshot" ESM time series to try to recover the true bistable system using the same array of methods as in Section 5.3, in which we analyze the ideal time series.

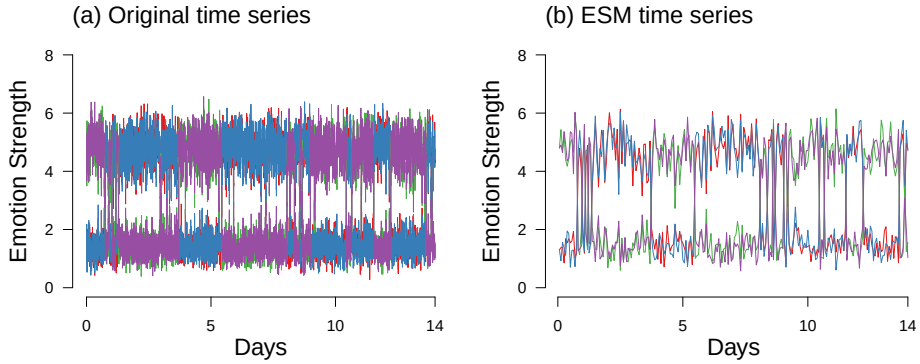


Figure 5.3: Panel (a) shows the original time series that was already shown in panel (a) of Figure 5.2. Panel (b) shows the ESM time series which was obtained by taking snapshots every 90 minutes in the series. Note that the ESM time series we analyze in Section 5.4 is much longer (1800 weeks) than the 14 day ESM time series shown here.

5.2.3 Qualitative Characteristics of the Model

In this section we discuss the key qualitative characteristics of the bistable system introduced in the previous section. We list these characteristics because most considered methods are models that are misspecified (i.e. they do not contain the true system as a special case). In such a situation one can only hope to recover some characteristics of the true system, and we therefore evaluate how well a method recovers the bistable system based on how well it recovers the following seven characteristics:

Global dynamics

1. Bistability (two stable fixed points)
2. Position of fixed points
3. Variability around fixed points
4. Frequency of transitions

Microdynamics

5. Suppressing effects between valences, reinforcing effects within valences
6. Relative size of suppressing/reinforcing effects
7. All parameters are independent of time and independent of variables outside the model

The first four characteristics describe the global dynamics of the dynamical system. The first is bistability, which means that the data-generating mechanism exhibits two stable fixed points. This is the case for the data-generating mechanism with stress set to 1, which we use to generate data from and aim to recover throughout the paper (see Figure 5.2, panels (a) and (c)). The second characteristic is the position of the fixed points, which are at (PE = 4.89, NE= 1.36) for the healthy fixed point, and (PE = 1.36, NE= 4.89) for the unhealthy fixed point. Third, we consider the variability around the different fixed points. Figure 5.2 shows that, for both fixed points, the variability of the emotions with lower values is smaller than the variability of the emotions with larger values. The fourth characteristic is the frequency of transitions between the area around the healthy fixed point and the area around the unhealthy fixed point. In the time series shown in Figure 5.2 we see that the system switches around 17 times.

The remaining three characteristics describe the microdynamics of the dynamical system. The fifth characteristic is that emotions of the same valence reinforce each other, while emotions of different valence suppress each other. The sixth characteristic is the fact that the size (absolute value) of the reinforcing effects (0.04) are smaller than the suppressing effects (0.2). The last characteristic is that all parameters in the system of differential equations are independent of time and independent of variables outside the model.

5.3 Recovering the Bistable System from Ideal Data

In this section we analyze the ideal time series to evaluate how well different methods recover the data-generating bistable system. The methods considered here primarily consist of the most popular models used in analyzing time series in clinical psychology and psychiatry, and some extensions thereof. In all but one instance, this entails the estimation of misspecified models, that is, models which do not contain the true bistable system as a special case. Thus we will focus our investigation on whether these models allow one to recover some of the characteristics of the true model, as outlined in Section 5.2.3.

We analyze each method in order of increasing complexity, moving from methods which may be helpful in recovering global characteristics alone to methods which are typically used with the aim of characterising the microdynamic structure. We begin by inspecting the time series using descriptive statistics (Section 5.3.1); in Section 5.3.2 we characterize the switching behavior in the system using a mean-switching Hidden Markov Model (Hosenfeld et al., 2015; Hamaker, Grasman, & Kamphuis, 2016). Next, in Section 5.3.3 we analyze the multivariate lag-0 (same time point) relationships using correlations and partial correlations with the popular Gaussian Graphical Model (GGM) (Epskamp, Waldorp, et al., 2018). In Section 5.3.4 we use the most popular approach to modeling microdynamics in experience sampling settings, the lag-1 Vector Auto-Regressive (VAR(1)) model (e.g. Bringmann et al., 2013; Pe et al., 2015; A. J. Fisher et al., 2017; Groen et al., 2019). Next, in Section 5.3.5 we evaluate the Threshold VAR model, an extension of the VAR model that allows the modeling of state-

switching behavior using time-varying parameters (Hamaker, Zhang, & van der Maas, 2009; Hamaker & Grasman, 2012). While all models so far are misspecified, we include one final method that is capable of recovering the full bistable system: a two-step model building approach based on direct estimation of differential equations from data, following the dynamic systems modeling approach of Boker, Deboeck, et al. (2010) and Chow (2019).

5.3.1 Descriptive Statistics

To get a rough overview of the behavior of the system, we inspect the time series plot of all four emotion variables shown above in Figure 5.2 panel (a). We see that at almost every time point the two positive emotion variables have high values around 5, and the two negative emotion variables have small values around 1, or the other way around. At the remaining time points, the variables seem to transition between those two states (see panel (b) in Figure 5.2). In addition, we see that the variables switch between states 17 times.

We can extract a considerable amount of information from simply inspecting the time series plot. There seem to be two stable states (fixed points), one in which positive emotions are high and negative emotions are low, and one in which the reverse is true. Further, we see that the variance is higher for the emotion with higher values, and we saw that the system switches around 17 times in the two week window. To get a more direct picture of the distribution of variables and possible fixed points we show the histograms for each variable in Figure 5.4:

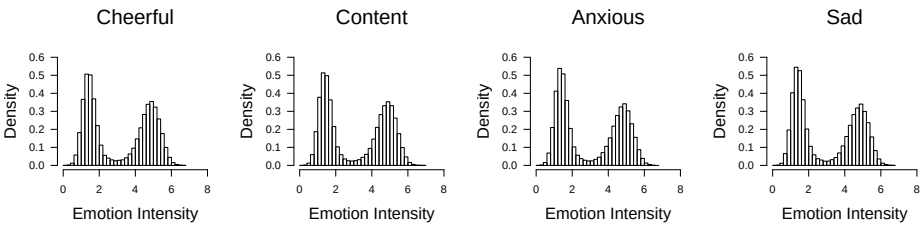


Figure 5.4: The histograms of the emotion intensity of the four modeled emotions Cheerful, Content, Anxious and Sad, for the ideal data.

We see that at most time points in the time series, each emotion either takes on values around 1 or around 5. This is what we would expect from inspecting the time series plot, however, the histograms give a more precise picture of the distributions and allow one to guess possible fixed points with greater precision. For instance, we could separate the two distributions (using a fixed threshold, or clustering algorithm) and take their means as estimates for the fixed points.

While eyeballing the data should be the first step in any time series analysis, the conclusions are subjective and do not allow us to quantify how certain we are about bistability and the switching frequency. We can quantify the observation that there are two states and that the system is switching between them by fitting a Hidden Markov Model (HMM) (e.g., Rabiner, 1989) to the data, which we will

do in the following section. Such quantification is especially valuable in more realistic situations, in which the two states are probably harder to separate than in our ideal simulated data.

5.3.2 Hidden Markov Model

In this section we fit a mean-switching Hidden Markov Model (HMM) in order to scrutinize the intuition that the system switches between two states and to quantify the switching frequency. The HMM models the observed data as consisting of K latent states or components, characterized by K multivariate Gaussian distributions, which may differ in their means μ_k and variances σ_k.[2] Each observation over time is drawn from one or other of these distributions, and the switching between these states is governed by a matrix of transition probabilities A. For more details about this model see Appendix 5.B.1.

Here we choose $K = 2$ components, and fit this model to our time series using the R-package *depmixS4* (Visser & Speekenbrink, 2010), obtaining the following parameter estimates

$$\hat{\mu}_1 = \begin{pmatrix} 1.47 \\ 1.46 \\ 4.71 \\ 4.71 \end{pmatrix}, \ \hat{\sigma}_1 = \begin{pmatrix} 0.41 \\ 0.40 \\ 0.63 \\ 0.62 \end{pmatrix}, \ \hat{\mu}_2 = \begin{pmatrix} 4.75 \\ 4.76 \\ 1.45 \\ 1.45 \end{pmatrix}, \ \hat{\sigma}_2 = \begin{pmatrix} 0.63 \\ 0.62 \\ 0.40 \\ 0.40 \end{pmatrix}, \hat{A} = \begin{pmatrix} 0.9996 & 0.0004 \\ 0.0004 & 0.9996 \end{pmatrix}.$$

We can see from the estimate $\hat{\mu}_1$ that in state 1 the means of positive emotions are low, and the means of negative emotions are high. We can therefore identify state 1 as the unhealthy state. We also see that the standard deviations of positive emotions are lower than for negative emotions in the unhealthy state which is what we already observed in the time series plot in Figure 5.2. Similarly, state 2 can be identified as the healthy state, with high means and standard deviations for positive emotions, and low for negative emotions. The transition matrix A indicates the probabilities of switching between states. We see that there is a very high probability for remaining in the same state ($\hat{A}_{11} = \hat{A}_{22} = 0.9996$), and a correspondingly low probability to switch states ($\hat{A}_{12} = \hat{A}_{21} = 0.0004$). This is what we would expect, because we take one measurement every six seconds, but the system changes states only a couple of times within the two week window. Multiplying the number of time points of the time series with the switching probability we obtain $201600 \times .0004 \approx 81$ switches, which is in the same order of magnitude of the eyeballed number of switches (17) reported in Section 5.2.2.

In addition to obtaining estimates of means and standard deviations of the two fixed points and the transition matrix, the HMM allows to predict the most likely state for each time point. We show the predicted states for the entire time series in Figure 5.5:

[2] In principle, the distributions may also differ with respective to their covariances, but in this analysis, we set all covariances to zero due to limitations of the software package used in estimation.

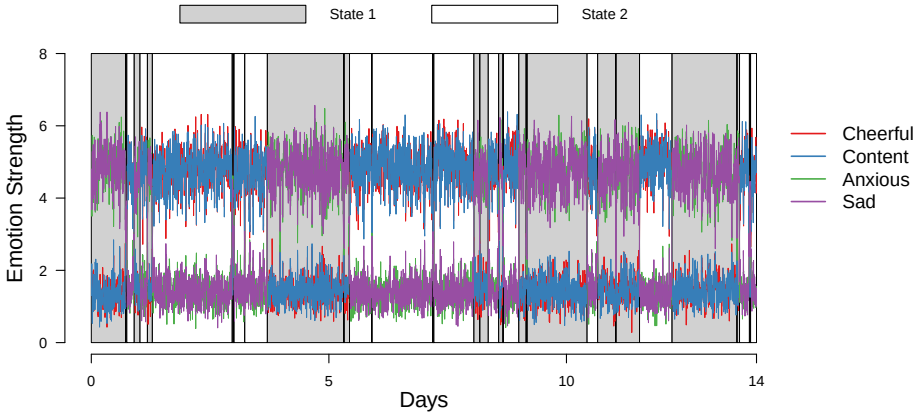


Figure 5.5: Time series of the four emotion variables, also shown in panel (a) of Figure 5.2, with background color indicating whether a given time point is assigned to the first or second component of the mean-switching HMM.

When inspecting the predicted states visually, it seems that the HMM captured the switches well. Next to the larger blocks in which the system stays at the same fixed point, it also identifies switches in which the system switches back and forth within only a few time points. These switches might have been missed when inspecting the time series visually alone.

Taking all results together, which characteristics of the bistable system did we recover with the HMM (based on the list in Section 5.2.3)? We obtained an estimate of the location (characteristic 2) and variance (characteristic 3) around two fixed points, which are very close to the healthy and unhealthy fixed points in the true bistable system. We also quantified the frequency of transitions in the transition matrix A. Since the transition frequency (characteristic 4) is not explicit in the true bistable model, there is no clear way to evaluate this estimate. However, the number of predicted transitions (81) is at least of the same order of magnitude as the number of transitions eyeballed from the entire time series (at least 17). Note that while a bistable HMM seems to fit the data well, we provided $K = 2$ as an *input* to the model, and therefore bistability (characteristic 1) cannot be considered a characteristic we recovered with this model. Instead of fixing a particular K, an optimal K can be obtained via model selection. However, in Appendix 5.B.2 we show that at least the standard approach to selecting K in mean-switching HMMs performs poorly since the data was not generated from a mean-switching HMM.

One additional way to visualize or ascertain how much of the true systems behavior a given model is able to capture is by generating new data from the estimated model parameters. In Figure 5.16 in Appendix 5.C.1 we generate a two week time series from the estimated mean-switching HMM and compare it to the original time series. We find that the data generated from the HMM is similar to the original data, except for two features: First, the system tends to switch between states somewhat more frequently, and second, there are no observations on the transitions between states.

The remaining three characteristics (5-7) are about the microdynamics of the true bistable system, that is, about how the components are related to each other. Clearly, the mean-switching HMM we used here cannot elucidate these characteristics since it does not model any dependencies between the four emotion variables. In the following sections we fit models that include such dependencies.

5.3.3 Lag-0 Relationships / Gaussian distribution

In this section we analyze the relationships between variables at the same time point. Figure 5.6 panel (a) displays the relationship between Content and Cheerful, two emotions of the same valence. We see that the observations cluster around two points, one close to (1, 1) with smaller variance, and one close to (5, 5) with larger variance. The red line indicates the best fitting regression line (correlation 0.98).

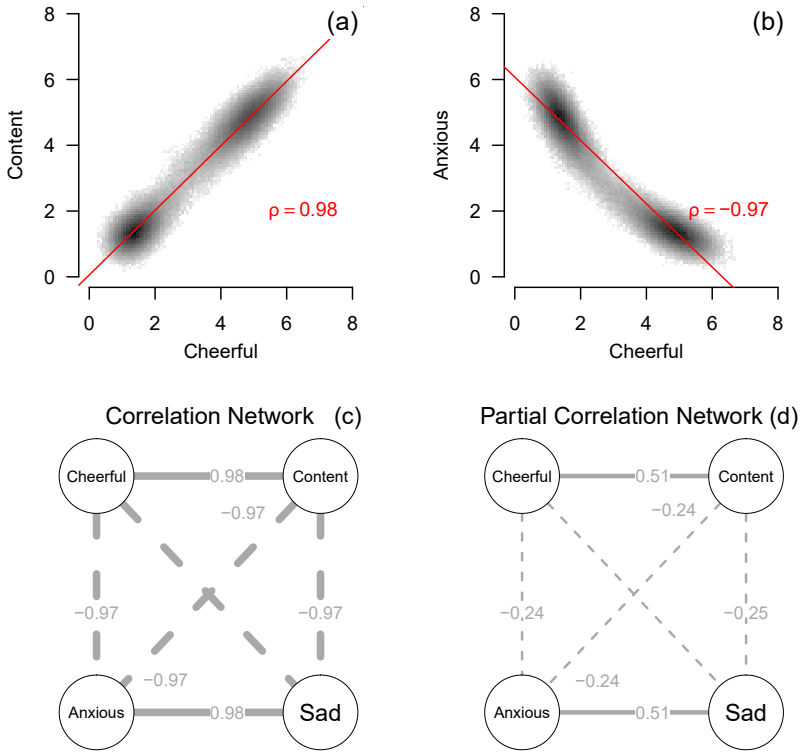


Figure 5.6: Panel (a) shows the relationship between Content and Cheerful, two emotions with the same valence, at the same time point. The red line indicates the best fitting regression model. Similarly, panel (b) shows the relationship between Anxious and Content, two emotions with different valence. Panel (c) displays the correlation matrix as a network, and panel (d) displays the partial correlation matrix as a network.

Panel (b) displays the relationship between Cheerful and Anxious, two emotions of different valence. We see that that the observations cluster around two points, one close to (1, 5) and the other one close to (5,1). The red line indicates the best fitting regression line (correlation $\rho = -0.97$).

Panel (c) displays the correlation network for all four emotion variables. As we have already seen in panel (a) and (b) there is a positive correlation ($\rho = 0.98$) between Content and Cheerful, and a negative correlation ($\rho = -0.97$) between Cheerful and Anxious. Due to the symmetry in the true bistable system, all correlations between emotions with the same valence are equal to $\rho = 0.98$ and all correlations between emotions with different valences are equal to $\rho = -0.97$. Panel (d) shows the partial correlation network (i.e., GGM). We see that the partial correlations between emotions with the same valence are equal to $\theta = 0.51$, and the partial correlations between emotions with different valences is equal to $\theta = -0.24$ or $\theta = -0.25$.

What can we learn from these results about the underlying bistable system? From inspecting the pairwise relationships of emotions with same and different valence in panels (a) and (b) one could guess the location and variance of possible fixed points, similarly to inspecting the histograms in Section 5.3.1. However, the 2-dimensional representation offers additional information about the stability landscape, for example the shape around the fixed points and the most likely paths to transition between them. When interpreting the correlations in panel (b) as "contemporaneous" relationships, we would conclude that there are strong positive linear relationships between emotions with the same valence, and similarly strong negative linear relationships between variables with different valences at a relatively short time scale. The partial correlations in (d) are smaller than the correlations, which is what one would expect since all correlations are high.

Using our knowledge about the true bistable system, which characteristics did we correctly recover? From inspecting the scatter plots in panels (a) and (b) one sees that most observations fall in one of two clusters indicating bistability (characteristic 1). Also, one can obtain rough estimates of the position of the fixed points (characteristic 2) and sees that the variances around the fixed points is different (characteristic 3). Note that the shape of the scatter plot in panel (b) is determined by the vector field in Figure 5.1 (c). The two clusters are exactly at the location of the two fixed points, and the observations between the clusters are both due to variance around the fixed points and switches between fixed points.

From the correlation and partial correlation networks, we correctly find that there are reinforcing effects within valences, and suppressing effects between valences (characteristic 5). However, the correlation network suggests that their relative size is equal, and the partial correlation network suggests that the reinforcing effects are stronger. In the true bistable system, however, the suppressing effects between valences are larger than the reinforcing effects within valences. Thus, judging the relative size of suppressing/reinforcing effects within/between valences from (partial) correlation would lead to incorrect conclusions.

In sum, inspecting scatter plots of pairwise relationships indicated bistability, and allowed us to obtain a rough estimate of the location of and variances around

the fixed points. The scatter plots also allowed one to get a projection of the stability landscape on two dimensions and thereby provided more information than histograms. While inspecting the scatter plots allows one to recover global dynamics of the true bistable system, one cannot infer the coupling between the emotion variables in the true bistable system from (partial) correlations. This is not too surprising since the Gaussian distribution is very restrictive in that it only models pairwise linear relationships (opposed to e.g., 3-way, 4-way, etc. interactions). In addition, it does not model any dependencies across time, which are the types of dependencies that constitute the microdynamics (characteristics 5-7) of the true model. In the next section, we inspect those dependencies across time and fit a Vector Autoregressive (VAR) model to the data, which captures temporal linear dependencies.

5.3.4 Lag-1 Relationships / VAR Model

In this section we aim to characterize the microdynamics between the four emotion variables by modeling the lagged relationships between them. Panels (a) and (b) of Figure 5.7 show the *marginal* relationship of Content at time t with Cheerful at the previous time point $t-1$, and Anxious at time t with Content at $t-1$, respectively. Although the data is generated from a dynamic model, the marginal lagged relationships look similar to the contemporaneous relationships shown in Figure 5.6: averaging over all other variables, lagged relationships within-valence are positive, and between-valence are negative. The reason is that the system largely stays around the two stable fixed points, and relative to the length of the time series, switches are infrequent. As such, the marginal relationships are largely driven by fixed point locations.

We can gain further insight into the dependencies between variables in our model by examining the *conditional* lagged relationships between pairs of emotions, that is, when keeping the other emotion variables at the previous time point(s) fixed. A popular model for such conditional lagged relationships is the first-order vector auto-regressive (VAR(1)) model. The VAR(1) model is one of the simplest multivariate dynamic models which can be fit to repeated measurement data, allowing linear relationships between all pairs of variables observed at consecutive measurement occasions t and $t-1$, that is

$$X_t = b + \Phi X_{t-1} + e_t \tag{5.2}$$

where b is a vector of intercepts, Φ is a matrix containing the auto-regressive (ϕ_{ii}) and cross-lagged ($\phi_{ij}, i \neq j$) effects, that is, conditional linear dependencies, and e_t is a vector of normally distributed residuals $e_t \sim \mathcal{N}(0, \Psi)$, which are independent across time, with residual variance-covariance matrix Ψ.

The VAR(1) model has been used widely to analyze experience sampling data in psychopathology research, particularly in the form of dynamic network analysis, wherein the Φ and Ψ matrices are used to construct directed and undirected network structures, respectively (e.g., Bringmann et al., 2013; Pe et al., 2015; Epskamp, Waldorp, et al., 2018). The VAR(1) model describes a system which fluctuates around a *single* stable fixed point: Stochastic input in the form of a

residual term pushes the system away from this fixed point, and the system returns to the fixed point with an exponential decay (Hamilton, 1994). The location of the stable fixed point is given by the mean vector μ, a function of the intercepts and the lagged relationships $\mu = (I - \Phi)^{-1} b$, where I is the identity matrix.

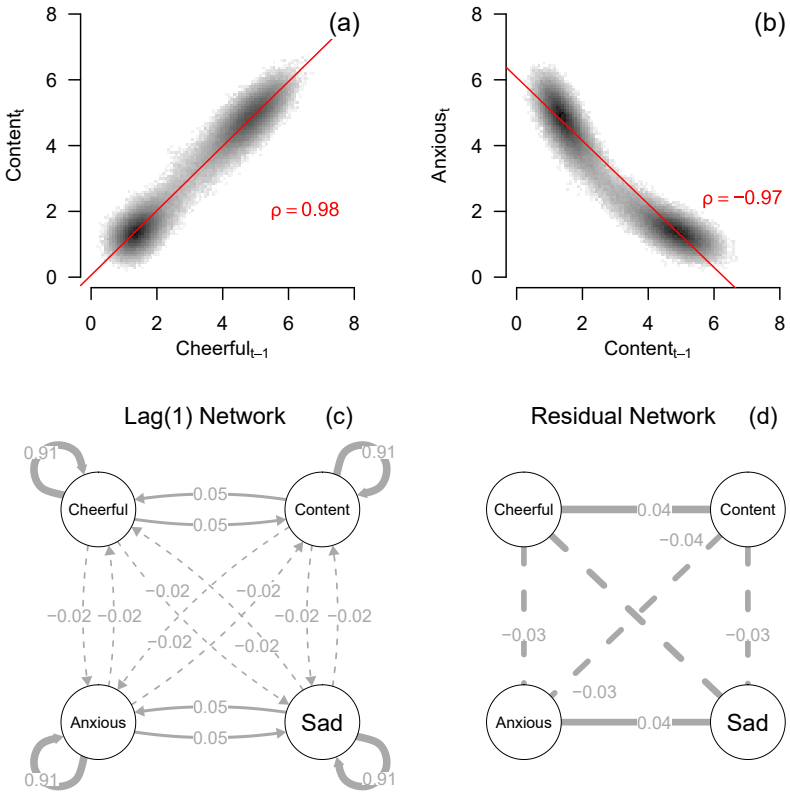


Figure 5.7: Panel (a) shows the relationship between Content and Cheerful, two emotions with the same valence, spaced one time point apart (at a lag of one). The red line indicates the best fitting regression model. Similarly, panel (b) shows the relationship between Anxious and Content, two emotions with different valence, at a lag of one. Panel (c) displays the matrix of lagged regression parameters, estimated from a VAR(1) model, as a network, and panel (d) displays the partial correlation matrix of the residuals of the VAR(1) model as a network. This latter network is often referred to as the contemporaneous network.

Panel (c) of Figure 5.7 displays the network of estimated lagged regression coefficients ($\hat{\Phi}$) between Cheerful, Content, Anxious and Sad. We can see that the auto-regressive parameters are large and positive for all four variables ($\hat{\phi}_{ii} = .91$). Furthermore, there are positive cross-lagged relationships between variables of the same valence ($\hat{\phi}_{12} = \hat{\phi}_{21} = \hat{\phi}_{43} = \hat{\phi}_{34} = .05$) and weaker, negative cross-lagged effects between variables of opposite valence ($\hat{\phi}_{13} = \hat{\phi}_{31} = \cdots = -.02$). All within-valence effects, and all between-valence effects, are of roughly equal

magnitude, respectively. In panel (d) we show the partial correlations of the residuals (i.e., standardized $\hat{\Psi}^{-1}$), sometimes referred to as the "contemporaneous" network or the residual GGM (Epskamp, Waldorp, et al., 2018). Here we see a similar pattern as above: The residuals have negative conditional relationships between-valence, and slightly greater in magnitude positive conditional relationships within-valence. Note that the residual variance in this case is quite low for each variable ($\hat{\Psi}_{ii} \approx 0.0185$ with in-sample explained variance of approximately 99 percent). The estimated fixed point (that is, the mean) is approximately $\hat{\mu}_1 = \hat{\mu}_2 = 3.16$ for Cheerful and Content, and $\hat{\mu}_3 = \hat{\mu}_4 = 3.04$ for Anxious and Sad.

Which characteristics of the bistable system can we recover based on the VAR(1) estimates? The strong auto-regressive effects correctly capture the strong linear auto-effects present in the true system, defined by the r parameters in Equation 5.2. The lagged regression parameters suggest that there are suppressing effects between valences, and reinforcing effects with valences, capturing characteristic number five of the data-generating mechanism (Section 5.2.3). However, the relative size of the suppressing and reinforcing effects is flipped in the VAR(1): The suppressing effects are in fact larger in absolute value than the reinforcing ones (see Section 5.2.1).

It is unclear what conclusions we can draw from the weak relationships present in the residual network — as there are no such additional instantaneous relationships present in the data-generating system. We assume a-priori when fitting the VAR(1) model that these parameters are independent of (i.e., constant across) time. Finally, the VAR(1) model describes a uni-stable system, precluding us from capturing any characteristics related to bistability. The dynamics implied by the VAR(1) are illustrated by generating new data from the estimated parameters, displayed in Appendix 5.C.2. The estimated location of the single fixed point is not equivalent to either of the two stable fixed points or the unstable fixed point in the true system.

Importantly, we can use our knowledge of the true bistable system to determine how we arrived at these observed parameter estimates. First, the estimated position of the fixed point (given by $\hat{\mu}$) is roughly halfway between the positions of the two stable fixed points, and is approximately equal to the sample mean for each variable, reflecting that the system spends roughly the same amount of time around each of the two stable fixed points. The main counter-intuitive result from the VAR(1) model is that the order of magnitude of the between- and within-valence relationships is different than in the true bistable system. In the true system, these pairwise relationships are non-linear, taking the form of interaction effects, and the VAR(1) model captures the best linear approximation of these non-linear relationships. As we can see from panel (a) of Figure 5.7, the linear approximation of the within-valence relationship is largely driven by the strong positive relationship present when both variables take on a high value, for instance when Cheerful and Content are both near the healthy fixed point. From panel (b) we can see that, in contrast, the linear between-valence relationship of Content on *Anxious* is in fact a mixture of the strong negative effect near the unhealthy fixed point (Content is low, Anxious is high) and the weaker effect near the healthy fixed point (Content is high, Anxious is low). Combined, this results

in higher linear within-valence relationships and lower linear between-valence relationships.

Finally, the residual covariances displayed in panel (d) of Figure 5.7 are produced by a combination of model-misspecification (linear approximation of non-linear relationships) and paths between each process at a shorter time scale than observed, due to the Euler steps used in data generation. We stress here that, even in the current idealized situation, it is not trivial to derive an exact explanation for the residual covariance structure, and so its utility in drawing conclusions about the underlying system should be approached with great caution.

In summary, the VAR(1) model gives us rather limited information regarding the core characteristics of the bistable model. In principle, the VAR(1) model is unable to capture any features which relate to bistability (characteristics 1-4), as one would expect from a model that exhibits only a single fixed point. What is perhaps more surprising is that, while the sign of the lagged relationships (characteristic 5), and their symmetries are captured, their relative ordering (characteristic 6) is not. This observation is critical: While we could expect that the VAR(1) model would not reproduce the global dynamics of the system, even when we have ideal data, the linear relationships in the VAR(1) model also fail to appropriately capture the local microdynamics in this instance. Fundamentally, this is due to the non-linear relationships which must be present in the underlying system in order to induce bi-stability. In general we would not expect that linear approximations of non-linear effects would preserve the same rank ordering. This observation has potentially major implications for the analysis of dynamic network structures, because many network metrics such as centrality metrics are strongly dependent on the relative ordering of effects. In the following we will examine if extending the VAR(1) model to allow for bi-stability brings us closer to recovering a more accurate characterization of the data-generating bistable system.

5.3.5 Threshold VAR Model

Regime-switching VAR(1) models extend the VAR(1) model to allow for observations to be drawn from two different conditional distributions for X_t given X_{t-1}, that is, two different *regimes*, described by two different sets of model parameters. These extensions in principle allow us to directly capture a notion of multi-stability, by interpreting the mean vector of each conditional distribution as a separate fixed point. Different extensions allow for different mechanisms by which to model the switch between these regimes.

One popular regime-switching VAR(1) model is the Threshold TVAR(1) model, where the system enters a different regime whenever a threshold value or values τ of an a-priori specified threshold variable z_t is crossed, written

$$
\begin{aligned}
X_t &= b^{(1)} + \Phi^{(1)} X_{t-1} + e_t^{(1)} && \text{if } \; z_t \leq \tau \\
X_t &= b^{(2)} + \Phi^{(2)} X_{t-1} + e_t^{(2)} && \text{if } \; z_t > \tau
\end{aligned}
$$

for a two-regime model with a single threshold, where the VAR(1) parameters are indexed by regime, with $e_t^{(r)} \sim \mathcal{N}(0, \Psi^{(r)})$, and mean vectors $\mu^{(r)} = (I - \Phi^{(r)})^{-1} b^{(r)}$ (Tong & Lim, 1980; Hamaker, Grasman, & Kamphuis, 2010). The threshold variable z_t may be an exogenous variable, or one of the variables in the VAR model. Here we choose to use Cheerful ($z_t = x_{1,t-1}$) as the threshold variable. The threshold value τ is a hyper-parameter that is estimated. Here, we estimate the TVAR(1) model using the R-package *tsDyn* (Fabio Di Narzo, Aznarte, & Stigler, 2009), which estimates τ using a grid search which selects the model with minimum summed squared residuals.

Figure 5.8 displays the main results from the estimated TVAR(1) model, in which the threshold is estimated as $\hat{\tau} = 2.811$. In panel (a) of Figure 5.8 we show the time-series with shading indicating which observations are below (grey) or above (white) the threshold. We can see that the estimated threshold nicely separates the time series into periods in which the system is in an unhealthy state (based on Cheerful values below the threshold) and a healthy state (Cheerful values above the threshold).

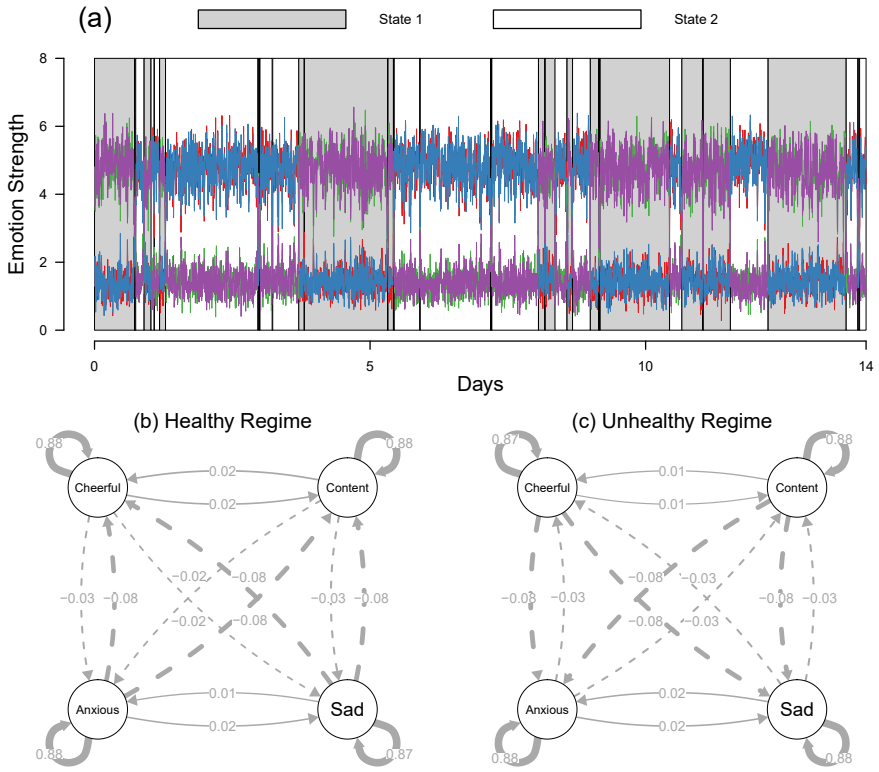


Figure 5.8: Panel (a) shows the two weeks of the time series, with observations shaded in either grey or white as a function of whether $x_{1,t-1}$ is above or below the threshold $\hat{\tau} = 2.811$. Panels (b) and (c) show the estimated VAR(1) parameters as lagged networks in the healthy (white) and unhealthy (grey) regimes respectively.

Inspecting the lagged networks for each regime in panels (b) and (c) of Figure 5.8 we can see that the auto-regressive effects, and the within-valence cross-lagged effects are pretty similar across both regimes. However, the cross-lagged effects between variables of opposite valence are different. In the healthy regime, negative valence emotions have much stronger cross-lagged effects on positive emotions ($\hat{\phi}^{(2)}_{13} = \hat{\phi}^{(2)}_{14} = \hat{\phi}^{(2)}_{23} = \hat{\phi}^{(2)}_{24} = -.08$), and vice versa for the unhealthy regime ($\hat{\phi}^{(1)}_{31} = \hat{\phi}^{(1)}_{41} = \hat{\phi}^{(1)}_{32} = \hat{\phi}^{(1)}_{42} = -.08$). Residual partial correlation networks for both regimes are shown in Appendix 5.D, which display a similar pattern to the regular VAR(1) model of weak positive residual partial correlation within-valence and weak negative residual partial correlation between-valence. For the TVAR, however, the residual covariance matrix is not symmetric across regimes: In the healthy regime there is a slightly higher covariance between positive emotions than negative emotions, and vice versa. The estimated means are given as $\hat{\mu}_2 = \{4.74, 4.75, 1.45, 1.46\}$ for the healthy state and $\hat{\mu}_1 = \{1.49, 1.48, 4.69, 4.69\}$ for the unhealthy state. Data generated by the TVAR(1) model estimates is shown in Figure 5.18 in Appendix 5.C.3. From this figure we can see that most of the global dynamics are well reproduced, although the system contains fewer switches between regimes than we would expect and there are fewer observations on the switches between states compared to the original time series.

Which characteristics of the bistable system do we recover on the basis of the TVAR(1) parameter estimates? First, the model picks up a number of characteristics related to the bistability of the system. The estimated mean vectors capture approximately the position of the two stable fixed points (characteristic 2), and the estimated threshold correctly captures the position of the unstable fixed point in the Cheerful dimension. However, note that bistability (characteristic 1) has been specified a-priori and therefore cannot be considered to be recovered by the model. Second, although the simulated data in Figure 5.18 (Appendix 5.C.3) exhibits less frequent switches between states than we would expect, we can see that the combination of state-dependent lagged parameters and residual variances does reproduce higher variability of positive emotion in the healthy state in comparison to the unhealthy state, and vice versa for negative emotions (characteristic 3). Finally, the lagged regression parameters in each regime correctly capture that there are reinforcing effects within valence, and suppressing effects between valence (characteristic 5).

The result that stands out in this analysis is the asymmetry in lagged regression coefficients across both regimes. This asymmetry would appear to indicate that the parameters relating processes either change over time or are all explicitly a (step) function of the Cheerful variable. This last result is striking because this intuitive interpretation does not correctly characterize the relationship between variables of different valences in the true bistable system. This is because we know that the dependencies in C are invariant over time and fully symmetric. However, the dependencies in C relate to pairwise interaction effects rather than linear dependencies in the VAR(1) model. For example, the relationship between Anxious, denoted x_3, and the rate of change of Content, $\frac{dx_2}{dt}$, depends both on the value of C_{23} and on the current value of x_2

$$\frac{dx_2}{dt} = r_2 x_2 + (C_{23} \times x_2) x_3 + \dots \ . \tag{5.3}$$

If we view x_2 as a moderator, we can see that, when x_2 is high, the effect of x_3 on the rate of change, given by $C_{23} \times x_2$, is relatively greater than when x_2 is low. In our system, separating the time-series into two regimes based on a threshold of 2.811 for the Cheerful emotion essentially means we condition on high values of x_1 and x_2 in the healthy regime, and low values in the unhealthy regime. This leads to the relatively stronger linear relationship from negative emotions to positive emotions in the healthy regime, and vice versa in the unhealthy regime. As such, we can see that the asymmetry in lagged relationships over time picked up by the TVAR(1) model is a characteristic of the true bistable system. Notably, however, the mechanism by which this asymmetry occurs is entirely due to nonlinear relationships between the observed variables and the similarity of variables that share the same valence, while the TVAR(1) modeler might be tempted to ascribe this entirely to the effect of the level of Cheerful.

To summarize, the TVAR(1) model allows us to recover global dynamics, and it recovers some aspects of the microdynamics. However, we saw that a naive interpretation of the TVAR(1) parameter estimates may easily lead to the incorrect conclusion that there is one time-varying variable which moderates the relationships between all variables. In addition, we provided bistability as an input to the model, and therefore cannot be considered a characteristic recovered from data. In principle one could perform model selection between TVAR(1) models with different numbers of components, however compared to the Mean switching HMM in Section 5.3.2, the run time for such a model comparison was unfeasible for the large data set used in our paper.

Furthermore, note that the threshold VAR(1) model does remarkably well for this specific system for the following reason: While TVAR(1) models have frequently been discussed in the literature (e.g. Warren, 2002; Hamaker et al., 2009, 2010; De Haan-Rietdijk, Gottman, Bergeman, & Hamaker, 2016) a major limitation of this method is the difficulty in choosing a threshold variable. In our data-generating mechanism, we know there to be an unstable fixed point defined in multivariate space, $x_1 = x_2 = x_3 = x_4 = 2.8$. It just so happens that in this case, almost always when we pass this position in one dimension (e.g., $x_1 > 2.8$) we also do so in all other dimensions (e.g., $x_2 > 2.8$, $x_3 < 2.8$, $x_4 < 2.8$). This means that the true mechanism of state-switching behavior is very well approximated by the univariate mechanism in the TVAR(1) model, for this choice of parameter values. In more general situations, the choice of threshold variable(s), and number of thresholds, is likely to be less trivial. While the TVAR(1) model does capture that there are suppressing and reinforcing effects between and within valences, it does not capture the relative size of these effects, and it may easily lead to the incorrect conclusion that there is a single time-varying variable which moderates all of the relationships between other variables in the system.

Finally, the TVAR(1) is only one of a variety of different regime-switching dynamic models which could be fitted to the data at hand. Another alternative would be the Markov-Switching (MS-)VAR model (Hamilton, 1989; Hamaker et

al., 2010; Hamaker & Grasman, 2012; Chow et al., 2018), a combination of the HMM and VAR models, in which the regime-switching behavior is determined by a random Markov process operating between latent categorical variables. While this model is more flexible than the Threshold VAR model, we show here the TVAR results for two reasons. First, in this instance the switching behavior will be less well approximated by the MS-VAR model, leading to even less straightforward conclusions about the data-generating process, but otherwise likely highly similar lagged parameter estimates. Second, while recent advances such as the *dynR* package (Ou et al., 2019) have made this model easier to estimate, it is still prohibitively difficult and time consuming to fit to data.[3]

Now that we have shown the capabilities and limitations of the TVAR model in recovering the bistable system, there are a few different avenues we could pursue to further increase our model complexity in the hope of recovering more and more of the features of underlying system. For example, both the TVAR and MS-VAR can be considered special cases of *time-varying parameter* models, that assume the true time-varying model is a partition between a finite set of components. Other types of time-varying VAR models assume the parameters are a smooth function of time (e.g., Haslbeck, Bringmann, & Waldorp, 2017). However, we would not expect these models to outperform the threshold VAR in this instance for two reasons. First, the threshold VAR model is already able to capture the major source of variation in parameters over time, that is, the step-like switches between stable states. Second, since these models are still based on fitting locally stationary VAR models, the fundamental limitations of approximating the dynamics with linear relationships remain. As such, in the next section we examine an approach which aims to recover the exact system of differential equations (DEs) from data, by allowing non-linear terms to enter into a step-wise model building procedure.

5.3.6 Differential Equation Model Building

In the previous sections we have shown that some models have been able to recover some characteristics of the true model, but that it is generally difficult to make inferences about the characteristics of the true system from these models. Also, since all of these models were misspecified they were fundamentally unable to recover the exact true bistable system. In this section, we aim to recover the exact system of differential equations (DEs) directly from the ideal time series.

5.3.6.1 Model Building Procedure

The structure of the true model is typically unknown in practice, and therefore has to be learned from the data. Chow (2019) describes a general methodology for building dynamic systems models which consists of two steps: In the first step, we approximate the first-order derivatives by taking difference scores between

[3] Despite numerous attempts and correspondence with the authors of the package, we were unable to get the model estimates for the dataset described here to converge.

consecutive measurement occasions, divided by the length of the time-interval between those occasions

$$\frac{dx_{i,t}}{dt} \approx \frac{x_{i,t+1} - x_{i,t}}{\Delta t} \tag{5.4}$$

where in each case, $\Delta t = .1$, as described in Section 5.2.2 (cf. Boker, Deboeck, et al., 2010).

In the second step, we use this approximate derivative as the outcome variable, and try to find a regression model that predicts this outcome variable as well as possible with as few parameters as possible. Here, we use results obtained from the statistical models in the previous sections as a starting point, and then follow the standard model-building approach of fitting models with increasing complexity and evaluating the improvement in out-of-sample fit.

From the descriptive statistics (Section 5.3.1) the marginal lag-0 (Section 5.3.3) and lag-1 (Section 5.3.4) relationships and the mean-switching Hidden Markov Model (Section 5.3.2), we saw that the system is bistable. From dynamical systems theory we know that bistability is only possible in the presence of non-linear terms (Strogatz, 2015). Similarly, we saw from the TVAR(1) model (Section 5.3.5) that the linear relationships between pairs of variables is dependent on where in the state-space other variables are located (i.e., below or above a given threshold). Both of these pieces of information suggest the presence of interaction effects between variables: However, we have no information about what specific interaction terms, or what other linear or non-linear dependencies should be in the model. Here, we start out with a main effects-only model, and then add more and more non-linear terms (interactions, quadratic effects, etc.). We evaluate the fit using the mean out-of-bag proportion of explained variance ($\mathcal{R}^2$) obtained from a 10-fold cross-validation scheme (see Appendix 5.E for details), and we choose the model that maximizes this value. If two models result in the same fit, we choose the model with less parameters.

The fit of the all models considered here is shown in Table 5.1. First, we test a baseline model (Model A) where each derivative is a linear function of all other variables. As we described above, the absence of interaction effects makes it unlikely that this is a suitable candidate model, but it gives us a baseline explained variance value of $\mathcal{R}^2 = 0.04664$. In Model B, we add to the baseline model all pairwise interactions between the outcome process (e.g., x_1 when the DV is dx_1/dt) and the other variables in the model x_j (i.e., $x_1 \times x_j$, $\forall j \in p$). Adding these pairwise interactions increases the variance explained to $\mathcal{R}^2 = 0.06874$. In Model C, we further extend this model by adding all possible pairwise interactions between all variables $x_i \times x_j$, $\forall (i, j) \in p$. However, we see that adding these parameters in fact leads to a slight decrease in explained variance, $\mathcal{R}^2 = 0.06870$, indicating overfitting. For brevity, we display only these three models, but adding further complexity to model in terms of additional interaction terms, quadratic or cubic terms also fails to increase the out-of-bag $\mathcal{R}^2$ (see Appendix 5.E). As such, we can take Model B to be our final model.

Model	$\frac{dx_{i,t}}{dt} \sim a_i + r_i x_i + \ldots$	q	$\mathcal{R}^2$
A	$\sum_{j \neq i} r_j x_j$	5	0.04464
B	$\sum_{j \neq i} R_{ij} x_j + \sum_j^p C_{ij} x_j x_i$	9	0.06874
C	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k$	15	0.06870

Table 5.1: Model fit results for each of the four models described in text. The second column gives the model equation for each variable, q denotes the number of parameters estimated per univariate regression model, and the final column indicates the mean proportion of explained variance $\mathcal{R}^2$, calculated on the hold-out sets of a 10-fold cross-validation scheme (for details see Appendix 5.E)

5.3.6.2 Dynamics and Data Generated by Final Model

We can see that the structure of Model B is highly similar to the structure of our data-generating model, with additional main effects between variables, that is, the linear effects denoted by the $p \times 1$ vector r in the true model is replaced by a $p \times p$ matrix R in our chosen model. Furthermore, we can see from the left panel of Figure 5.9 that the parameter estimates are highly similar, but not exactly equal to the data-generating parameters:

$$\hat{a} = \begin{bmatrix} 1.40 & 1.37 & 1.25 & 1.27 \end{bmatrix}^T$$

$$\hat{\sigma} = \begin{bmatrix} 1.35 & 1.34 & 1.34 & 1.34 \end{bmatrix}^T$$

$$\hat{R} = \begin{bmatrix} 0.88 & 0.02 & -0.01 & 0.01 \\ 0.03 & 0.95 & 0.01 & 0.00 \\ -0.02 & 0.05 & 0.96 & 0.04 \\ 0.04 & -0.01 & 0.08 & 0.91 \end{bmatrix}$$

$$\hat{C} = \begin{bmatrix} -0.18 & 0.04 & -0.17 & -0.18 \\ 0.03 & -0.19 & -0.19 & -0.19 \\ -0.18 & -0.19 & -0.19 & 0.03 \\ -0.19 & -0.18 & 0.02 & -0.18 \end{bmatrix}$$

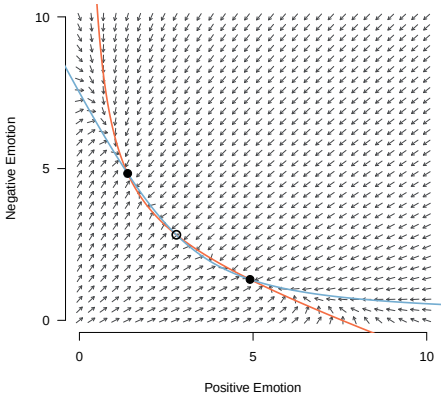


Figure 5.9: Left panel: the parameters estimated from the ideal data. Right panel: the vector field defined by the estimated parameters in the left panel. Solid points indicate stable fixed points and empty points indicate unstable fixed points. The solid lines indicate the values at which derivative of positive emotion (orange) and negative emotion (light blue) is equal to zero. At the points at which the two lines meet, both derivatives are equal to zero and the system remains in this (stable) state.

While we would not expect to recover the exact parameters of the true model with a different functional form, we see that the signs, size and relative orderings of parameters in the estimated $\hat{C}$ matrix are quite accurate. Based on these parameters, we recover that there are suppressing effects between valences and re-

inforcing effects within valences (characteristic 5), that the reinforcing effects are smaller in absolute value than the suppressing effects (characteristic 6), and by capturing approximately the correct functional form, we capture that the micro-dynamic parameters are dependent only on variables inside the model (characteristic 7). Furthermore, we can see that false positive (i.e., off-diagonal) elements of $\hat{R}$ are of a much smaller size than the true positive diagonal elements. The full parameter estimates, with standard errors and p-values are shown in Appendix 5.E.

Beyond inspecting the estimated parameters, we can judge how good of an approximation of the true bistable system our estimated model represents by comparing the dynamics implied by that model to that of the true system. The dynamics of a differential equation model are described by its vector field, which we depict for Model B in the right panel of Figure 5.9. To construct this vector field we use the same two-dimensional approximation (positive and negative emotion) as we did in Section 5.2.1 (see Appendix 5.A for details). The orange and light blue lines are solution lines which indicate the locations where the rate of change in one dimension (orange for no change in positive emotion, light blue for no change in negative emotions) is zero. The points at which these solutions line cross determine the fixed points. We can see that our model correctly identifies three fixed points in this range of values: one stable healthy ($x_1 = x_2 = 4.91, x_3 = x_4 = 1.34$), one stable unhealthy ($x_1 = x_2 = 1.39, x_3 = x_4 = 4.84$), and one unstable fixed point approximately halfway between those two ($x_1 = x_2 = 2.79, x_3 = x_4 = 2.82$). If we compare these global dynamics to the global dynamics of the true bistable system depicted in Figure 5.1 in Section 5.2.1, we see that Model B very accurately reproduces these dynamics, approximating the position of the fixed points in the true system closely. From this we can say that the estimated DE model captures characteristics 1 (bistability) and 2 (location of the fixed points) in the true system.

An additional way to evaluate whether the dynamics of the estimated model are similar to the dynamics of the true model is to generate data from the estimated model and compare this data to the original data. Figure 5.10 shows a time series generated from Model B using a step size of .1. We can see that the data looks very similar to the original data generated from the true bistable system, in that the fixed points are at roughly the same location, there is a difference in variance across the high and low emotion value fixed points (characteristic 3), and there is a similar number transitions (around 14) between the healthy and unhealthy state (characteristic 4). Thus, even though we did not exactly recover the set of true parameters exactly, we seem to have recovered a model that is equal to the true model in all relevant aspects, capturing all of the seven characteristics listed in Section 5.2.3.

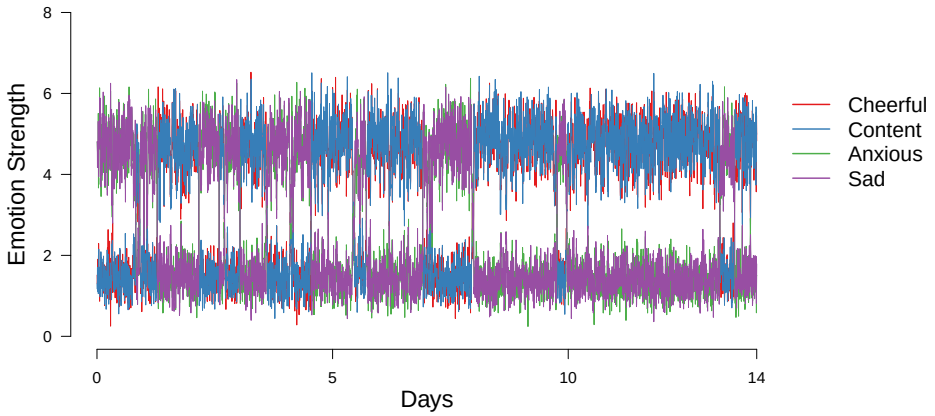


Figure 5.10: Data generated from the estimated DE model, with the same initial values as the observed data

5.3.6.3 Exact Recovery of Model Parameters

While this model building procedure performed extremely well in this scenario, the findings here should be approached with a note of caution. Observe that, despite negligible sampling error, we do not succeed in recovering the exact parameter estimates. The reason for this is that, while data is generated using an Euler step of $\Delta t = .01$, our ideal time series is created by sub-sampling with $\Delta t = .1$. While this is an unrealistically high sampling frequency, it still means that we cannot estimate the derivative perfectly. As the sampling frequency becomes lower, we would expect the quality of this approximation to degrade.

In theory, to recover the data-generating parameters, we would need to fit the *integral solution* form of the differential equation (Strogatz, 2015), as this describes the relationships between observed variables spaced Δt apart, as implied by the differential equation. It is well known that this integral solution may contain a seemingly different set of dependency relationships than the differential equation. For instance, variables which are independent in the DE form may be dependent in the integral form, and the signs and relative orderings of these dependencies may change depending on the value of Δt (Ryan & Hamaker, 2019; Kuiper & Ryan, 2018; Aalen et al., 2012). Because methods based on approximating integral solutions are expected to suffer from similar problems as the two-step DE estimation procedure, and because these methods are difficult to apply in practice, we limit ourselves to the two-step approach in this paper (see section 5.5.3 for a further discussion).

5.3.7 Summary: Analysis of Ideal Time Series

In this section we aimed to recover characteristics of the true bistable system from the ideal time series with 10 measurements each minute using a number of time series analysis tools. Table 5.2 provides a rough summary of which method recovered which characteristics of the true bistable system.

	Bistability (1)	Position (2)	Variance (3)	Transitions (4)	Suppr./Reinf. (5)	Relative Size (6)	Time-constant (7)
Data Visualization	✓	✓	✓	✗	✗	✗	✗
HMM	✓*	✓	✓	✓	✗	✗	✗
Lag-0 / GGM	✗	✗	✗	✗	✓	✗	✓*
Lag-1 / VAR(1)	✗	✗	✗	✗	✓	✗	✓*
TVAR(1)	✓*	✓	✓	✓	✓	✗	✗
DE-Estimation	✓	✓	✓	✓	✓	✓	✓*

Table 5.2: Summary of which method recovered which of the seven qualitative characteristics listed in Section 5.2.3 from the ideal time series. The first four characteristics are global dynamics, thee last three are microdynamics. The check marks with asterisk indicate that the method includes the characteristic as a model assumption, and can therefore not be considered recovered from the time series.

We showed that data visualization (Histograms and the pairwise marginal relationships in Sections 5.3.3 and 5.3.4) revealed bistability and provided a rough estimate of the position of and variances around the fixed points. However, when comparing the eye-balled number of switches with the estimates of the HMM, we saw that we missed instances in which the system quickly switched back and forth. The Mean switching HMM recovered all global dynamics, however, we provided bistability as a model assumption, which is why we mark the check mark at the first characteristic with an asterisk.

Turning to methods that capture dependencies between variables, the analysis of lag-0 relationships with the GGM and the analysis of lag-1 relationships with the VAR(1) model (and a GGM on its residuals) fundamentally cannot recover any global dynamics of the bistable system, but they recovered some microdynamics: the characteristic that within valence effect are reinforcing, and between valence effects are suppressing; and that the parameters are constant across time, however this is again an assumption of the model and therefore cannot be considered recovered from the data. The TVAR(1) model was able to recover all global dynamics with the same caveat as in the HMM, that bistability is a model assumption and not recovered from data. Similarly to the VAR model, it recovered the reinforcing/suppressing characteristic. However, a naive interpretation of the model parameters would lead one to conclude that the parameters are time-varying. Finally, the DE-estimation method was able to recover all microdynamics reasonably well, which implies that it also recovered all global dynamics.

The purpose of this section was to establish whether or not each method can recover, in principle, some aspect of the bistable system. To do this we used a highly idealized dataset, with an unrealistically high sampling frequency. As such, the performance of each method described above can be considered an upper bound on its performance in any more realistic scenario. It remains to be seen exactly how the performance of each method, and in general our ability to

recover global and microdynamic characteristics of the system, changes when a more realistic sampling frequency is used.

5.4 Recovering the Bistable Systems from ESM Data

In this section we analyze a time series that is similar to the ideal time series in all aspects, except that the system is sampled every 90 minutes instead of every six seconds (see Section 5.2.2). This allows us to investigate how the ability of each method to recover (some characteristic of) the bistable system is affected by having only a low sampling frequency time series, as it is typical for ESM studies.

5.4.1 Descriptive Statistics, HMM and Lag-0 Relationships

The descriptive statistics, such as histograms, and the lag-0 relationships obtained from the ESM times series (Figures 5.20 and 5.21 respectively in Appendix 5.F) are essentially identical to those obtained from the ideal time series, depicted in Figures 5.4 and 5.6 in Section 5.3.1. This makes sense: we have exactly the same amount of data points, sampled from the same system as in the ideal time series case. The only difference is that in the ESM data set 900 time points are "missing" between each measurement of the ESM time series. However, because lag-0 relations do not pick up on any temporal dependence, the lower sampling frequency does not affect the lag-0 relations. While this suggests that lag-0 relations are robust against low sampling frequency, it also puts their utility to infer the dynamics of an underlying dynamical system into question.

The parameter estimates obtained by fitting a two-component mean-switching Hidden Markov Model on the ESM dataset were

$$\hat{\mu}_1 = \begin{pmatrix} 1.47 \\ 1.46 \\ 4.71 \\ 4.71 \end{pmatrix}, \ \hat{\sigma}_1 = \begin{pmatrix} 0.41 \\ 0.41 \\ 0.63 \\ 0.63 \end{pmatrix}, \ \hat{\mu}_2 = \begin{pmatrix} 4.71 \\ 4.71 \\ 1.47 \\ 1.47 \end{pmatrix}, \ \hat{\sigma}_2 = \begin{pmatrix} 0.64 \\ 0.64 \\ 0.41 \\ 0.41 \end{pmatrix}, \ \hat{A} = \begin{pmatrix} 0.915 & 0.085 \\ 0.090 & 0.910 \end{pmatrix}.$$

and the predicted states for two weeks of the time series are shown in Appendix 5.F Figure 5.22.

We see a very similar pattern of results as obtained from the HMM fit to the ideal time series in Section 5.3.2, with the means and standard deviations of state 1 and state 2 reflecting the unhealthy and healthy states respectively. However, the parameters of the estimated transition matrix A for this time series show substantially higher switching probabilities, $\hat{A}_{12} = .085$ and $\hat{A}_{21} = .090$. As we can see from Figure 5.3, although the sub sampled ESM time series contains only 224 observations for a two-week period, rather than 201600, the sampling frequency is still high enough to capture each of the 17 switches between states in this period. That means that, although the amount of transitions that occur over a period of time remains the same, the number of measurement occasions between any two transitions is lower, which results in a higher transition probability. We

can see that this higher transition probability captures the number of transitions over two weeks quite accurately — the model predicts between $224 \times .090 \approx 20$ and $224 \times .085 \approx 18$ switches on average over a two week period. As such, the HMM fitted on the ESM time series still allows us to estimate the location of and variance around the two fixed points (characteristics 2 and 3), and approximate the frequency of transitions between these two fixed points (characteristic 4). In fact, the transition probabilities appear to be even more accurate than the ideal case — most likely this numerical imprecision in the ideal case is because the number of transitions relative to total time series was so low that slight changes in the transition probability value lead to very different predictions regarding the number of transitions over 201600 time points.

5.4.2 Lag-1 Relationships and VAR model

When analyzing the lagged relationships in the ESM time series, we begin to see some striking differences from the analysis of the ideal time series. Panels (a) and (b) of Figure 5.11 show the *marginal* relationship of Content observed at time t with Cheerful at the previously observed time point $t-1$, and Anxious at time t with Content at time $t-1$, respectively. Focusing on panel (a), we see that the density of the lagged variables takes on a square-like shape, and each quadrant seems to be filled with a roughly circular density. This is in contrast to the density of the lagged relationships in the ideal data displayed in Figure 5.6 in Section 5.3.4, which was described by two elliptical shapes at the two fixed points.

How can we explain this pattern? In the ideal data, the two elliptical shapes indicate that Content_{t-1} and Content_t tend to be near the same fixed point (two shapes), and that the two variables are positively correlated (elliptical shape). Now, in the ESM time series, we still have most of the density in the upper-right and the bottom-left quadrant, indicating that if Content_{t-1} is at the healthy (unhealthy) fixed point, it is very likely that Content_t is also at the healthy (unhealthy) fixed point (noting that $t-1$ now reflects a 90 min instead of 6 second time interval). However, we now also observe density at the top-left and bottom-right quadrant. These densities represent the situation in which Content_{t-1} is in the healthy (unhealthy) state, but Content_t is in the unhealthy (healthy) state. This situation is created when a switch between states falls within the 90min interval between two ESM measurements. Next, we focus on the shape of the density *within* each of the quadrants: we see that each of the densities have roughly a circular shape, which indicates that Content_{t-1} and Content_t are uncorrelated at each fixed point. This makes sense: in the ESM time series 900 time points are missing between each pair of measurement, which means that the that there is essentially no temporal dependence anymore between the variables. The relationship between Anxious_{t-1} and Content_t can be explained in an analogous way. Before fitting the VAR model below, this already shows us that it is futile to recover the microdynamics of the bistable system from these data.

Panel (c) of Figure 5.11 displays the network of estimated lagged regression coefficients from a VAR(1) model fit to the data. If we were to use these to in-

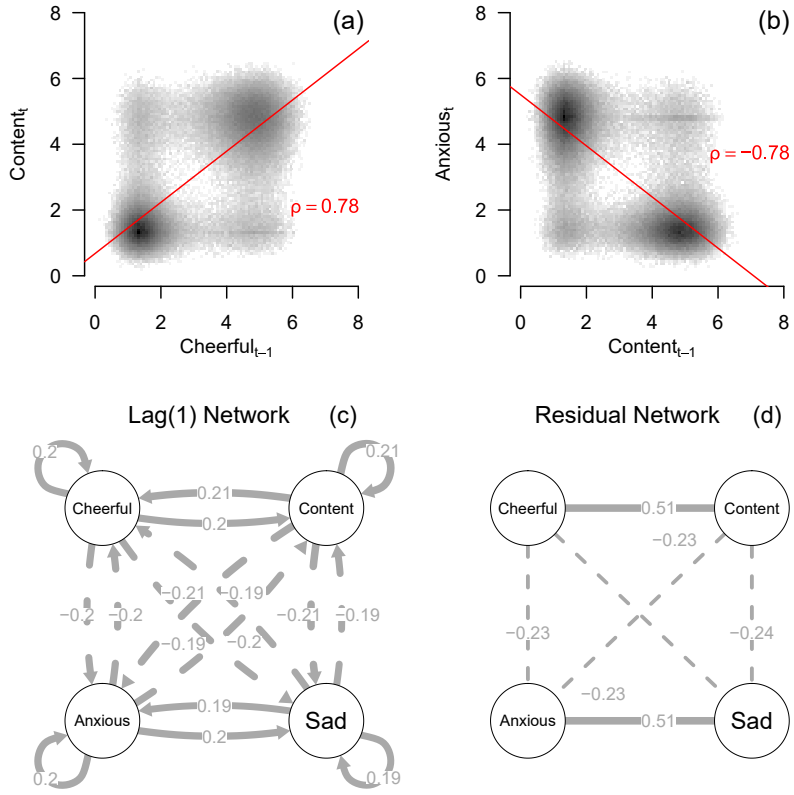


Figure 5.11: Panel (a) shows the relationship between Content and Cheerful, two emotions with the same valence, spaced one measurement occasion apart (i.e. at a lag of one but with 90 minutes between measurements) for the ESM dataset. The red line indicates the best fitting regression model. Similarly, panel (b) shows the relationship between Anxious and Content, two emotions with different valence, at a lag of one. Panel (c) displays the matrix of lagged regression parameters, estimated from a VAR(1) model, as a network, and panel (d) displays the partial correlation matrix of the residuals of the VAR(1) model as a network. This latter network is often referred to as the "contemporaneous" network.

fer the microdynamic characteristics of the system, we would manage to recover the signs of effects between variables: Negative, suppressing effects between-valence, and positive, reinforcing effects within-valence (characteristic 5, see Section 5.2.3). However, we again fail to recover the relative size of the suppressing and reinforcing effects. In fact, in this case, all of the estimated auto-regressive and cross lagged effects have approximately equal absolute value $|\hat{\phi}_{ij}| \approx .2$. This means that we also fail to recover the strong auto-regressive relationships encoded by the r parameters in the data-generating model, and reflected by the strong auto-regressive effects estimated by the VAR(1) model in the ideal setting. In panel (d) we can see that, as was the case for the ideal time series, we obtain

positive residual partial correlations within-valence and negative residual correlations between-valence, although the magnitude of these correlations is now quite high, $\hat{\theta} = 0.51$ and $\hat{\theta} = -0.23$ respectively. In addition, note that the residual variances of each variable in the model is considerably higher than the ideal case, and approximately equal for all variables ($\hat{\Psi}_{ii} \approx 1.1$, explained variance ≈ 0.62). As we would expect, taking only every 900th measurement from the ideal time series means that the predictive power of the VAR model decreases.

How can we reconcile these parameter estimates with what we know of the underlying bistable system? Although we may be tempted to interpret the VAR(1) parameters as reflecting the microdynamic structure, we have already seen in the analysis of the marginal relationships above that the large time-interval between observed measurements means that such an interpretation would be incorrect. Instead, the VAR(1) parameter values in the present situation are fully determined by the global characteristics of the system. Essentially, the estimated lagged relationships reflect that, at a time-scale of 90 minutes, the dynamics of the system from one observation to the next can be boiled down to four possibilities: Either the entire system stays in the same state (healthy-healthy or unhealthy-unhealthy) or moves from one state to the other (healthy-unhealthy or unhealthy-healthy). Since 1) the most likely behavior is that the system stays near the same fixed point, and 2) those two fixed points are defined as high-positive and low-negative emotions, or low-positive and high-negative emotions, we end up with positive within-valence relationships (e.g., if Cheerful now is near the high fixed point, it is likely that Content later will be high too) and negative between-valence relationships (e.g., if Anxiety now is high, it is likely that Content later will be low). All of the auto-regressive and cross-lagged relationships are of equal value, as essentially all variables have the same value in predicting what fixed point each other variable will be near at the next measurement occasion: Enough time elapses between measurement occasions that even the auto-regressive effect is only as predictive as the cross-lagged effects. As noted above, this interpretation is also reflected in the joint densities in panels (a) and (b) of Figure 5.11. Here, each density takes the appearance of four quadrants of uncorrelated Gaussian distributions, indicating that the microdynamic dependencies present in the ideal times series are totally absent from the ESM time series.

In summary, having a realistic sampling frequency results in the VAR(1) model providing even less information about the characteristics of the bistable model than in the ideal scenario. The longer time-interval between observations implies that interpreting VAR(1) parameters as reflecting truly microdynamic behavior would be incorrect. Parameters interpreted as reflecting microdynamics in fact must be interpreted as reflecting the global characteristics of the system. Although the sign of the microdynamic relationships (characteristic 5) is recovered, in this instance it happens that the pattern of microdynamic relationships has the same valence as the pattern of relationships at a longer time-scale, that is, the movement of the process between fixed points. Thus, while in the ideal case the VAR parameters were a mixture of the microdynamics (around each fixed point), and global characteristics (i.e., position of the two fixed points), in the ESM time series these parameters are only reflective of the global characteristics.

5.4.3 Threshold VAR Model

We saw in the previous sections that inferring global characteristics using ESM data was somewhat successful, but that inferring microdynamic characteristics using a VAR(1) model was impossible. For the ideal time series, we saw that the threshold VAR(1) model was in principle able to capture some microdynamic and some global characteristics, and so in this section we examine how well that performance generalises to our emulated ESM data. We use the same threshold variable (Cheerful, X_1) and model specification as described in Section 5.3.5.

Figure 5.12 displays the main results from the TVAR(1) model estimated on the ESM data: The estimated threshold is $\hat{\tau} = 2.796$, very close to the estimated threshold in the ideal case, and from panel (a) of Figure 5.12 we can see that this threshold value does well in separating the time-series into the healthy and unhealthy states. Inspecting the lagged networks for each regime in panels (b) and (c) of Figure 5.12 we see a similar general pattern of results as the lagged networks for the ideal time series in Figure 5.8 in section 5.3.5. In the healthy regime, the negative variables have much stronger cross-lagged effects on the positive variables, and vice versa for the unhealthy regime. However, we see even more differences between regimes in this case than we did for the ideal time series. For instance, in the healthy regime, the within-valence and auto-regressive relationships for the negative variables is much stronger than for the positive variables, a pattern which is flipped for the unhealthy regime. In both regimes, the within-valence cross-lagged parameters are roughly equal to the auto-regressive effects of the variables involved. The estimated means of each regime are given as $\hat{\mu}_2 = \{4.31, 4.31, 1.87, 1.87\}$ for the healthy state and $\hat{\mu}_1 = \{1.74, 1.71, 4.44, 4.62\}$ for the unhealthy state.

We can see from this that the TVAR(1) model for the ESM data succeeds in recovering some global characteristics of the system. Specifically, the estimated mean vectors capture approximately the position of the two stable fixed points (characteristic 2), and the estimated threshold correctly captures the position of the unstable fixed point in the Cheerful dimension. However, the recovery of this characteristic comes with the same caveats as described in Section 5.3.5, in that the use of a univariate threshold for this particular configuration of the true system happens to be a good approximation of the unstable fixed point in multi-dimensional space.

Regarding the microdynamics, the lagged parameters in each regime approximately capture that there are reinforcing effects within valence and suppressing effects between valence (characteristic 5). Otherwise, however, the recovery of microdynamic relationships performs worse than for the ideal time series, as expected. As was the case for the VAR(1) model, the regime-specific lagged parameters here again reflect global characteristics at the 90 minute time-scale rather than microdynamics. Partitioning the joint densities in panels (a) and (b) of Figure 5.11 using a threshold does not aid us in any way to reproduce microdynamic dependencies which are absent due to the low sampling frequency. Thus, the asymmetry in parameter values across regimes has to be a function of the global characteristics, influenced by both the different variances around the fixed points

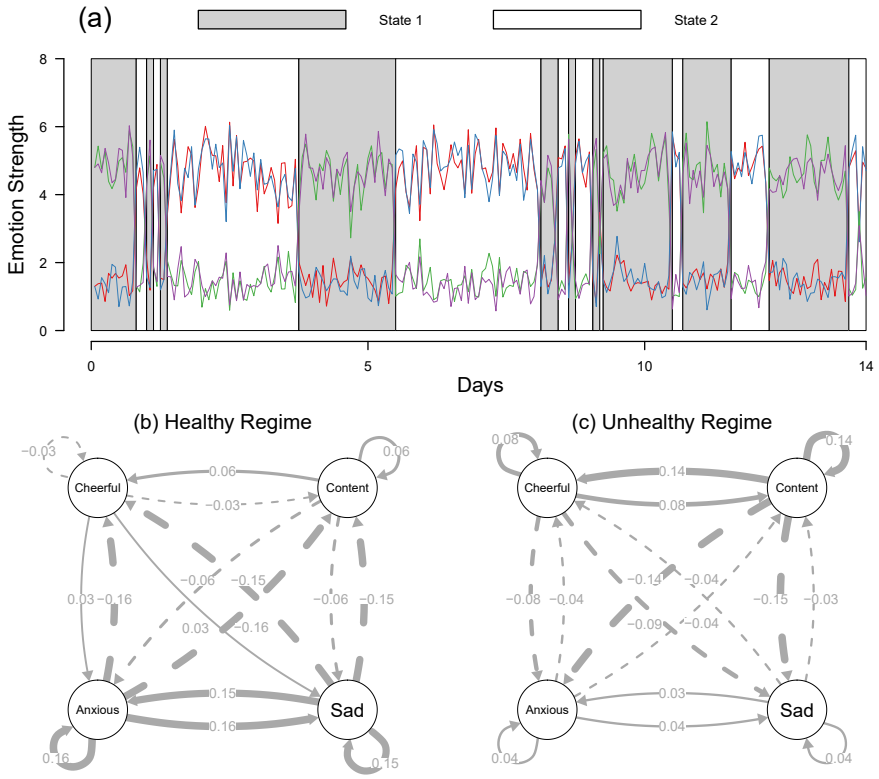


Figure 5.12: Panel (a) shows the first two weeks of the time series, with observations shaded in either grey or white as a function of whether $x_{1,t-1}$ is above or below the threshold $\hat{\tau} = 2.796$. Panels (b) and (c) show the estimated VAR(1) parameters as lagged networks in the healthy (white) and unhealthy (grey) regimes respectively.

in each state (i.e., high variance for positive emotions, low variance for negative emotions in the healthy state, and vice versa) and those observations which jump from one fixed point to the other across consecutive measurement occasions, as discussed in the previous section.

As we did throughout Section 5.3, we could evaluate how well this model describes the bistable system by generating data from it. Notably, the dynamics defined by $\Phi^{(1)}$ and $\Phi^{(2)}$ reflect an *unstable* system in both regimes: The eigenvalues of both contain a value outside the unit circle (i.e., with absolute value greater than one) (Hamilton, 1994). This means that, if we were to generate data using these parameters, the time series would always diverge towards infinity. This instability also precludes us from making any statement regarding the variance of positive and negative emotions in each regime (characteristic 3), as the long run variances implied by the model are infinite. As such, we can say that overall, the set of estimated parameters for the TVAR(1) based on the ESM time series are a poor characterisation of the microdynamics of the model at any time-scale.

In summary, the TVAR(1) model fitted on the ESM time series still picks up a global characteristic of the system, but the recovery of microdynamic characteristics fails. In fact, the relationship between the estimated lagged parameters and the characteristics of the system was much more opaque than in the ideal data case, and our ability to generalize from the estimated parameters to the behavior of the system at any time scale was considerably worse than in the ideal case. Again here, we should note that the only difference between the ideal and ESM time series is the sampling frequency. Fundamentally, the results here indicate that, if we do not have a sufficiently high sampling frequency, then fitting increasingly complex models, or extensions to simpler models such as the TVAR(1), does not aid us in recovering the characteristics we are interested in: Even when we have an arbitrarily large *number* of observations, we fail to recover basic characteristics of the microdynamics due to the spacing between measurements.

5.4.4 Differential Equation Model Building

In this section we will examine whether the DE model-building procedure described in Section 5.3.6 also succeeds in recovering the bistable system when applied to the emulated ESM dataset. Recall from Section 5.3.6 that for the ideal time series, this method succeeded in recovering the microdynamics of the system: The global characteristics were also considered to be recovered as the global characteristics implied by the estimated model (bistability, position of fixed points) matched up with the actual global characteristics of the underlying system.

5.4.4.1 Model Building Procedure

Similarly to Section 5.3.6 we first estimate the derivatives directly from the data by differencing the time series, and then search for the best fitting model by fitting a series of regression models with increasing complexity. Table 5.3 displays the fit of seven increasingly complex regression models, evaluated using the mean out-of-bag proportion of explained variance $\mathcal{R}^2$.

Model A ($\mathcal{R}^2 = 0.13991$), Model B ($\mathcal{R}^2 = .16827$) and Model C ($\mathcal{R}^2 = 0.16928$) are the same models as introduced in Section 5.3.6. However, since we did not observe a clear drop in $\mathcal{R}^2$ as we increased model complexity from Model A to Model C, we also assess the fit of four additional models. Model D adds cubic main effects $x_1^3, \dots x_4^3$ as predictors, increasing the model fit to $\mathcal{R}^2 = 0.19455$. Model E adds four three-way interactions ($x_i \times x_j \times x_k, i \neq j \neq k$) to this, further increasing the model fit to $\mathcal{R}^2 = 0.19940$. Adding yet more three-way interactions ($x_i \times x_j \times x_k, \forall (i, j, k) \in p$) in Model F still increases model fit ($\mathcal{R}^2 = 0.19940$), as does adding all possible four-way interactions in Model G ($\mathcal{R}^2 = 0.20420$). As it is not possible to specify more unique product interaction terms, we consider Model G to be our final model.

Model	$\frac{dx_{i,t}}{dt} \sim a + r_i x_i + \ldots$	q	$\mathcal{R}^2$
A	$\sum_{j \neq i} r_j x_j$	5	0.13991
B	$\sum_{j \neq i} R_{ij} x_j + \sum_j^p C_{ij} x_j x_i$	9	0.16827
C	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k$	15	0.16928
D	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_j^p \gamma_j x_j^3$	19	0.19455
E	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_j^p \gamma_j x_j^3 + \sum_{j \neq k \neq l}^p \zeta_j (x_j x_k x_l)$	23	0.19801
F	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_j^p \gamma_j x_j^3 + \sum_{(j,k,l)}^p \zeta_j (x_j x_k x_l)$	35	0.19940
G	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_j^p \gamma_j x_j^3 + \sum_{(j,k,l)}^p \zeta_j (x_j x_k x_l) + \sum_{(j,k,l,m)}^p \eta_j (x_j x_k x_l x_m)$	70	0.20420

Table 5.3: Model fit results for each of the seven models described in text, for the emulated snapshot ESM data. The second column gives the model equation for each variable, q denotes the number of parameters estimated per univariate regression model, and the final column indicates the mean proportion of explained variance $\mathcal{R}^2$, calculated on the hold-out sets of a 10-fold cross-validation scheme (for details see Appendix 5.E)

5.4.4.2 Dynamics and Data Generated by Final Model

Clearly, the model-building procedure for the emulated ESM data failed to recover the functional form of the true bistable system. Furthermore, we have arrived at a final model which is so complex ($4 \times 70 = 280$ vs. $4 \times 6 = 24$ parameters in the true model) that it is close to uninterpretable. In theory we could continue adding complexity to the model in the form of non-linear transformations or spline functions, which we know to be absent from the data-generating mechanism, but which may improve model fit. However, this would make the model even more difficult to interpret.

In the left panel of Figure 5.13 we present the estimated parameters that are also contained in the true model, with full parameter estimates and standard errors shown in Appendix 5.E.3. We can see that the estimates deviate widely from the parameters in the true bistable system. In addition, the estimated parameters in the C matrix fail to capture the sign and relative ordering of all parameters in the true C matrix, though a full evaluation of whether suppressing and reinforcing effects of different sizes are present (i.e., characteristics 5 and 6) is infeasible due to the large number of parameters present in the model. Thus, we can say that this approach fails to recover the microdynamics of the system at least to the degree that they can be interpreted.

While the system did not recover the microdynamics in the sense that it captures the qualitative characteristics of the true bistable system, it could still be the case that this more complex system exhibits global characteristics that are similar to the true bistable system. Similarly to Section 5.3.6, we can evaluate these dynamics by inspecting its vector field, shown on the right-hand side of Figure 5.13. As in the vector field obtained from the ideal data (Figure 5.9), the intersections of the two solution lines (blue and orange) indicate the position of the different fixed points in the shown range of the state space. These fixed points

$$\hat{a} = \begin{bmatrix} -0.07 & -0.16 & 0.31 & 0.20 \end{bmatrix}^T$$

$$\hat{\sigma} = \begin{bmatrix} 0.01 & 0.01 & 0.01 & 0.01 \end{bmatrix}^T$$

$$\hat{R} = \begin{bmatrix} 0.19 & 0.07 & -0.06 & -0.01 \\ 0.16 & 0.18 & 0.01 & -0.04 \\ -0.09 & -0.29 & -0.05 & 0.01 \\ -0.12 & -0.22 & -0.07 & 0.10 \end{bmatrix}$$

$$\hat{C} = \begin{bmatrix} -0.02 & -0.11 & -0.02 & -0.05 \\ -0.10 & -0.02 & -0.01 & -0.05 \\ 0.00 & 0.05 & -0.01 & 0.01 \\ 0.00 & 0.02 & -0.04 & -0.03 \end{bmatrix}$$

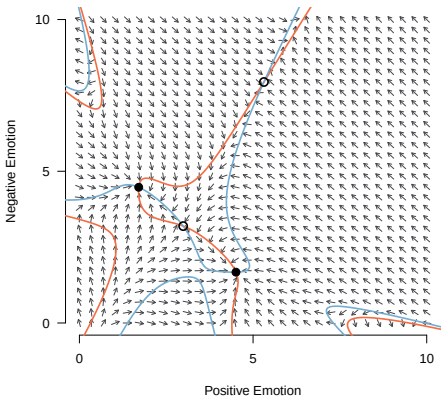


Figure 5.13: Left panel: the parameters estimated from the snapshot ESM time series. Right panel: the vector field defined by the estimated parameters. Solid points indicate stable fixed points and empty points indicate unstable fixed points. The solid lines indicate the values at which derivative of positive emotion (orange) and negative emotion (light blue) is equal to zero. At the points at which the two lines meet, both derivatives are equal to zero and the system remains in this (stable) state.

are further denoted by dots, with filled dots indicating a stable fixed point, and empty dots indicating an unstable fixed point.

We can immediately see from Figure 5.13 that the stability landscape is much more complex than the one of the true bistable system, with high-degree polynomial solution lines, and with four rather than three fixed points. Interestingly, the system correctly identifies that there are two stable fixed points relating to the healthy state ($x_1 = x_2 = 4.51, x_3 = x_4 = 1.67$) and the unhealthy state ($x_1 = x_2 = 1.71, x_3 = x_4 = 4.47$), and that there is an unstable fixed point approximately half-way between those two ($x_1 = x_2 = 2.99, x_3 = x_4 = 3.19$). Despite having an entirely different functional form, the estimated model *does* capture two stable fixed points (characteristic 1) and the approximate position of those fixed points (characteristic 2). This shows that Model G performs well in capturing the characteristics of the system for emotion values that were observed in the time series, that is, near the two stable fixed points.

Crucially, however, we cannot say that this system recovers the global dynamics of the true system, not least because the system contains an additional unstable fixed point at ($x_1 = x_2 = 5.31, x_3 = x_4 = 7.94$), which is not present in the true bistable system. The presence of this unstable fixed point means that if, for instance, both negative and positive emotions take on a high value simultaneously, then the system enters an unstable region and diverges to infinity. If we examine the behavior of the system even further outside the range of observed values ($-\infty > X > 0$ and $10 < X < \infty$) even more fixed points and regions of stability and instability can be found. We can further demonstrate these dynamics by generating data from Model G. Figure 5.14 shows a time series generated from

the difference-form of Model G (i.e., with a step size equal to that of the observed data).[4] We see that the process moves between the healthy and unhealthy fixed point for the first ten days, exhibiting the bistable behavior we see in the true system. However around the eleventh day, the stochastic input is large enough to move the system to an unstable region in the vector field and which leads the system to diverge.

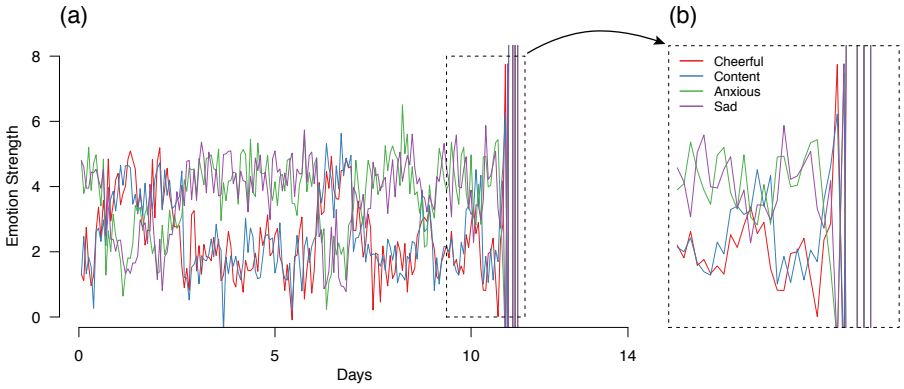


Figure 5.14: Data generated from the estimated DE model, with the same initial values as the "ideal" data

Note that the complexity of the final model here is not a result of over-fitting the data, as we performed model selection based on the out-of-bag $\mathcal{R}^2$, an approximation of the out-of-sample $\mathcal{R}^2$. Rather, the complexity of this model can be attributed to two factors. First, due to the low sampling frequency, our approximation of the derivative at each point in time is poor. The second, as we discussed in Section 5.3.6.3, is that given the spacing between observations, the best one can hope for is to approximate the integral solution to the data-generating equation, which is likely of a highly complex functional form. The ability of the misspecified Model G to reproduce some characteristics in regions where we have observed data can be attributed to the high flexibility afforded by the many non-linear terms. In that sense, this behavior is highly comparable to the problem of using a high-degree polynomial regression model to make predictions outside of the range of observed values. The vector field in Figure 5.13 is constructed by obtaining predicted values for the derivatives across a grid of input values and as such, it is unsurprising that the vector field is accurate where the input values are close to the observed data, and inaccurate elsewhere.

In summary, we do not at all recover the functional form or parameters of the system; we do recover some of the global characteristics and behavior of the

[4]This is obtained by re-fitting the differential equation using the unscaled difference $x_{i,t+1} - x_{i,t}$ as the outcome variable, leading to equivalent results with parameters approximately scaled by $dt = 90$. The residual variance used is the estimated residual variance scaled down to .65 the magnitude, to account for the non-normal residual distribution. Using the estimated residual standard deviation results in shocks which immediately move the system into an unstable region.

system in the region where we have observations, capturing that there are two stable fixed points and one unstable fixed point, and their locations. However, the estimated model also implies the presence of at least one extra unstable fixed point, which has major implications for the dynamics of the model, implying divergent behavior. Thus, the estimated model implies fundamentally different microdynamic and global characteristics. Based on the simulated data in Figure 5.14, it does not seem that we correctly capture the variability around these fixed points, or the frequency of transitions, as any reasonable simulation of data from this model eventually leads the system to diverge. Crucially, we fail in recovering an interpretable approximation of the data-generating model. As such, it is not feasible to assess whether there are truly suppressing effects between valences and reinforcing effects within valences, or the relative size of these effects (characteristics 5 and 6).

5.4.5 Summary: Analysis of ESM Time Series

In this section, we aimed to investigate to which extent lowering the sampling frequency affects the the ability of our considered methods to recover the bistable system. Our findings are summarized in Table 5.4.

	Bistability (1)	Position (2)	Variance (3)	Transitions (4)	Suppr./Reinf. (5)	Relative Size (6)	Time-constant (7)
Data Visualization	✓	✓	✓	✗	✗	✗	✗
HMM	✓*	✓	✓	✓	✗	✗	✗
Lag-0 / GGM	✗	✗	✗	✗	✗	✗	✓*
Lag-1 / VAR(1)	✗	✗	✗	✗	✗	✗	✓*
TVAR(1)	✓*	✓	✓	✓	✗	✗	✗
DE-Estimation	✗	✗	✗	✗	✗	✗	✗

Table 5.4: Summary of which method recovered which of the seven qualitative characteristics listed in Section 5.2.3 from the ESM time series. The first four characteristics are global dynamics, the last three are microdynamics. The check marks with asterisk indicate that the method includes the characteristic as a model assumption, and can therefore not be considered recovered from the time series.

Our main findings are that, in general, we remain able to recover global characteristics of the system using simple methods, but that we are completely unable to recover any of the microdynamics. We saw that each approach which aimed to capture microdynamic characteristics either deteriorated dramatically in performance (for the VAR and TVAR approaches) or broke down altogether (for the DE model building approach) as soon as we applied them to a time series obtained with a more realistic sampling frequency. This is despite the fact that the time series we used in this section can be considered a highly idealized approximation to ESM time series, in terms of the number of observations and the quality

of measurements, suggesting that sampling frequency is a fundamental barrier to inference which needs further investigation. As a side result, we can say that lowering the sampling frequency typically made it much more difficult to interpret and understand the results of different methods: In particular for methods which involved lagged relationships of some kind (i.e. the VAR, TVAR and DE approaches), it was difficult to ascertain precisely what features of the system the estimated parameters reflected.

The recovery of global characteristics was more successful. Using data visualization and the Hidden Markov Model it was still possible in principle to learn about the position, variance around and frequency of transitions between fixed points, and the threshold estimate form the TVAR model succeeded in capturing the unstable fixed point. Finally, the predictions made by the best-fitting differential equation model did allow us to get some tentative indication of bistable behavior, and the possible location of stable fixed points. However, the resulting model suffered from a high degree of complexity, limiting both substantive interpretation and our ability to extrapolate the model parameters to predict the behavior of the system under different conditions.

In summary, the results in this section call into question to what extent it is possible to investigate moment-to-moment microdynamics using data sampled at a rate typical of ESM studies. We have showed that interpreting model estimates from ESM time series as reflecting the microdynamics can be highly misleading, when the process of interest is varying at a higher frequency than the sampling frequency. Although the recovery of global characteristics is more promising, we remind the reader that the time series considered here is still highly idealized, with essentially infinite sampling size, and so the performance of these methods should be considered an upper bound on performance in any realistic situation.

5.5 Discussion

In this paper we explored to what extent dynamical systems models can be recovered from psychological time series by investigating two successive questions. First, how well does a set of popular and more advanced methods recover (characteristics of) a basic bistable system with an ideal data set sampled at extremely high sampling frequency (every six seconds)? And second, how is the performance of each method affected when reducing the sampling frequency to one measurement every 90min, which is typical for ESM studies.

When analyzing the ideal time series we found that the popular VAR model (and the GGM fitted on its residuals) can in principle not recover the global dynamics of the true bistable system, and only recovers some of its microdynamics. However, we showed that descriptive statistics, data visualization and statistical models which are based on mixtures (the HMM and threshold VAR) were able to capture the global dynamics of the bistable system. The only method that recovered the full bistable system was a differential equation (DE) model building procedure. Reducing the sampling frequency from every six seconds to every 90 minutes affected the considered methods differently. The VAR model and its ex-

tensions no longer recover any microdynamics, and the DE-estimation procedure fails. However, descriptives, data visualization and appropriate statistical models still recover the global dynamics. Overall, our analysis therefore suggests that it is neither possible to estimate dynamical systems directly from realistic time series, nor is it possible to reliably infer its microdynamics from the parameter estimates of statistical models.

5.5.1 Implications for Complex Systems Approaches to Studying Mental Disorders

Our results raise fundamental questions about how to study mental disorders from a complex systems perspective. First, they show that it is unclear what exactly one can in principle conclude from statistical models estimated from psychological time series about an underlying dynamical system. Clearly, these models are always misspecified (i.e., do not include the true system as a special case), so one cannot hope to directly recover the underlying dynamical system. More surprisingly, however, recovering the qualitative characteristics of the true system also turned out to be difficult. While it was possible to recover the global dynamics, no statistical model correctly recovered the microdynamics. For example, the VAR model fundamentally cannot capture the global characteristics (e.g., location of fixed points) of the true bistable system and only recovered some of its microdynamics (e.g., reinforcing vs. suppressing effect between two variables). This is a problem for the emerging framework of studying mental disorders as complex systems, because one is typically interested in the microdynamics (the "mechanics") of a disorder because one hopes to intervene on them. In contrast, it is usually less clear how interventions can target global dynamics, since they can be seen as the aggregate behavior implied by the microdynamics. Especially the failure of the popular VAR model to correctly recover the qualitative nature of the microdynamics in the true model is concerning, because it calls into question whether it allows any reliable conclusions about an underlying dynamical system. It therefore seems to be an open question how useful VAR models and other statistical models are to studying mental disorders from a complex systems perspective.

Second, the analysis of the ESM time series raises the question of which process can be recovered with which sampling frequency. While we were still able to recover the global dynamics of the system, each method that provides some approximation of the microdynamics was strongly affected by sampling only every 90min instead of every six seconds. The qualitative characteristics of the VAR and TVAR models were even less in agreement with those of the true model, and the DE-estimation method, which was the only fully successful method in the ideal data case, returned a model with uninterpretable parameters and incorrect global- and microdynamics. Thus, our results suggest what also seems intuitive: It is extremely difficult — perhaps impossible — to recover microdynamics at a time scale that is much smaller than the sampling frequency. This intuition is also in line with sampling theorems from the field of signal processing. For example, the Nyquist-Shannon sampling theorem states that a sine wave (a process

much simpler than our bistable system) that completes one cycle within, say, 2 minutes, has to be sampled at least every minute to be recovered (e.g., Marks, 2012; Papoulis & Pillai, 2002). This suggests that it is futile to try use a time series sampled every 90 minutes to directly recover dynamics of emotions that operate on a time scale of seconds or minutes (Houben, Van Den Noortgate, & Kuppens, 2015) or even from moment to moment (Wichers et al., 2015). However, this also means that ESM time series can certainly be used to recover processes that unfold at a time scale of several hours or days.

To summarize, we identified two fundamental barriers to studying mental disorders from a complex systems perspective. First, even with extremely high sampling frequency it is generally unclear how to make inferences from a statistical model to an unspecified dynamical systems model. Second, the sampling frequency of the data collection constrains the type of processes one can recover. Specifically, a process can only be recovered if the sampling frequency is sufficiently high. Clearly, these are profound problems every empirical discipline struggles with and no simple answers can be expected. Indeed, they might imply that studying some aspects of mental disorders will always remain out of reach. That said, we believe that much progress *can* be made by studying mental disorders as complex systems and that acknowledging and studying the above issues allows one to do so more efficiently. As a way forward, in the following section we suggest a new research strategy based on proposing substantively plausible dynamical systems, which opens up avenues to creatively tackle the two problems identified in this section.

5.5.2 Moving Forward: Proposing Plausible Dynamical Systems Models

A more abstract perspective on the first problem identified in the previous section is the following: We have parameters of a statistical model which we estimated from a time series sampled from some system, and we hope to infer some characteristics (e.g., global or microdynamics) of the data-generating system from them. The problem, however, is that the mapping from parameters of statistical model to the parameters and structure (and the implied dynamics) of the true model is unknown. Thus, this inference cannot be made. The main reason this mapping is unknown is the trivial reason that no true dynamical system model is specified.

We propose that, in order to overcome this fundamental problem, researchers must begin the research process by proposing a "first guess" model of the dynamical system. While this is clearly difficult and the validity of this model should certainly be questioned, this approach has one major advantage: It is much clearer how to draw conclusions from descriptive statistics, data visualizations or statistical models about the underlying dynamical systems model. This is because one can generate time series from the "first guess" model and fit a statistical model of choice; that way, one always knows which statistical model is implied by the dynamical systems model. This implied model can then be compared to the model fitted to corresponding empirical data. If the implied model

and the empirical model are in agreement, we have tentative evidence that the dynamical system model is correct; if not, we can use the nature of the disagreement to improve the dynamical system model. Clearly, this modeling approach, which is typical to more quantitative disciplines such as physics, chemistry and biology, is different to the statistical modeling framework most psychological researchers are familiar with. On the one hand these formal dynamical systems models are harder to build, since they cannot be estimated directly from the data. On the other hand, they are powerful enough to be plausible for complex phenomena such as mental disorders, and have additional benefits such as synthesizing knowledge, revealing unknowns, laying open hidden assumptions and enabling checking of the internal consistency of a model (Epstein, 2008; Lewandowsky & Farrell, 2010; Smaldino, 2017).

This modeling approach also allows to tackle the problem of sampling frequencies that are too low to recover the process of interest directly, as one can generate a time series from the specified dynamical systems model and reduce the sampling frequency to a level that is also available in empirical data. Then, similarly to above, one can again compute the statistical model of choice that is implied by the dynamical systems model with a given sampling frequency, compare it to the corresponding model fit on empirical data, and in the case of disagreement adapt the dynamical systems model accordingly. Of course, this approach is not a panacea. Less information is available when the sampling frequency is low, which makes model identification more difficult. However, specifying an initial dynamical systems model allows one to gauge how difficult it is to recover a given type of process on a given time scale with a given sampling frequency.

In addition, starting out with a dynamical systems model also allows to study the *measurement function* that defines the mapping from the variables in the dynamical systems to the obtained measurements, a topic we only touched on briefly in this paper. In our emulated ESM time series we took the measurement function to return the exact values of variables at the time point of measurement. However, different questions imply different measurement functions. For example, if the phrasing of a particular question refers to the entire period since the last measurement, one could instead define the measurement as a function of the variable values since the last measurement, such as the average. Next to formalizing which experiences an ESM question refers to exactly, defining a measurement function also allows to formalize known response and memory biases, such as the recency effect (Ebbinghaus, 1913/2013).

Finally, having a plausible dynamical systems model allows one to explicitly address a behavior that has been largely ignored in the psychological time series modeling literature, that is, the fact that humans sleep. Sleep interacts with essentially everything physiological and psychological, is part of the definition of several mental disorders (e.g., Major Depression) and related to many more (e.g., Walker, 2017). Thus, for many mental disorders, it seems necessary for a plausible model to include sleep. This may also allow using existing data in new ways, because data around the "day-night shift" does not have to be excluded anymore, but instead can be used to test hypotheses about the sleep-related assumptions

of the dynamical systems model.

Clearly, this brief outline of the proposed modeling approach leaves many important questions unanswered: Where should the initial "first guess" dynamical system come from? How to formalize different substantive aspects in a dynamical systems model? Which statistical models should one choose to test which implications of the dynamical systems model? Given some disagreement between predicted and observed statistics, how should one adapt the existing dynamical systems model? These and other questions are difficult ones and answering them requires the combined creativity of a large research community. Nonetheless, a more detailed account of our proposed new modeling approach would be desirable. However, since such a detailed account is beyond the scope of the present work, we address it in a forthcoming paper.

5.5.3 Limitations

Several limitations of our work require discussion. First, our goal was to explore to which extent one can recover (bistable) dynamical systems for mental disorders from psychological time series. However, we only studied a single bistable system. Therefore, it could be that the fundamental problems identified in the paper and summarized in Section 5.5.1 are in fact a particularity of the chosen bistable system. This, however, seems extremely unlikely: First, because we identify the problems in our paper as examples of well-known issues such as model misspecification and sampling systems with a sampling frequency that is sufficient for recovery. Second, the bistable system we chose is arguably the simplest bistable system for four variables one can find. Choosing a different model therefore results most likely in choosing a more complex model, and our intuition is that the methodological difficulties discussed in this paper become more and not less relevant in such models.

Second, a more specific criticism of our bistable system could be that the time scale of the process is unrealistically small, and we therefore exaggerated the problem of recovering dynamics of psychological processes from ESM time series. We agree that it is possible that some psychological processes are easier to recover from ESM data than the dynamical system used in this paper. Thus, strictly speaking, we only showed that it is impossible to recover a system if the sampling frequency does not appropriately match the time scale of the system. In principle, it is therefore an open question whether there is a mismatch between the time scale of the system of interest and the available sampling frequency. However, intuition — and the sampling theorems such as the one mentioned in Section 5.5.1 — strongly suggest that it is impossible (or at least very difficult) to recover a process that operates at a time scale of seconds or minutes from an ESM time series that is measured every 1.5 hours. Clearly, however, our investigation is only a first treatment of the important topic of sampling frequency, and much work on it is required to establish a tight connection between psychological time series and dynamical systems models.

Third, one could reverse the argument in the previous two paragraphs and argue that our model is so ideal that many analyses perform better than in most

realistic applications. This is certainly the case for the Threshold VAR model, which performs well only because of the simple dynamics of the bistable system as we discussed in Section 5.3.5. Other examples are the descriptive statistics and data visualization which may not be as insightful if fixed points are closer to each other and if there is more noise in the system. Also, the two-step approach to estimating the differential equations in Section 5.3.6 may work less well for a more complicated model. Thus, we would agree with this assessment, however chose to use a simple bistable system in order to make the paper more accessible to applied researchers.

Fourth, we analyzed a bistable system whose structural parameters do not change over time. However, much of the framework of considering mental disorders as complex systems is based on the idea that pathology is defined with respect to a structural change in the underlying system, and therefore structural change is of central interest. We expect that structural change renders the recovery of a system more difficult, and we therefore did not include this feature in order to keep the paper at a reasonable length. However, we believe that future methodological research into how to recover such structural changes both in principle and with realistic time series would be extremely helpful to better understand phenomena such as early warning signals (Scheffer et al., 2009; van de Leemput et al., 2014) and more generally structural change in mental disorders.

Fifth, in order to estimate a differential equation from data, we took a rather simple two-step approach based on local linear approximation of the derivative (cf. Boker, Deboeck, et al., 2010). This approach involves first estimating the derivative itself using scaled difference scores, and then using this derivative as an outcome variable in a regression model. While this method benefits from being extremely simple to implement, we could expect that it would perform poorly in the presence of low sampling frequency as the quality of the derivative approximation degrades (as we noted in Section 5.3.6.3 and observed in Section 5.4.4). There are multiple alternative approaches to estimating DE equations which we did not consider here. For example, approaches based on numerical integration of the DE equation during estimation, such as implemented in *dynR* (Ou et al., 2019) and *stan* (Carpenter et al., 2017) (with additional functionality in the *ctsem* package; Driver et al., 2017) may in general perform better than the two-step procedure when the sampling frequency is low. However, for the analysis shown in the present paper, neither the *ctsem* nor *dynr* package performed better than the two-step approach. In general, however, more research is needed to map out which method deals best with the problem of low sampling frequencies.

Lastly, throughout our paper we studied how well certain analysis methods can recover the true bistable system *in principle*. We did this by studying the population properties of these methods, that is, the situation in which one has essentially infinite sample size, which we approximated with a huge number (201600) of measurements. This was necessary in order to study the more fundamental questions of (1) whether a given method can recover our bistable system in principle and (2) whether a given method can recover our bistable system based on a time series with realistic sampling frequency. We did this because it would be meaningless to study the performance of a method as a function of sample size,

if the method alredy fails with infinite sample size. Clearly, however, to apply any of the methods we studied in practice, one has to know how reliable they are with which sample size, and much more research is necessary to map our these sample size requirements (e.g., Dablander, Ryan, & Haslbeck, 2019).

5.5.4 Summary

In the present paper we identified two fundamental problems involved in studying mental disorders from a complex systems perspective: first, it is generally unclear what to conclude from a statistical model about an unspecified underlying complex systems model. Second, if the sampling frequency of a time series is not high enough, it is futile to attempt to recover the microdynamics of the underlying complex system. In response to these problems, we proposed a new modeling strategy that takes an initial substantively plausible dynamical systems model as a starting point, and develops the dynamical systems model by testing its predictions. In this approach it is much clearer what we can learn from data and statistical models about an underlying dynamical system, and in addition it provides avenues to move the field forward by formalizing the sampling process, measurement, response and memory biases, measurement reactivity and the influence of sleep.

Appendix 5.A Determining Fixed Points

In this section we show how to compute the fixed points of the deterministic part of our model, which we report in Section 5.2.1. The fixed points of a set of differential equations is found by setting all equations to zero and solving that system. In our case this means solving the nonlinear system of equations:

$$
\begin{aligned}
0 &= r_1 x_1 + \sum_{j=1}^{4} C_{1j} x_j x_1 + a_1 \\
0 &= r_2 x_2 + \sum_{j=1}^{4} C_{2j} x_j x_2 + a_2 \\
0 &= r_3 x_3 + \sum_{j=1}^{4} C_{3j} x_j x_3 + a_3 \\
0 &= r_4 x_4 + \sum_{j=1}^{4} C_{4j} x_j x_4 + a_4
\end{aligned}
$$

Since we have $r_1, r_2 = 1$ and $a = [1.6, 1.6, 1.6, 1.6]$ in all studied situations, we fill in those values and write out the summation:

$$
\begin{aligned}
0 &= x_1 + C_{11} x_1 x_1 + C_{12} x_1 x_2 + C_{31} x_1 x_3 + C_{41} x_1 x_4 + 1.6 \\
0 &= x_2 + C_{21} x_2 x_1 + C_{22} x_2 x_2 + C_{23} x_2 x_3 + C_{24} x_2 x_4 + 1.6 \\
0 &= r_3 x_3 + C_{31} x_3 x_1 + C_{32} x_3 x_2 + C_{33} x_3 x_3 + C_{34} x_3 x_4 + 1.6 \\
0 &= r_4 x_4 + C_{41} x_4 x_1 + C_{42} x_4 x_2 + C_{43} x_4 x_3 + C_{44} x_4 x_4 + 1.6
\end{aligned}
$$

We can exploit the symmetries in r and C to simplify finding the fixed points. The derivatives of x_1 and x_2 are actually identical, and the derivatives of x_3 and x_4 are identical. Thus also their integrals are identical. Thus, we can substitute x_1 into x_2, and x_3 into x_4 to arrive at a simpler 2-dimensional system. Making the substitutions, and filling in the parameter values, the differential equations then reduce to

$$
\begin{aligned}
0 &= 1 x_1 - 0.2 x_1^2 + 0.04 x_1^2 - 0.4 x_1 x_2 + 1.6 \\
0 &= r_3 x_3 - 0.2 x_3^2 - 0.4 x_3 x_1 + 0.04 x_3^2 + 1.6
\end{aligned}
$$

where r_3 is the stress level for which the fixed points should be computed.

We now solve these systems for a number of stress values (r_3) using Mathematica (Wolfram Research, Inc., 2019). This way, we computed the fixed points shown in Table 5.5, which are displayed in panel (a) of Figure 5.1 in Section 5.2.1.

Stress	Healthy:PE	Healthy:NE	Unhealthy:PE	Unhealthy:NE	Unstable:PE	Unstable:NE
0.90	5.28	1.15				
0.91	5.26	1.16				
0.91	5.24	1.17				
0.92	5.22	1.18				
0.93	5.19	1.19				
0.93	5.17	1.20				
0.94	5.15	1.22				
0.95	5.12	1.23				
0.95	5.10	1.24				
0.96	5.08	1.26				
0.90	5.28	1.15				
0.91	5.26	1.16				
0.91	5.24	1.17				
0.92	5.22	1.18				
0.93	5.19	1.19				
0.93	5.17	1.20				
0.94	5.15	1.22				
0.95	5.12	1.23				
0.95	5.10	1.24				
0.96	5.07	1.26	1.83	3.96	2.03	3.66
0.97	5.05	1.27	1.66	4.25	2.25	3.30
0.97	5.02	1.29	1.57	4.42	2.39	3.22
0.98	4.99	1.31	1.50	4.56	2.50	3.10
0.99	4.96	1.33	1.45	4.69	2.61	2.99
0.99	4.92	1.34	1.40	4.79	2.71	2.89
1.00	4.89	1.36	1.36	4.89	2.80	2.80
1.01	4.85	1.39	1.33	4.98	2.90	2.72
1.01	4.80	1.41	1.30	5.06	2.99	2.64
1.02	4.76	1.44	1.27	5.15	3.09	2.56
1.03	4.71	1.47	1.24	5.23	3.19	2.48
1.03	4.65	1.50	1.22	5.30	3.29	2.40
1.04	4.59	1.54	1.19	5.38	3.40	2.32
1.05	4.51	1.58	1.17	5.45	3.52	2.23
1.05	4.41	1.65	1.15	5.52	3.67	2.12
1.06	4.24	1.75	1.13	5.59	3.87	1.98
1.06			1.12	5.63		
1.07			1.11	5.66		
1.07			1.09	5.72		
1.08			1.08	5.79		
1.09			1.06	5.85		
1.09			1.04	5.91		
1.10			1.03	5.98		

Table 5.5: Fixed points of the emotion model for different values of stress (rows), rounded to two decimals. The 2nd and 3rd columns refer to the fixed points of the healthy fixed points for positive and negative emotions; the 4th and 5th columns reger to the unhealthy fixed points; and the last two columns refer to the unstable fixed point.

Appendix 5.B Mean-Switching Hidden Markov Model

In this appendix we provide additional details with respect to the specification of the mean-switching Hidden Markov Model, and using model selection to obtain the number of components, described in Section 5.3.2

5.B.1 Model Specification

The mean-switching Hidden Markov Model is denoted

$$P(X, S|\mu, \sigma) = \pi_i \mathcal{N}(X_1) \prod_{t=1}^{T-1} A_{ji} \mathcal{N}(X_{t+1}),$$

where $X = \{X_1, \ldots, X_T\}$ is a matrix of $p-$variate elements X_j, $S \in \{1, \ldots, K\}^T$ is a vector of length T indicating the state at each time point, π_i is the probability of being in state $i \in \{1, \ldots, K\}$, $A_{i,j}$ is the probability of transitioning from state i to state j, and μ, σ parameterize the multivariate Gaussian distribution $\mathcal{N}$ with zero covariances.

In Section 5.3.2 we chose $K = 2$ components, and fix the covariances of the Gaussian distribution to zero. Since we model four variables, this gives us 2×4 means and 2×4 standard deviations. The transition matrix A has three parameters since the last one is determined by the remaining three. Similarly, the marginal probabilities π_1, π_2 are determined by A and therefore do not count as additional parameters. We therefore fit a model with 19 freely estimated parameters.

5.B.2 Model Selection for Mean-Switching HMM

In Section 5.3.2 we inserted bistability as an assumption in the model by specifying that the HMM exhibits two states, and therefore the HMM does not provide us any evidence with respect to which number of states represents the data best. This can be done by performing model selection between HMMs with different numbers of states.

A popular way to select between mean-switching HMMs / Gaussian mixtures is the Bayesian Information Criterion (BIC) (Schwarz et al., 1978), because it has been shown to consistent in estimating Gaussian mixtures (Leroux, 1992), and has outperformed other information criteria (including the AIC) in simulations (R. J. Steele & Raftery, 2010). Here we fit HMMs with $K \in \{1, \ldots, 10\}$ and report the BIC values in Figure 5.15.
We see that the BIC is highest for $K = 1$ and then decreases for larger K, however the the change in BIC becomes less and less when adding additional states. Since we know from the true bistable system that the number of states is $K = 2$, we see that the BIC does not select the true number of states. The reason is that the BIC has been shown to be a consistent estimator of K if the data is generated from

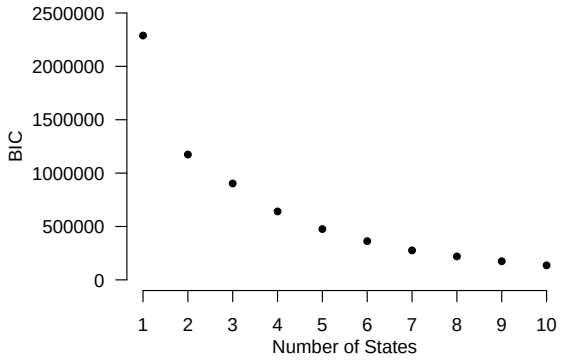


Figure 5.15: The figure depicts BIC values as a function of the number of states K, for HMMs fitted to the ideal data.

a Gaussian mixture. However, in the present case the data is generated from a bistable dynamical system. This failed attempt at model selection based on statistical models again highlights the problems of using misspecified statistical models to make inferences about dynamical systems models.

Appendix 5.C Data Generated from Estimated Models

In this Appendix we show data generated from estimated models for the time period of two weeks of the original time series.

5.C.1 Mean Switching Hidden Markov Model

Figure 5.16 displays a time series of two weeks generated from the Mean switching HMM estimated in Section 5.3.2:

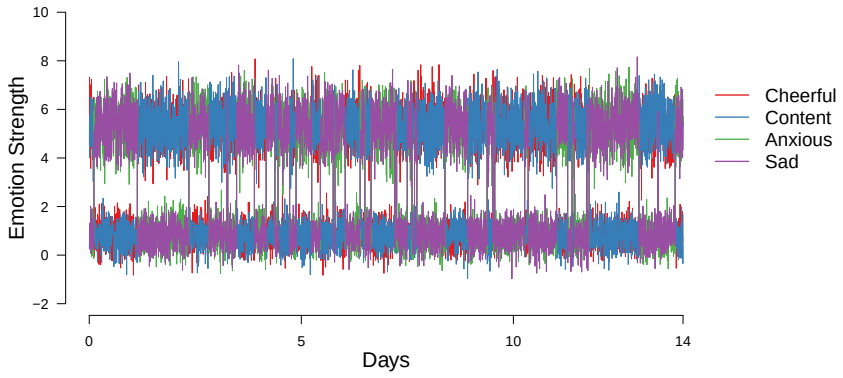


Figure 5.16: A time series of two weeks generated from the HMM estimated in Section 5.3.2.

The generated time series looks similar to the original data in that it switches between the two fixed points at around (1,6) and (6,1). However, there are also differences. In the original data there are less switches that lead to a long-lived change in fixed point, but more switches that are very short-lived. Second, due to the form of the Mean-Switching HMM, there are no "intermediate" observations leading from one fixed point to the other. These observations exist in the original time series (see panel (b) in Figure 5.2).

5.C.2 First-order Vector Autoregressive (VAR(1)) model

Figure 5.17 displays a time series of two weeks generated from the VAR(1) model in Section 5.3.4:

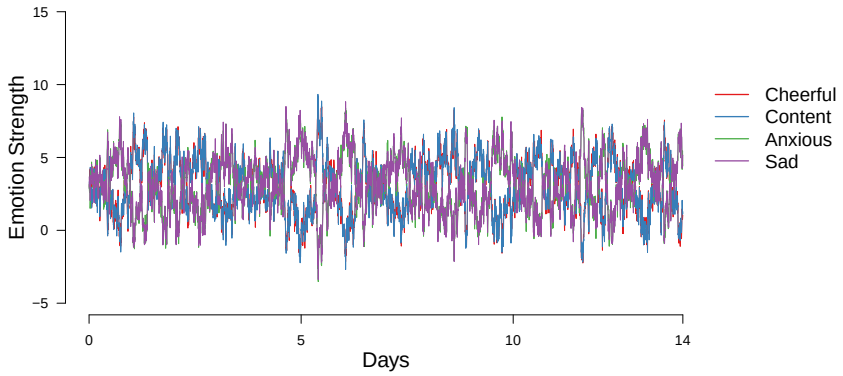


Figure 5.17: A time series of two weeks generated from the VAR(1) model estimated in Section 5.3.4.

The generated data does not show bistability, which is expected because the VAR(1) model exhibits only a single fixed point. What looks approximately like oscillating behavior is a result of the high auto-regressive effects present in the estimated VAR(1) model: given a stochastic input, the high auto-regressive effects ensure that the system is slow to eventually return to equilibrium. This oscillating behavior is also evident in the eigenvalues of $\mathbf{\Phi}$, which consist of one complex conjugate pair (Strogatz, 2015).

5.C.3 Threshold VAR(1) Model

Figure 5.18 displays a time series of two weeks generated from the TVAR(1) model in Section 5.3.5. The data generated from the TVAR(1) model looks similar to the original time series in that the position of the fixed points and the variance around them is very similar. However, the system seems to switch less often between states, and similarly to the data generated from the HMM above, there are much fewer observations on the transitions between states.

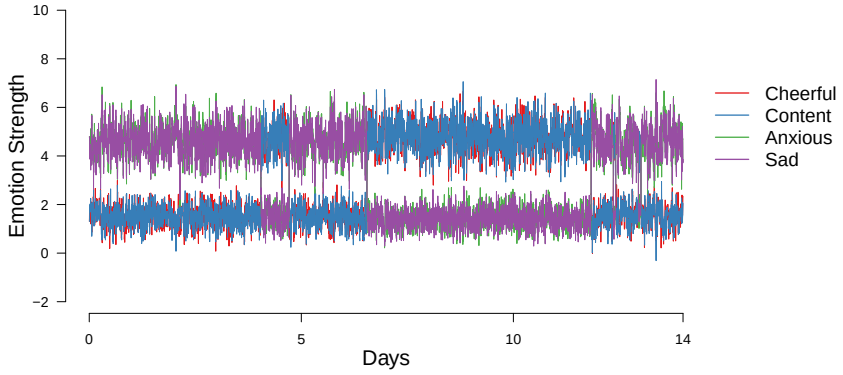


Figure 5.18: A time series of two weeks generated from the TVAR(1) model estimated in Section 5.3.5.

Appendix 5.D Residual Partial Correlations TVAR(1)

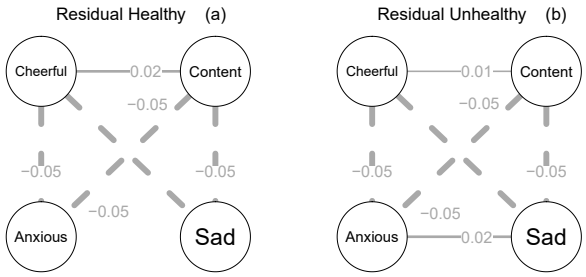


Figure 5.19: Residual partial correlation networks for both regimes in the TVAR model described in Section 5.3.5 in the main text.

Appendix 5.E Differential Equation Model Building

In this appendix we present additional information relating to the two-step DE model building procedure utilized in Sections 5.3 and 5.4. This includes details on how model fit is computed, as well as full model fit results and parameter estimates for each of the models described in the main text.

5.E.1 Evaluating Model Fit

The fit of each model is evaluated with the mean out-of-bag explained variance, referred to throughout as R^2. This metric is calculated using 10-fold cross-

validation. First, the given dataset is randomly partitioned into ten mutually exclusive training and test sets. Second, for each partitioned dataset, regression models A through G, (defined by the expression in the second column of Table 5.6) are fit to the training set four times, once each of the four outcome variables $\hat{dx_i/dt}, \forall i \in \{1,2,3,4\}$. Third, the resulting parameters are then used to predict the values of the outcome variable in the test set $\hat{dx_i/dt}$. The variance of the resulting residuals $VAR(dx_i/dt - \hat{dx_i/dt})$ is then divided by the variance of the outcome variable in the test set, $VAR(dx_i/dt)$ yielding an out-of-bag variance explained for variable i based on model m in partition k, $R^2_{i,k,m}$. Averaging the explained variance across each of the partitions yields an average explained variance for variable i in model m, $R^2_{i,m}$, and averaging this number across all four outcome variables yields the average out-of-bag explained variance for model m.

5.E.2 Ideal Data

In Table 5.6 we show the fit of models A through G for the ideal dataset analysis in Section 5.3.6. In Table 5.7 we show the full parameter estimates, standard errors and $p-$values for the selected model, Model C.

Model	$\frac{dx_{i,t}}{dt} \sim a + r_i x_i + \ldots$	q	R^2
A	$\sum_{j \neq i} r_j x_j$	5	.04464
B	$\sum_{j \neq i} R_{ij} x_j + \sum_j^p C_{ij} x_j x_i$	9	.06874
C	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k$	15	.06870
D	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_j^p \gamma_j x_j^3$	19	.06871
E	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_j^p \gamma_j x_j^3 + \sum_{j \neq k \neq l}^p \zeta_j (x_j x_k x_l)$	23	.06870
F	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_j^p \gamma_j x_j^3 + \sum_{(j,k,l)}^p \zeta_j (x_j x_k x_l)$	35	.06860
G	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_j^p \gamma_j x_j^3 + \sum_{(j,k,l)}^p \zeta_j (x_j x_k x_l) + \sum_{(j,k,l,m)}^p \eta_j (x_j x_k x_l x_m)$	70	.06846

Table 5.6: Model fit results for each of the seven models described in text in Section 5.3.6 for the ideal dataset. The second column gives the model equation for each variable, q denotes the number of parameters estimated per univariate regression model, and the final column indicates R^2, the explained variance, as calculated based on the prediction error on a hold-out set, using 10-fold cross-validation.

	dx_1/dt			dx_2/dt			dx_3/dt			dx_4/dt		
	Est	SE	p	Est	SE	p	Est	SE	p	Est	SE	p
a	1.40	0.13	<.01	1.37	0.12	<.01	1.25	0.12	<.01	1.27	0.12	<.01
$x1$	0.88	0.05	<.01	0.03	0.02	0.19	-0.02	0.02	0.33	0.04	0.02	0.05
$x2$	0.02	0.02	0.34	0.95	0.05	<.01	0.05	0.02	0.02	-0.01	0.02	0.57
$x3$	-0.01	0.02	0.72	0.01	0.02	0.68	0.96	0.05	<.01	0.08	0.02	<.01
$x4$	0.01	0.02	0.51	<.01	0.02	0.80	0.04	0.02	0.10	0.91	0.05	<.01
$x1 \times x1$	-0.18	0.01	<.01	-	-	-	-	-	-	-	-	-
$x1 \times x2$	0.04	0.01	<.01	0.03	0.01	<.01	-	-	-	-	-	-
$x1 \times x3$	-0.17	0.01	<.01	-	-	-	-0.18	0.01	<.01	-	-	-
$x1 \times x4$	-0.18	0.01	<.01	-	-	-	-	-	-	-0.19	0.01	<.01
$x2 \times x2$	-	-	-	-0.19	0.01	<.01	-	-	-	-	-	-
$x2 \times x3$	-	-	-	-0.19	0.01	<.01	-0.19	0.01	<.01	-	-	-
$x2 \times x4$	-	-	-	-0.19	0.01	<.01	-	-	-	-0.18	0.01	<.01
$x3 \times x3$	-	-	-	-	-	-	-0.19	0.01	<.01	-	-	-
$x3 \times x4$	-	-	-	-	-	-	0.03	0.01	<.01	0.02	0.01	<.01
$x4 \times x4$	-	-	-	-	-	-	-	-	-	-0.18	0.01	<.01

Table 5.7: Full parameter estimates, standard errors and p-values for Model B in Section 5.3.6, for the DE model fit to ideal data.

5.E.3 ESM Data

In Table 5.8 we show the fit of models A through G for the emulated ESM dataset analysis, from Section 5.4.4. In Table 5.9 we show the full parameter estimates, standard errors and p–values for the selected model, Model G.

Model	$\frac{dx_{i,t}}{dt} \sim a + r_i x_i + \ldots$	q	R^2
A	$\sum_{j \neq i} r_j x_j$	5	0.13991
B	$\sum_{j \neq i} R_{ij} x_j + \sum_j^p C_{ij} x_j x_i$	9	0.16827
C	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k$	15	0.16928
D	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_j^p \gamma_j x_j^3$	19	0.19455
E	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_j^p \gamma_j x_j^3 + \sum_{j \neq k \neq l}^p \zeta_j (x_j x_k x_l)$	23	0.19801
F	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_j^p \gamma_j x_j^3 + \sum_{(j,k,l)}^p \zeta_j (x_j x_k x_l)$	35	0.19940
G	$\sum_{j \neq i} R_{ij} x_j + \sum_{(j,k)}^p \beta_j x_j x_k + \sum_j^p \gamma_j x_j^3 + \sum_{(j,k,l)}^p \zeta_j (x_j x_k x_l) + \sum_{(j,k,l,m)}^p \eta_j (x_j x_k x_l x_m)$	70	0.20420

Table 5.8: Model fit results for each of the seven models described in text, for the ESM time series, described in Section 5.4. The second column gives the model equation for each variable, q denotes the number of parameters estimated per univariate regression model. The final two columns indicate R^2, the explained variance, as calculated based on the prediction error on a hold-out set, using 10-fold cross-validation, for the snapshot ESM data and the mean-aggregated ESM data, respectively. R^2 for Model G in the mean-aggregated ESM data case was not available due to multicollinearity problems encountered when fitting the model.

	dx_1/dt			dx_2/dt			dx_3/dt			dx_4/dt		
	Est	SE	p	Est	SE	p	Est	SE	p	Est	SE	p
(Intercept)	-0.07	0.73	0.92	-0.16	0.73	0.82	0.31	0.74	0.68	0.20	0.74	0.78
$x1$	0.19	0.31	0.53	0.16	0.31	0.61	-0.09	0.31	0.78	-0.12	0.31	0.71
$x2$	0.07	0.30	0.80	0.18	0.30	0.54	-0.29	0.30	0.34	-0.21	0.30	0.48
$x3$	-0.06	0.31	0.86	0.01	0.31	0.96	-0.05	0.31	0.88	-0.07	0.31	0.82
$x4$	-0.01	0.31	0.97	-0.04	0.31	0.89	0.01	0.31	0.98	0.10	0.31	0.74
$x1 \times x1$	-0.02	0.06	0.78	-0.01	0.06	0.87	-0.03	0.06	0.65	-0.01	0.06	0.81
$x1 \times x2$	-0.11	0.09	0.21	-0.10	0.09	0.22	0.14	0.09	0.10	0.13	0.09	0.13
$x1 \times x3$	-0.02	0.10	0.86	-0.04	0.10	0.66	<.01	0.10	0.99	0.02	0.10	0.80
$x1 \times x4$	-0.05	0.10	0.59	<.01	0.10	0.99	0.01	0.10	0.94	<.01	0.10	0.98
$x2 \times x2$	0.01	0.06	0.83	-0.02	0.06	0.76	0.02	0.06	0.72	0.01	0.06	0.82
$x2 \times x3$	<.01	0.10	0.98	-0.01	0.10	0.94	0.05	0.10	0.58	0.06	0.10	0.57
$x2 \times x4$	-0.01	0.10	0.95	-0.05	0.10	0.60	0.06	0.10	0.57	0.02	0.10	0.86
$x3 \times x3$	0.02	0.06	0.68	0.01	0.06	0.87	-0.01	0.06	0.87	0.02	0.06	0.73
$x3 \times x4$	0.01	0.09	0.92	0.01	0.09	0.92	0.01	0.09	0.94	-0.04	0.09	0.62
$x4 \times x4$	0.02	0.06	0.80	0.03	0.06	0.68	-0.03	0.06	0.69	-0.03	0.06	0.68
$x1 \times x1 \times x1$	<.01	0.01	0.71	<.01	0.01	0.92	0.01	0.01	0.38	0.01	0.01	0.33
$x1 \times x1 \times x2$	0.02	0.01	0.19	0.01	0.01	0.48	-0.01	0.01	0.33	-0.02	0.01	0.15
$x1 \times x1 \times x3$	<.01	0.01	0.95	0.01	0.01	0.61	0.01	0.01	0.64	<.01	0.01	0.82
$x1 \times x1 \times x4$	<.01	0.01	0.78	<.01	0.01	0.71	<.01	0.01	0.73	<.01	0.01	0.81
$x1 \times x2 \times x2$	<.01	0.01	0.87	0.01	0.01	0.49	-0.01	0.01	0.39	<.01	0.01	0.76
$x1 \times x2 \times x3$	0.01	0.02	0.58	0.01	0.02	0.69	-0.02	0.02	0.32	-0.02	0.02	0.27
$x1 \times x2 \times x4$	0.02	0.02	0.25	0.02	0.02	0.22	-0.03	0.02	0.21	-0.02	0.02	0.34
$x1 \times x3 \times x3$	<.01	0.01	0.79	<.01	0.01	0.96	<.01	0.01	0.81	<.01	0.01	0.98
$x1 \times x3 \times x4$	0.01	0.02	0.73	<.01	0.02	0.80	<.01	0.02	0.99	<.01	0.02	0.99
$x1 \times x4 \times x4$	<.01	0.01	0.98	-0.01	0.01	0.62	<.01	0.01	0.80	<.01	0.01	0.75
$x2 \times x2 \times x2$	<.01	0.01	0.92	<.01	0.01	0.96	<.01	0.01	0.98	<.01	0.01	0.82
$x2 \times x2 \times x3$	<.01	0.01	1.00	<.01	0.01	0.87	<.01	0.01	0.79	<.01	0.01	0.79
$x2 \times x2 \times x4$	-0.01	0.01	0.70	<.01	0.01	0.90	<.01	0.01	0.89	<.01	0.01	0.96
$x2 \times x3 \times x3$	<.01	0.01	0.92	<.01	0.01	0.93	<.01	0.01	0.89	-0.01	0.01	0.50
$x2 \times x3 \times x4$	-0.01	0.02	0.70	<.01	0.02	0.79	<.01	0.02	0.83	0.01	0.02	0.59
$x2 \times x4 \times x4$	<.01	0.01	0.91	0.01	0.01	0.63	<.01	0.01	0.83	<.01	0.01	0.78
$x3 \times x3 \times x3$	<.01	0.01	0.95	<.01	0.01	0.93	<.01	0.01	0.97	<.01	0.01	0.51
$x3 \times x3 \times x4$	-0.01	0.01	0.40	-0.01	0.01	0.68	0.01	0.01	0.63	0.01	0.01	0.39
$x3 \times x4 \times x4$	0.01	0.01	0.44	<.01	0.01	0.76	-0.01	0.01	0.56	<.01	0.01	0.73
$x4 \times x4 \times x4$	-0.01	0.01	0.36	-0.01	0.01	0.42	0.01	0.01	0.33	0.01	0.01	0.40
$x1 \times x1 \times x1 \times x1$	<.01	<.01	0.94	<.01	<.01	0.71	<.01	<.01	0.91	<.01	<.01	0.48
$x1 \times x1 \times x1 \times x2$	<.01	<.01	0.82	<.01	<.01	0.55	<.01	<.01	0.52	<.01	<.01	0.92
$x1 \times x1 \times x1 \times x3$	<.01	<.01	0.99	<.01	<.01	0.74	<.01	<.01	0.46	<.01	<.01	0.40
$x1 \times x1 \times x1 \times x4$	<.01	<.01	0.80	<.01	<.01	0.78	<.01	<.01	0.83	<.01	<.01	0.88
$x1 \times x1 \times x2 \times x2$	<.01	<.01	0.25	<.01	<.01	0.23	<.01	<.01	0.15	<.01	<.01	0.36
$x1 \times x1 \times x2 \times x3$	<.01	<.01	0.70	<.01	<.01	0.66	<.01	<.01	0.43	<.01	<.01	0.22
$x1 \times x1 \times x2 \times x4$	<.01	<.01	0.33	<.01	<.01	0.76	<.01	<.01	0.84	<.01	<.01	0.81
$x1 \times x1 \times x3 \times x3$	<.01	<.01	0.37	<.01	<.01	0.44	<.01	<.01	0.19	<.01	<.01	0.17
$x1 \times x1 \times x3 \times x4$	<.01	<.01	0.27	<.01	<.01	0.16	<.01	<.01	0.36	<.01	<.01	0.19
$x1 \times x1 \times x4 \times x4$	<.01	<.01	0.47	<.01	<.01	0.12	<.01	<.01	0.23	<.01	<.01	0.17
$x1 \times x2 \times x2 \times x2$	<.01	<.01	0.35	<.01	<.01	0.51	<.01	<.01	0.43	<.01	<.01	0.52
$x1 \times x2 \times x2 \times x3$	<.01	<.01	0.96	<.01	<.01	0.92	<.01	<.01	0.97	<.01	<.01	0.80
$x1 \times x2 \times x2 \times x4$	<.01	<.01	0.82	<.01	<.01	0.35	<.01	<.01	0.28	<.01	<.01	0.44
$x1 \times x2 \times x3 \times x3$	<.01	<.01	0.28	<.01	<.01	0.18	<.01	<.01	0.10	<.01	<.01	0.04
$x1 \times x2 \times x3 \times x4$	<.01	<.01	0.39	<.01	<.01	0.16	<.01	<.01	0.27	<.01	<.01	0.12
$x1 \times x2 \times x4 \times x4$	<.01	<.01	0.13	<.01	<.01	0.05	<.01	<.01	0.07	<.01	<.01	0.06
$x1 \times x3 \times x3 \times x3$	<.01	<.01	0.76	<.01	<.01	0.90	<.01	<.01	0.99	<.01	<.01	0.98
$x1 \times x3 \times x3 \times x4$	<.01	<.01	0.96	<.01	<.01	0.99	<.01	<.01	0.66	<.01	<.01	0.80
$x1 \times x3 \times x4 \times x4$	<.01	<.01	0.82	<.01	<.01	0.94	<.01	<.01	0.74	<.01	<.01	0.88
$x1 \times x4 \times x4 \times x4$	<.01	<.01	0.75	<.01	<.01	0.53	<.01	<.01	0.52	<.01	<.01	0.59
$x2 \times x2 \times x2 \times x2$	<.01	<.01	0.52	<.01	<.01	0.57	<.01	<.01	0.56	<.01	<.01	0.50
$x2 \times x2 \times x2 \times x3$	<.01	<.01	0.76	<.01	<.01	0.60	<.01	<.01	0.64	<.01	<.01	0.49
$x2 \times x2 \times x2 \times x4$	<.01	<.01	0.71	<.01	<.01	0.68	<.01	<.01	0.63	<.01	<.01	0.69
$x2 \times x2 \times x3 \times x3$	<.01	<.01	0.47	<.01	<.01	0.32	<.01	<.01	0.30	<.01	<.01	0.44
$x2 \times x2 \times x3 \times x4$	<.01	<.01	0.57	<.01	<.01	0.26	<.01	<.01	0.18	<.01	<.01	0.29
$x2 \times x2 \times x4 \times x4$	<.01	<.01	0.41	<.01	<.01	0.43	<.01	<.01	0.37	<.01	<.01	0.36
$x2 \times x3 \times x3 \times x3$	<.01	<.01	0.64	<.01	<.01	0.80	<.01	<.01	0.89	<.01	<.01	0.31
$x2 \times x3 \times x3 \times x4$	<.01	<.01	0.28	<.01	<.01	0.48	<.01	<.01	0.76	<.01	<.01	0.36
$x2 \times x3 \times x4 \times x4$	<.01	<.01	0.55	<.01	<.01	0.77	<.01	<.01	0.74	<.01	<.01	0.77
$x2 \times x4 \times x4 \times x4$	<.01	<.01	0.68	<.01	<.01	0.95	<.01	<.01	0.86	<.01	<.01	0.91
$x3 \times x3 \times x3 \times x3$	<.01	<.01	0.70	<.01	<.01	0.99	<.01	<.01	0.95	<.01	<.01	0.64
$x3 \times x3 \times x3 \times x4$	<.01	<.01	0.44	<.01	<.01	0.83	<.01	<.01	0.96	<.01	<.01	0.83
$x3 \times x3 \times x4 \times x4$	<.01	<.01	0.79	<.01	<.01	0.96	<.01	<.01	0.77	<.01	<.01	0.72
$x3 \times x4 \times x4 \times x4$	<.01	<.01	0.75	<.01	<.01	0.76	<.01	<.01	0.48	<.01	<.01	0.54
$x4 \times x4 \times x4 \times x4$	<.01	<.01	0.30	<.01	<.01	0.33	<.01	<.01	0.18	<.01	<.01	0.23

Table 5.9: Full parameter estimates, standard errors and p-values for Model G in Section 5.4.4, for the DE model fit to the emulated ESM data.

Appendix 5.F Additional Results ESM Time Series

In this appendix, we provide additional figures to visualize the results of the statistical models fit to the ESM time series in Section 5.4.

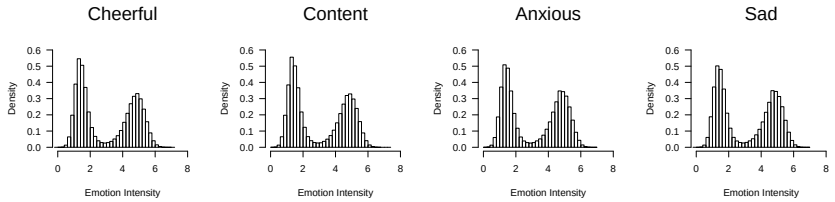


Figure 5.20: The histograms of the emotion intensity of the four modeled emotions Cheerful, Content, Anxious and Sad, for the ESM data

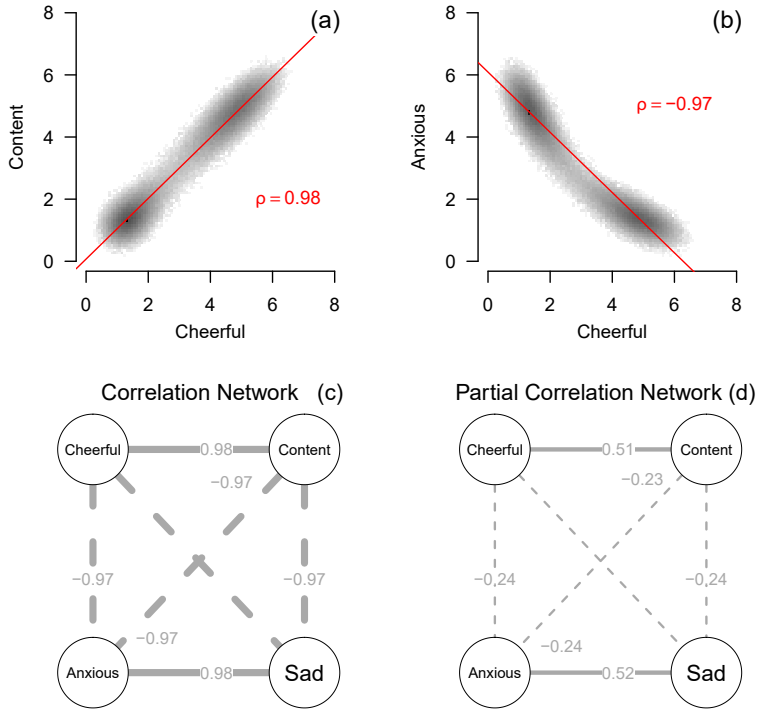


Figure 5.21: Panel (a) shows the relationship between Content and Cheerful, two emotions with the same valence, at the same time point The red line indicates the best fitting regression model, for ESM time series. Similarly, panel (b) shows the relationship between Anxious and Content, two emotions with different valence. Panel (c) displays the correlation matrix as a network, and panel (d) displays the partial correlation matrix as a network.

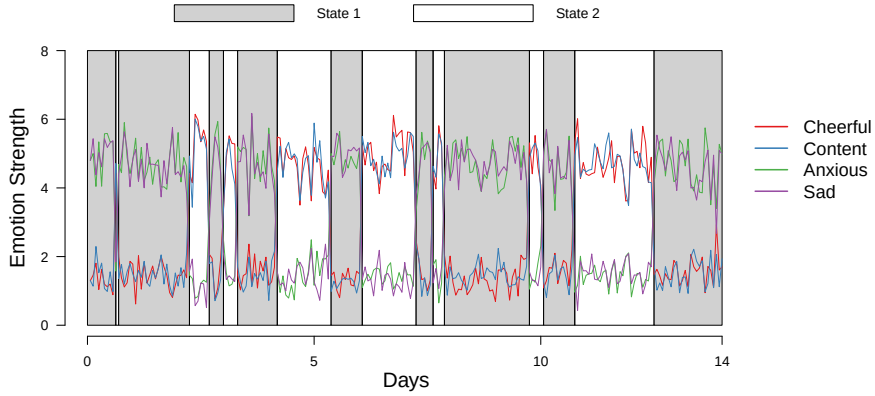


Figure 5.22: Time series of the four emotion variables, also shown in panel (a) of Figure 5.2, with background color indicating whether a given time point is assigned to the first or second component of the mean-switching HMM estimated from the ESM dataset.

Chapter 6

Modeling Psychopathology: From Data Models to Formal Theories

Abstract

Over the past decade there has been a surge of empirical research investigating mental disorders as complex systems. In this paper, we investigate how to best make use of this growing body of empirical research and move the field toward its fundamental aims of explaining, predicting, and controlling psychopathology. We first review the contemporary philosophy of science literature on scientific theories and argue that fully achieving the aims of explanation, prediction, and control requires that we construct formal theories of mental disorders: theories expressed in the language of mathematics or a computational programming language. We then investigate three routes by which one can use empirical findings (i.e. data models) to construct formal theories: (a) using data models themselves as formal theories, (b) using data models to infer formal theories, and (c) comparing empirical data models to theory-implied data models in order to evaluate and refine an existing formal theory. We argue that the third approach is the most promising path forward and conclude by expanding on this approach, proposing a framework for theory construction that details how to best use empirical research to generate, develop, and test formal theories of mental disorders.

This chapter has been adapted from: Haslbeck, J. M. B.*, Ryan, O.*, Robinaugh, D.J.*, Waldorp, L.J. and Borsboom, D. (under review). Modeling Psychopathology: From Data Models to Formal Theories. Pre-print: https://psyarxiv.com/jgm7f/. Author contributions: JMBH, OR and DJR are considered joint first authors and contributed equally to this project. LJW and DB helped develop the ideas in the project, discussed progress and provided textual feedback.

6.1 Introduction

Mental disorders are complex phenomena: highly heterogeneous and massively multifactorial (e.g., Kendler, 2019). Confronted with this complex etiological and ontological picture, researchers have increasingly called for approaches to psychiatric research that embrace this complexity (Gardner & Kleinman, 2019). The "network approach" to psychopathology addresses these calls, conceptualizing mental disorders as complex systems of interacting symptoms (e.g., Borsboom & Cramer, 2013; Schmittmann et al., 2013; Borsboom, 2017). From this perspective, symptoms are not caused by an underlying disorder, rather the symptoms themselves and the causal relations among them constitute the disorder.

In recent years, empirical research within the network approach literature has rapidly grown (for reviews see e.g., Robinaugh, Hoekstra, et al., 2019; Contreras et al., 2019). Most of this work employs statistical models that allow researchers to study the multivariate dependencies among symptoms, thereby providing rich information about the relationships among those symptoms. However, this quickly expanding empirical literature has raised a critical question: how can we best make use of this growing number of empirical findings to advance the fundamental aims of psychiatric science? This problem is not unique to the network approach. Psychiatry has produced countless empirical findings, yet genuine progress in our efforts to explain, predict, and control mental disorders has remained stubbornly out of reach.

In this paper, we will argue that empirical research can best advance these aims by supporting the development of scientific theories. We will begin in Section 6.2 by discussing the nature of scientific theories and how they achieve the explanation, prediction and control sought by psychiatric science. We will argue that to fully achieve these aims, psychiatry requires theories formalized as mathematical or computational models. In Section 6.3, we will explore how models estimated from data can best be used to develop formal theories. We examine three possible routes from data model to formal theory: first, treating data models themselves as formal theories; second, drawing inferences from data models to generate a formal theory; and third, using data models to develop formal theories with an abductive approach. We will argue that the third approach is the most promising path forward. In Section 6.4, we will expand on this approach and propose a framework for theory construction, detailing how best to use empirical research to advance the generation, development, and testing of scientific theories of mental disorders.

6.2 The Nature and Importance of Formal Theories

In this section we will examine the nature of scientific theories and how they support explanation, prediction, and control. We will begin by introducing four key concepts that we will use throughout the remainder of the paper: theory, target system, data, and data models. We will illustrate each of these concepts using the example of panic disorder.

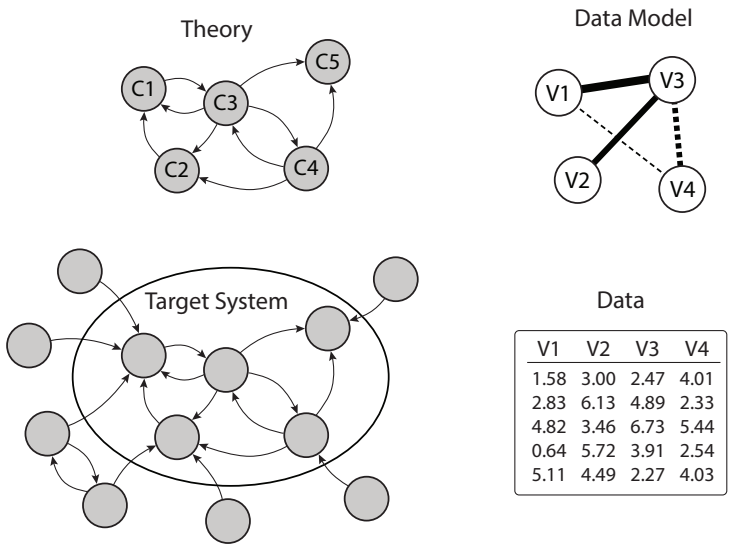


Figure 6.1: The figure illustrates the concepts target system, theory, data and data model. The target system is the system consisting of interacting components that gives rise to phenomena. Phenomena are robust features of the world captured by data models. Theories represent the structure of the target system, proposing a set of components C and the relations among them and positing that they give rise to the phenomena. Data for variables V are obtained by probing the target system.

6.2.1 Theories and Target Systems

Theories seek to explain phenomena: stable, recurrent, and general features of the world (Bogen & Woodward, 1988; Haig, 2008, 2014) such as the melting point of lead, the orbit of planets, and the tendency for some individuals to experience recurrent panic attacks. Well developed theories can predict these phenomena and show how they can be controlled. Although the precise nature of theories remains a subject of ongoing debate among philosophers of science, the past half century has seen a growing consensus that theories are best understood as *models*.[1] Specifically, theories are models that aim to represent *target systems*: the particular parts of the real world that give rise to the phenomena of interest. We use the word "system" here because we assume that the part of the real world giving rise to any psychiatric phenomena can be partitioned into components and the relations among them. We use the term "target", because it is this system that a theory aspires to represent (cf. Elliott-Graves, 2014).

In psychiatry, the most common phenomena to be explained are symptoms

[1] The precise relationship between theories and models is muddied by inconsistent and often conflicting use of these terms across time, disciplines, and scientists (for a brief history of models and their relation to theory, see Bailer-Jones, 2009). In this paper, we will adopt the perspective that theories are models (Suárez & Pero, 2019). However, the core arguments presented in this paper do not require this precise conceptualization of theories and would similarly hold for pragmatic accounts that regard models as an intermediary between theory and the real world (e.g., Bailer-Jones, 2009; Cartwright, 1983).

and syndromes. For example, researchers seek to explain the tendency for some individuals to experience panic attacks and the tendency for recurrent panic attacks to be accompanied by persistent worry about those attacks and avoidance of situations in which they may occur (Spitzer, Kroenke, & Williams, 1980). The target system in psychiatric research comprises the components of the real world that give rise to these symptoms and syndromes, and may include genetic, neurobiological, physiological, emotional, cognitive, behavioral or social components. Psychiatric theories aim to represent these target systems, positing a specific set of components and relationships among them that give rise to the phenomena of interest. For example, researchers have generated numerous theories of panic disorder, specifying a set of components that they believe interact to give rise to panic attacks and panic disorder. Among these, perhaps the most influential is Clark's cognitive model of panic attacks, which posits that "if [stimuli] are perceived as a threat, a state of mild apprehension results. This state is accompanied by a wide range of body sensations. If these anxiety-produced sensations are interpreted in a catastrophic fashion, a further increase in apprehension occurs. This produces a further increase in body sensations and so on round in a vicious circle which culminates in a panic attack" (Clark, 1986). This cognitive theory of panic attacks specifies components (e.g., bodily sensations and a state of apprehension) and the relations among them (e.g., the "vicious cycle" of positive causal effects), positing that this is the target system that gives rise to panic attacks.

Because theories represent the target system, we can reason from theory in order to draw conclusions about the target system. It is this capacity for *surrogative reasoning* (Swoyer, 1991) that allows theories to explain, predict, and control. For example, we can explain the rise and fall of predator and prey populations in the real world by appealing to the relationships between components specified in mathematical models representing these populations (H. I. Freedman, 1980; Nguyen & Frigg, 2017). We can predict what will occur when two atoms collide by deriving the expected outcome from models of particle physics (Higgs, 1964). We can determine how to intervene to prevent panic attacks by appealing to the relationships posited in the cognitive model of panic attacks, determining that an intervention modifying a patient's "catastrophic misinterpretations" should prevent the "vicious cycle" between arousal and perceived threat, thereby circumventing panic attacks (Clark, 1986). It is this ability to support surrogative reasoning that makes theories such powerful tools.

6.2.2 The Importance of Formal Theories

Surrogative reasoning relies on a theory's structure: its components and the relations among them (Pero, 2015; Suárez & Pero, 2019). This structure can be expressed in a written or spoken language (i.e. *verbal theory*) or in the language of mathematics or computation (i.e. *formal theory*). For example, a verbal theory would state that the rate of change in an object's temperature is proportional to the difference between its temperature and the temperature of its environment. A formal theory would instead express this relationship as a mathematical equation, such as $\frac{dT}{dt} = -k(T - E)$, where $\frac{dT}{dt}$ is the rate of

change in temperature, T is the object's temperature, and E is the temperature of the environment; or in a computational programming language, such as: `for(t in 1:end) { T[t+1] = T[t]-k*(T[t]-E) }`.

Expressing a theory in a mathematical or computational programming language gives formal theories many advantages over verbal theories (e.g., Smith & Conrey, 2007; Epstein, 2008; Lewandowsky & Farrell, 2010; Smaldino, 2017). There is one advantage especially relevant to the current paper: Formalization enables precise deduction of the behavior implied by the theory. Verbal theories can, of course, also be used to deduce theory-implied behavior. However, due to the vagaries of language, verbal theories are typically imprecise, thereby precluding their ability to make exact predictions. For example, the verbal theory of temperature cooling described in the previous paragraph allows for some general sense of how the object's temperature will evolve over time, but cannot be used to make precise predictions about how it will change or where temperature will be at any given point in time. Indeed, because of the imprecision of verbal theories, there are often multiple ways in which those theories could be interpreted and implemented, each with a potentially divergent prediction about how the target system will evolve over time. Consider the interpersonal theory of suicide, which posits that suicide arises from the simultaneous experience of perceived burdensomeness and thwarted belongingness (Van Orden et al., 2010). This theory fails to specify many aspects of this causal structure, such as the strength of these effects or the duration for which they must overlap before suicidal behavior arises (Hjelmeland & Loa Knizek, 2018). As a result, there are many possible implementations of that verbal theory, each of which could potentially lead to a different prediction about when suicidal behavior should be expected to arise. This imprecision thus substantially limits the theories ability to support surrogative reasoning and the degree to which we can empirically test the theory.

In contrast to most verbal theories, formal theories are precise in their implementation as the mathematical notation or code in a computer programming language forces one to be specific about the structure of the theory (e.g., specifying the precise effect of one component on another). The precision of formal theories allows for the provision of singular and precise predictions about how the target system will behave. These predictions can either be obtained analytically from the mathematical equation or computed by implementing the formula in a programming language. For example, we can use the formal theory of cooling to predict the exact temperature of our object at any given point in time. Similarly, a formal implementation of the interpersonal theory of suicide would make precise predictions that could inform the prediction of suicide attempts. In other words, formal theories substantially strengthen surrogative reasoning, the very characteristic of scientific theories upon which we wish to capitalize.[2]

[2]It is, of course, possible to express verbal theories with the same level of precision as is provided by a mathematical equation (e.g., there are very few equations in the Principia, yet the laws Newton describes are not lacking in precision). Nonetheless, the specificity required by mathematics or computational programming makes them more amenable to expressing theories precisely and has the considerable practical advantage of supporting the derivation of predictions from the theory.

6.2.2.1 A Formal Theory of Panic Disorder

The cognitive model of panic attacks posited by Clark is a verbal theory and is limited by the imprecision characteristic of most verbal theories. Indeed, in two recent papers, Fukano and Gunji (Fukano & Gunji, 2012) and Robinaugh and colleagues (Robinaugh, Haslbeck, et al., 2019) independently proposed two distinct formal implementations of this theory, taking the verbal theory and expressing it in differential equations. Notably, these distinct implementations of the same verbal theory make divergent predictions about when panic attacks should occur, illustrating the limitations of failing to precisely specify the theory (for further detail, see Robinaugh, Haslbeck, et al., 2019).

In this paper, we will make extensive use of the formal theory proposed by Robinaugh and colleagues. A complete description of the generation of this theory can be found in Robinaugh, Haslbeck, et al. (2019). For the purposes of this paper, it is sufficient to note that the aim in developing this model was to take extant verbal theories, especially cognitive behavioral theories, and express them in the language of mathematics. For example, Clark's verbal theory posits that a perception of threat can lead to arousal-related bodily sensations. However, the actual form and strength of this effect remain unspecified. In our mathematical model, we used a differential equation to precisely define this effect: $\frac{dA}{dt} = \alpha(\nu T - A)$. In this equation, there is a linear effect of Perceived Threat (T) on the rate of change of Arousal (A), with the strength of this effect specified by the parameter ν. The product of ν and T is the value Arousal is pulled toward: if νT is smaller than the current level of Arousal, $\frac{dA}{dt}$ will be negative and Arousal will *decrease* toward νT; if νT is greater than Arousal, $\frac{dA}{dt}$ is positive and Arousal *increases* toward νT. Each model component was defined as a differential equation in this way (see middle panel in Figure 6.2).

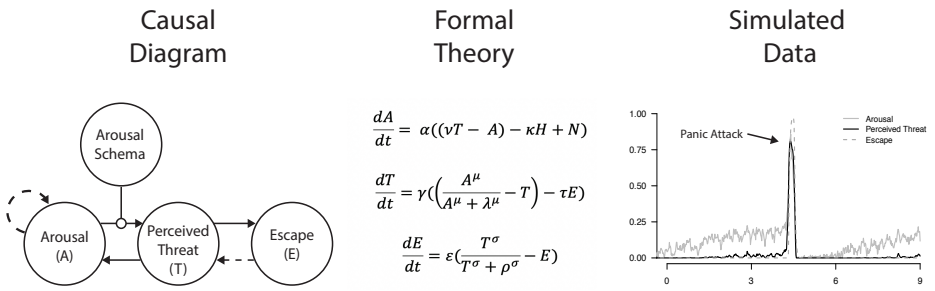


Figure 6.2: The left panel displays the key components of the theory proposed by Robinaugh, Haslbeck, et al. (2019) at play during panic attacks: Arousal, Perceived Threat, Escape Behavior and arousal schema. The arrows indicate the direct causal relationships which are posited to operate between these components in the formal theory. The middle panel displays the formal theory that specifies the precise nature of the relations among these components. The right panel depicts the simulated behavior implied by the theory.

By specifying the structure of the theory in this way, we are able to solve the system numerically, thereby deducing the theory's predictions about how the

target system will behave. For example, the theory shows that when the effect of Arousal on Perceived Threat is sufficiently strong, the positive feedback between these components is sufficient to send the system into runaway positive feedback, producing the characteristic surge of arousal, perceived threat, and escape behavior that we refer to as a panic attack (see right panel in Figure 6.2). As this example illustrates, specifying the theory as a computational model substantially strengthens our ability to deduce the behavior implied by the theory. A full realization of a theory's usefulness thus all but requires that theory be formalized. For that reason, we believe the ultimate goal of psychiatric research should not only be the production of theories, but the production of formal theories.

6.2.3 Data and Data Models

Our brief overview of the philosophy of science literature on theories suggests that if our aim is the explanation, prediction, and control of mental disorders, what we are after are well-developed formal theories: mathematical or computational models that represent the target system. The key question then becomes: how can we best determine such a formal theory?

The answer to this question will, of course, involve the collection and analysis of *data*. Empirical data plays at least two key roles in the development of formal theories. First, data gathered about the target system are key to establishing what our theories must explain. Yet, theories typically do not aim to explain data directly. Data are sensitive to the context in which they are acquired and subject to myriad causal influences that are not of core interest (Woodward, 2011). For example, panic disorder researchers collect data from diagnostic interviews, self-report symptom inventories, assessments of physiological arousal during panic attacks, time-series data, and a host of other methods. Data about panic attacks gathered using these methods will be influenced not only by the experienced attacks, but also by recall biases, response biases, sensor errors, and simple human error. Accordingly, theories do not aim to account for specific "raw" data. Rather, theories explain phenomena identified through robust patterns in the data that cannot be attributed to the particular manner in which the data were collected (e.g., researcher biases, measurement error, methodological artifacts, etc.). To identify these empirical regularities in data, researchers use *data models*, which are representations of the data (Suppes, 1962; Kellen, 2019). Data models can take many forms. These can range from the most basic canonical descriptive tools, such as a mean score, a correlation, or a fitted curve, to more complex statistical tools which are common in different areas of psychology and beyond; such as structural equation models, time-series models, hierarchical models, network models, mixture models, loglinear models and so forth. Essentially, we can consider a data model to be any (statistical) model that summarizes the data in some way. Thus, data models, particularly robust and replicable data models, play a key role in determining what a theory must explain.

Second, data models also inform our understanding of the components and the relations among them that are posited to give rise to a phenomenon (i.e. the theory's structure). It is this role which we will focus on in the next part of the

paper. This role is especially noteworthy in the context of the data models most commonly used in the network approach literature: the Ising model, the Gaussian Graphical Model, and the Vector Autoregressive model. In Section 6.3 we will describe each of these models in more detail, but here it is sufficient to note their most salient feature: these analyses estimate the structure of relationships among a set of variables; specifically, the structure of conditional dependence relationships (see Figure 6.1; Top Right). There is a strong intuitive appeal to these analyses as they seem to hold the promise of directly informing the very thing we are after: the structure of relations among components of the mental disorder (see Figure 6.1; Top Left). In Section 6.3, our overarching aim will be to critically evaluate that promise and determine how best to use (network) data models to guide the development of theories about specific mental disorders.

6.3 Identifying Formal Theories from Data

In this section we will explore how data models can best contribute to the development of formal theories. We will do so within the broader theoretical framework of conceptualizing mental disorders as complex systems and will focus on three data models that have become popular among researchers adopting this framework: the Ising model, the Gaussian Graphical Model (GGM), and the Vector Autoregressive (VAR) model. Specifically, we evaluate three routes that make use of data models in different ways to obtain a formal theory. We believe that the first two routes describe how data models are currently used in the literature, and the third route is an alternative that addresses some of the shortcomings of the first two approaches.

The first route arrives at formal theories directly by treating these data models as formal theories. In this case, the transition from data model to formal theory is largely an act of interpretation. Instead of interpreting a data model as a representation of the data, we interpret it as a representation of the target system (see Figure 6.3, Left Panel). Specifically, the variables of the data model are treated as the components of the target system, and the statistical relationships are treated as the structural relationships among the components. From this perspective, research is carried out by conducting an empirical study, estimating a data model, and treating the data model as a theory. If viable, this approach would be extremely powerful, because a well-developed theory would be just one well-designed study away. We evaluate this route in Section 6.3.1.

The second possible route arrives at formal theories by drawing inferences from data models (Figure 6.3, Middle Panel). That is, the data model is not directly treated as a theory, but rather is used to inform the theory. From this perspective, research is carried out by conducting an empirical study, estimating a data model, and using the data model to infer characteristics of the target system, thereby informing the development of a theory. For example, one could observe a conditional dependence relationship between two variables and infer the presence of a causal relationship between the corresponding components in the target system. To evaluate this approach we need to know the data-generating

target system. We do so in Section 6.3.2 by treating the Panic Model introduced in the previous section as the target system of interest, simulating data from that target system, and examining how well inferences drawn from data models can be used to inform our understanding of the target system.

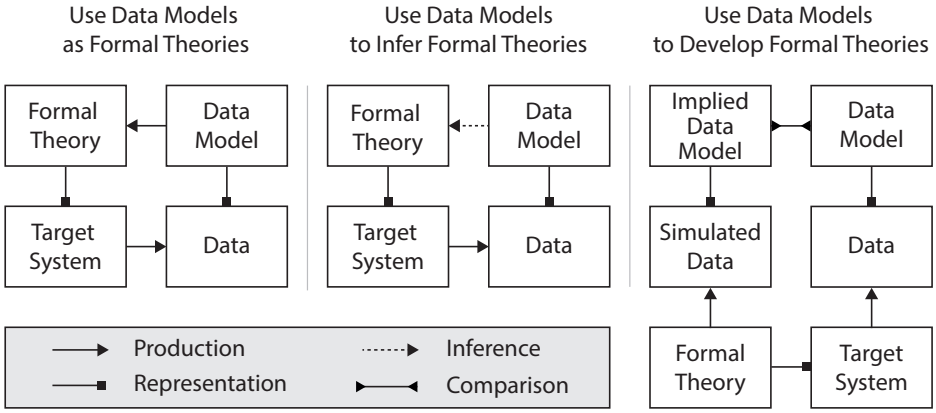


Figure 6.3: The figure provides an overview of three routes to developing formal theories using data models. In the left panel, data models are treated as formal theories. In the middle panel, data models are used to draw inferences about the target system and, thereby, to generate formal theories of that system. In the right panel, data models used to develop formal theories by deducing implied data models and comparing them with empirical data models.

The third possible route puts formal theories at the heart of theory development. From this perspective, research is carried out by first generating an initial formal theory. From this formal theory we simulate data which we use to obtain the theory-implied data model. We subsequently compare the implied data model with the empirical data model, and adapt the formal theory based on the discrepancy between the two. This route thus leverages the "immense deductive fertility" of formal theories to make precise predictions that clarify how the model must be revised to be brought in line with empirical data (Meehl, 1978). From this perspective, formal theory is not only the ultimate goal of the research process, but also plays an active role in theory development. We evaluate this route in Section 6.3.3 by deriving predicted data models from a formal theory of Panic disorder, and showing how the model can be improved by comparing the predicted data models to empirical data models.

6.3.1 Using Data Models as Formal Theories

If data models are to serve as formal theories of a target system, the properties of those data models must be able to represent the properties we expect in the target system. Accordingly, in this section, we discuss the properties we expect in the target systems of mental disorders from the complex systems perspective (Section 6.3.1.1) and evaluate whether these properties are captured by the properties

of three data models: the VAR model, the GGM, and the Ising model (Section 6.3.1.2).

6.3.1.1 Properties of Mental Disorder Target Systems

Target systems consist of components and the relations among them. From the network perspective there are a number of properties we would expect to be present in the target systems of mental disorders. First, feedback loops among components are likely present. Researchers have frequently posited "vicious cycles", where the initial activation of one component (e.g., arousal) elicits activation of other components (e.g., perceived threat) and, in turn, is reinforced by the activation of those components. Second, causal effects between components are likely to be asymmetrical. That is, the effect of component A on component B may differ from the effect of component B on component A. For example, it is unlikely that concentration has the same effect on sleep as sleep has on concentration or that compulsions have the same effect on obsessions that obsessions have on compulsions.

Third, interactions among components are likely to occur at different time scales. For example, the effect of intrusive memories on physiological reactivity in Post-traumatic Stress Disorder is likely to occur on a time scale of seconds to minutes, whereas an effect of energy on depressed mood may play out over the course of hours to days, and the effect of appetite on weight gain may occur on a time scale of days to weeks. Fourth, it is likely that there are higher order interactions among components. For example, the presence of sleep difficulties may strengthen the effect of feelings of worthlessness on depressed mood or the effect of intrusive trauma memories on physiological reactivity. If data models are to serve as formal theories of the target system, they must be able to represent these types of causal structures.

We would further suggest that most, perhaps all, mental disorder target systems are likely to have multiple stable states, that is, states into which the system settles and will remain in the absence of external perturbation. In the simplest case, the system will be characterized by the presence of two stables states: an unhealthy state (i.e. a state of elevated symptom activation, such as a depressive episode), and a healthy state (e.g., a state without elevated symptom activation). In other cases, there may be multiple stable states (e.g., healthy, depressed, and manic states in Bipolar Disorder). The presence of multiple stable states is, in turn, accompanied by other behavior often observed in mental disorders, including spontaneous recovery and sudden shifts into or out of a state of psychopathology, further suggesting that a model of any given mental disorder will almost certainly need to able to produce alternative stable states.

6.3.1.2 Comparing Target System Properties with Data Model Properties

The first model we will consider is the VAR model. The VAR model for multivariate continuous time series data linearly relates each variable at time point t to all other variables and itself at previous time points (Hamilton, 1994), typically the time point immediately prior $t-1$ (i.e. a first order VAR, or VAR(1), model; e.g.,

Bringmann et al., 2013; Pe et al., 2015; A. J. Fisher et al., 2017; Snippe et al., 2017; Groen et al., 2019). The estimated lagged effects of the VAR models indicate conditional dependence relationships among variables over time. The dynamic of the VAR model is such that the variables are perturbed by random input (typically Gaussian noise) and the variables return to their means, which represent the single stable state of the system.

As depicted in Figure 6.4, the VAR model is able to represent some key characteristics likely to be present in mental disorder target systems. Most notably, it allows for feedback loops. Variables can affect themselves both directly (e.g., $X_t \rightarrow X_{t+1}$), or via their effects on other variables in the system (e.g., $X_t \rightarrow Y_{t+1} \rightarrow X_{t+2}$). The VAR model also allows for asymmetric relationships, since the effect $X_t \rightarrow Y_{t+1}$ does not have to be the same effect as $Y_t \rightarrow X_{t+1}$ in direction or magnitude. However, because the lag-size (i.e. the distance between time points) is fixed and consistent across all relationships, the VAR model does not allow for different time scales. Moreover, because the VAR model only includes relations between pairs of variables, it is unable to represent higher-order interactions involving more than two variables. Finally, the VAR model has a single stable state defined by its mean vector and thus cannot represent multiple stable states of a system, such as a healthy state and unhealthy state.

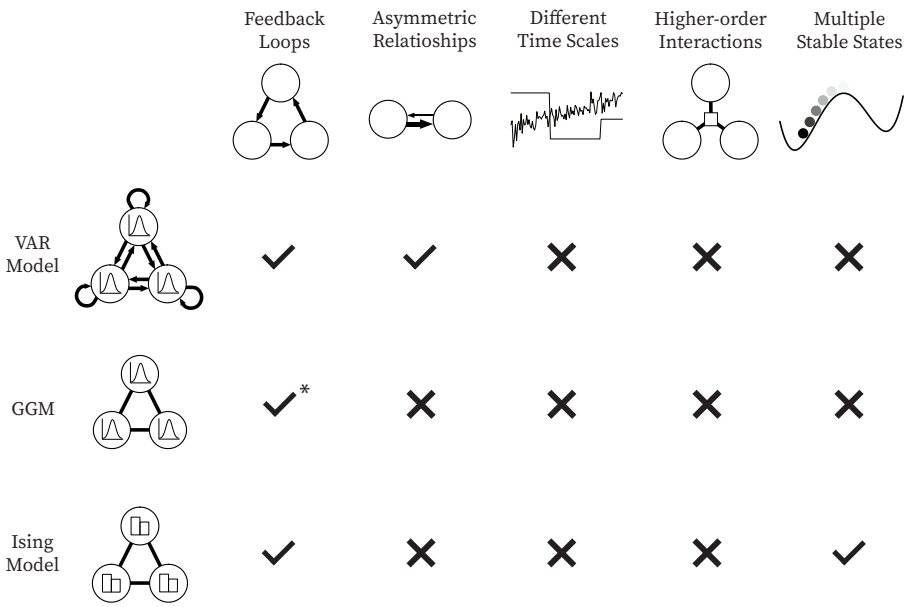


Figure 6.4: The figure shows whether the five properties of mental disorders discussed above can be represented by the three most popular network data models, the VAR model, the GGM, and the Ising model with Glauber dynamics. Note that there is a check mark at feedback loops for GGMs because one could in principle endow the GGM with a dynamic similar to the Ising model, which would essentially lead to a restricted VAR model but with symmetric relations. The asterisk is present because this endowment of dynamics is not done in practice.

The second model we will consider is the Gaussian Graphical Model (GGM). The GGM linearly relates pairs of variables in either cross-sectional (Haslbeck & Fried, 2017) or time series data (Epskamp, Waldorp, et al., 2018). In the case of time series data the GGM models the relationships between variables at the same time point. Because it does not model any dependency across time, it is typically not considered a dynamic model and, thus, could not be used to represent the behavior of a mental disorder target system as it evolves over time. In principle the GGM could be augmented by a dynamic rule similar to one commonly used with the Ising model (i.e. "Glauber dynamics", see below). However, in that case, the GGM would become a model similar to, but more limited than, the VAR model described above (e.g., it would be limited to symmetric relationships). Accordingly, the GGM is similarly unable to represent key features we expect to observe in a mental disorder target system.

The final model we will consider is the Ising model. The Ising model again represents pairwise conditional dependence relations between variables (Ising, 1925), however, it is a model for multivariate binary data. While the original Ising model does not model dependencies over time, it can be turned into a dynamic model by augmenting it with Glauber dynamics (Glauber, 1963).[3] Like the VAR model, the Ising model is able to represent feedback loops. Moreover, due to its non-linear form it is able to exhibit multiple stable states (and the behavior that accompanies such stable states, such as hysteresis and sudden shifts in levels of symptom activation, see e.g., Cramer et al., 2016; Lunansky, van Borkulo, & Borsboom, 2019; Dalege et al., 2016). It is perhaps not surprising then, that the Ising model is used as a theoretical model across many sciences (Stutz & Williams, 1999), and to our knowledge, is the only of the three data models examined here that has been used as a formal theory of a mental disorder target system (Cramer et al., 2016). Unfortunately, the Ising model falls short in its ability to represent the remaining characteristics likely to be present in mental disorders. The relationships in the Ising model are exclusively symmetric; with the standard Glauber dynamics, there is only a single time scale; and the Ising model includes exclusively pairwise relationships, precluding any representation of higher-order interactions.

6.3.1.3 Data Models as Formal Theories?

The analysis in this section shows that the VAR, GGM, and Ising models are unable to represent most key properties we would expect in the target systems giving rise to mental disorders, and therefore cannot serve as formal theories for those disorders. Of course, more complex models would be able to produce more of the characteristics likely to be present in mental disorders. For example, one could extend the VAR model with higher-order interactions or a latent state (Tong & Lim, 1980; Hamaker et al., 2010), thereby allowing it to represent multiple

[3] Glauber dynamics work as follows: After specifying an initial value for each variable, one randomly picks a variable X_i at $t = 1$ and takes a draw from the distribution of X_i conditioned on the values of all other variables. This value (either 0 or 1) is set to be the new value of X_i and then the same process is repeated, thereby allowing the model to evolve over time.

stable states. However, estimating data models is subject to fundamental constraints. More complex models require more data, and larger sample sizes which are often unavailable in psychiatric research. For example, around 90 observations (about 2.5 weeks of a typical ESM study) are needed for a VAR model to outperform the much simpler AR model (Dablander et al., 2019). Models more complex than the VAR model would require even more data to be estimated reliably. In addition, the sampling frequency (e.g., measurement every 2 hours) might be too low to capture the structure of the target system of interest (Haslbeck & Ryan, 2019). In this situation a data model still contains some information about the target system, but cannot capture the structure of the target system to the extent that it can serve as a formal theory. Even if large amounts of high frequency data were widely available, it is unclear how to estimate many complex models. For example, one could extend the Ising model with a second time scale (e.g., Lunansky et al., 2019), but it would be unclear how to estimate such a model from data. Finally, even if such models could be estimated, more complex models are often uninterpretable. For example, nonparametric models (e.g., splines; Friedman, Hastie, & Tibshirani, 2001, p. 139), which can capture extremely complex behavior, typically consist of thousands of parameters, none of which can be interpreted individually. Accordingly, it is unlikely that any data model estimated from the type of data typically available in psychiatric research will be both interpretable and capable of capturing the characteristics of psychopathology in such a way that would allow it to serve as a formal theory of a mental disorder.

6.3.2 Using Data Models to Infer Formal Theories

An alternative route from data models to formal theories is to use data models to draw inferences about a target system, inferences that we can use to construct a formal theory. There is good reason to think that this approach could work. Because the data are generated by the target system, and data models summarize these data, the parameters of any data model certainly *somehow* reflect characteristics of the target system. This means that it should be possible, in principle, to infer something about the target system and its characteristics from data and data models. Although we have seen already that the GGM, Ising and VAR models cannot directly reproduce the key characteristics of the target system, their parameters could potentially still yield insights into the structure or patterns of relationships between components. In line with this intuition, it has frequently been suggested that the GGM, the Ising model, and the VAR models can serve as "hypothesis-generating tools" for the causal structure of the target system (e.g., Borsboom & Cramer, 2013; van Rooijen et al., 2017; Fried & Cramer, 2017; Epskamp, van Borkulo, et al., 2018; Epskamp, Waldorp, et al., 2018; Jones, Mair, Riemann, Mugno, & McNally, 2018).

Although this approach seems intuitive, in practice it is unclear how this inference from data model to target system should work. For example, if we observe a strong negative cross-lagged effect of X_t on Y_{t+1} in a VAR model, what does that imply for the causal relationship between the corresponding components in the target system? A precise answer to this question would require a rule that con-

nects parameters in particular data models to the structure of the target system. For some simple systems, such a rule is available. For example, if the target system can be represented as a Directed Acyclic Graph (DAG), then under certain circumstances its structure can be inferred from conditional (in)dependence relations between its components: Conditional independence implies causal independence, and conditional dependence implies either direct causal dependence or a common effect (Pearl, 2009; Ryan, Bringmann, & Schuurman, 2019). However, it is generally unclear how we can use the parameters of typical data models to make inferences about the types of non-linear dynamic systems we expect in a psychiatric context (although Mooij, Janzing, & Schölkopf, 2013 and Forré & Mooij, 2018 have established some links in this regard). The consequences of this are twofold. First, any inference from data model to target system must rely instead on some simplified heuristic(s) in an attempt to approximate the link between the two. Second, it is unclear how well the combination of common data models and simple heuristics perform in allowing us to make inferences about the target system.

In this section, we evaluate whether the three data models introduced above can be used to make inferences about mental disorder target systems. To do this, we treat the Panic Model discussed in Section 6.2 as the data-generating target system and compare the causal structure inferred from the data models to the true causal structure. To yield these inferences we use a very simple and intuitive set of heuristics: a) if two variables are conditionally dependent in the data model, we will infer that the corresponding components in the target system are directly causally dependent; b) if there is a positive linear relationship, we will infer that the causal relation between the corresponding components is positive (i.e. reinforcing); c) if there is a negative linear relationship, we will infer that the causal relationship among components is negative (i.e. suppressing).

6.3.2.1 Inferring the Panic System from Network Data Models

To be able to evaluate the success of the simple heuristics described above, we must first represent the structure of the Panic Model (see Section 6.2) in the structure of a square matrix, that is, in the same form as the parameters of the VAR, GGM, and Ising models. Since the relationships between components are formalized through *differential equations*, a natural choice is to represent the Panic Model as a network of moment-to-moment dependencies, drawing an arrow $X \rightarrow Y$ if the rate of change of Y is directly dependent on the value of X (known as a *local dependence graph*; Didelez, 2007). Figure 6.5 (a) displays these moment-to-moment dependencies. Note that this structure cannot capture many aspects of the true model, such as the presence of two time scales or the moderating effect of Arousal Schema (AS) (see Section 6.2 for details). It is, thus, already clear that the models cannot recover the *exact* causal structure of the Panic Model. Nonetheless, we can still investigate whether applying the simple heuristics to these three data models allows us to infer this less detailed pattern of direct causal dependencies.

We next compare this true causal structure to the causal structure inferred based on the three data models. To obtain the three data models, we first gen-

erate data from the target system (See Appendix 6.A). Specifically, we use four weeks of minute-to-minute time-series data for 1000 individuals. These individuals differ in their initial value of Arousal Schema, with the distribution chosen so that the proportion of individuals for whom a panic attack is possible was equivalent to the lifetime history prevalence of panic attacks in the general population (R. R. Freedman, Ianni, Ettedgui, & Puthezhath, 1985). For the VAR model analysis, we create a single-subject experience-sampling-type dataset by choosing the individual who experiences the most (16) panic attacks in the four-week period. To emulate ESM measurements, we divide the four week period into 90-minute intervals, taking the average of each component in that interval, yielding 448 measurements. For the GGM analysis, we create a continuous cross-sectional dataset by taking the mean of each component for each individual over the four weeks. For the Ising model analysis, we obtain cross-sectional binary data by taking a median split of those same variables. The resulting VAR, GGM and Ising model networks are displayed in Figure 6.5 panels (b), (c) and (d), respectively.[4]

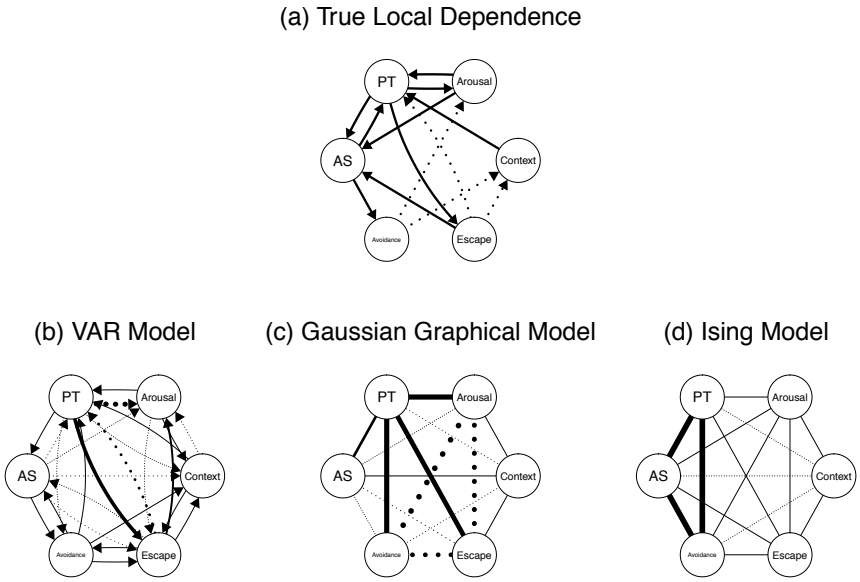


Figure 6.5: Panel (a) shows the true model in terms of local dependencies between components; panel (b) shows the VAR model estimated from ESM data sampled from the true model; panel (c) shows the GGM estimated from the cross-sectional data of 1000 individuals, generated from the true model; panel (d) shows the Ising model estimated on the same data after being binarized with a median split. Solid edges indicate positive relationships, dotted indicate negative relationships. For panels (b) to (d), the widths of edges are proportional to the absolute value of the corresponding parameter. Note that in panel (b) we do not depict the estimated auto-regressive parameters as the primary interest is in inferring relationships between variables.

[4]Note that in the Ising model the parameter estimates are somewhat unstable due to near-deterministic relationships between some binarized variables.

We will focus our evaluation on two important causal dependencies in the target system: the positive (i.e. reinforcing) moment-to-moment feedback loop between Perceived Threat and Arousal, and the positive effect of Arousal Schema (i.e. beliefs that arousal-related bodily sensations are dangerous) on Avoidance (i.e. efforts to avoid situations or stimuli that may elicit panic attacks). In the VAR model (panel (b) in Figure 6.5) we see a lagged positive relationship of Arousal to Perceived Threat, a strong *negative* lagged relationship from Perceived Threat to Arousal, and a weak positive effect of Arousal Schema on Avoidance. Applying the heuristics, we would infer a reinforcing relationship from Arousal to Perceived Threat, a suppressing relationship from Perceived Threat to Arousal, and a reinforcing effect of Arousal Schema on Avoidance. In the GGM (panel (c) in Figure 6.5) we see a positive conditional dependency between mean values of Arousal and Perceived Threat, but we also see a weak negative dependency between mean values of Arousal Schema and Avoidance. Applying the heuristics to the GGM, we would infer a reinforcing relationship between Arousal and Perceived Threat, and a suppressing relationship between Arousal Schema and Avoidance. Finally, in the Ising model (panel (d) in Figure 6.5), we see a strong positive dependency between Arousal Schema and Avoidance, and a very weak positive relationship between PT and Arousal. This leads us to infer two reinforcing relationships, between Arousal and Perceived Threat, and Arousal Schema and Avoidance.

For the VAR model, the heuristics yield one correct and one incorrect inference. For the GGM, we make exactly the opposite inferences, with again one correct and one incorrect. In the Ising Model, we yield two correct inferences. However, inspecting the rest of the Ising Model edges we can see a variety of incorrect inferences about other relationships, with independent components in the target system connected by strong edges in the Ising model, and the valences of various true dependencies flipped. At best, we can say that in each of the three network models, some dependencies do reflect the presence and/or direction of direct causal relationships, and some do not. Unfortunately, it is not possible to distinguish which inferences are trustworthy and which are not without knowing the target system, and in any real research context, the target system will be unknown. Consequently, these data models cannot be used to confidently and reliably draw inferences about the target system using these simple heuristics.

6.3.2.2 The Mapping between Data Model and Target System

Importantly, our inability to draw accurate inferences from these data models is not a shortcoming of the data models themselves. Each data model correctly captures some form of statistical dependency between the components in a particular domain (e.g., lagged 90 minute windows). Moreover, the statistical dependencies in the data models are produced by causal dependencies in the target system, so we know there is *some* mapping from the causal dependencies in the target system to statistical dependencies in the data model. The fundamental barrier to inference is that the form of this mapping is unknown and considerably more complex than the simple heuristics we have used to draw inferences here.

For example, consider the relationships between Perceived Threat and Arousal. The VAR model (panel (b) in Figure 6.5), identifies a negative lagged relationship from Perceived Threat to Arousal in the data generated by the target system. Yet in the target system, this effect is positive. This "discrepancy" occurs because of a very specific dynamic between these components: After a panic attack (i.e. a brief surge of Perceived Threat and Arousal) there is a "recovery" period in which arousal dips below its mean level for a period of time. As a result, when we average observations over a 90 minute window, a high average level of Perceived Threat is followed by a low average level of Arousal whenever a panic attack occurs. That same property of the system produces the observed findings for the GGM and Ising Model through yet another mapping (for details, see Appendix 6.B).

As this example illustrates, the mapping between target system and data model is intricate, and it is unlikely that any simple heuristics can be used successfully to work backwards from the data model to the exact relationships in the target system. We can expect this problem to arise whenever we use relatively simple statistical models to directly infer characteristics or properties of a complex system (c.f. the problem of under-determination or indistinguishability; Eberhardt, 2013; Spirtes, 2010). Indeed, the same problem arises even for simpler dynamical systems when analyzed with more advanced statistical methods (e.g., Haslbeck & Ryan, 2019). Of course, in principle, it must be possible to make valid inferences from data and data models to some properties of a target system using a more principled notion of how one maps to the other. For example, under a variety of assumptions, it has been shown that certain conditional dependency relationships can potentially be used to infer patterns of local causal dependencies in certain types of dynamic system (Mooij et al., 2013; Bongers & Mooij, 2018; Forré & Mooij, 2018). However, the applicability of these methods to the type of target system we expect to give rise to psychopathology (see Section 6.3.1) is as yet unclear and even under the strict assumptions under which they have been examined, these methods still do not recover the full structure of the target system.[5] This means that the intricacy of the mapping between target system and data model currently precludes us from making reliable inferences about the target system. Accordingly, we cannot use those inferences to build formal theories.

[5]Specifically, Mooij et al. (2013) and Bongers and Mooij (2018) have shown that cyclic causal models can be conceptualized as encoding causal dependencies between the equilibrium positions of deterministic differential equations and differential equations with random initial values. Forré and Mooij (2018) formally link the conditional dependencies between equilibrium position values to the causal dependencies in these cyclic causal models using a considerably more complex mapping rule than that which holds for DAGs. Their applicability to the current context is limited in the sense that 1) to our knowledge these rules have not been extended to dynamic systems with time-varying stochastic terms (SDEs) as we would expect to see in complex psychological systems (and on which the Panic Model is based), and 2) the use of these methods is reliant on data that reflects equilibrium positions. Future developments in this area may prove to yield useful tools for psychological theory development however, and we consider this area to be ripe for future research beyond the current paper.

6.3.3 Using Data Models to Develop Formal Theories

In Section 6.3.2, we saw that the mapping between target system and data model is intricate and would be nearly impossible to discern when the target system is unknown. However, we also saw that when the target system is known, we can determine exactly which data models the target system will produce. Indeed, this is precisely what we did when we simulated data and fit data models to it in the previous section. In this section we consider a third route to formal theories, which makes use of this ability to determine which data models are implied by a given target system (or formal theory).

This third route works as follows. First and foremost, we must propose *some* initial formal theory which we take as a representation of the target system. The quality or accuracy of this representation may be good or bad, but crucially the theory must be formalized in such a way as to yield unambiguous predictions (see Section 6.2.2). Second, we can use this initial theory to deduce a theory-implied data model. This can be done by simulating data from the formal theory and fitting the data model of interest. Third, we can learn about and adapt the formal theory by comparing implied data models with their empirical counterparts. This approach is represented in schematic form in the right-hand panel of Figure 6.3. It can be seen as a form of inference, but it is *abductive inference*: inference to the best explanation (Haig, 2005). We first infer the best explanation for the core phenomena to generate an initial theory. We then infer the best explanation for any discrepancies between empirical and theory-implied data models, inferences which inform subsequent theory development. Given the importance of abductive inference to this approach, we adopt this term to refer to this third route. In this sub-section, we will illustrate this approach using the example of panic disorder (for an overview, see Figure 6.6).

6.3.3.1 Obtaining Theory-Implied and Empirical Data Models

In this section, we will treat the Panic Model introduced in Section 6.2 as our initial formal theory, which should represent the target system that gives rise to panic disorder (Figure 6.6, bottom row). The Panic Model can be used to simulate data and, in turn, to derive predictions made by the theory in the form of theory-implied data models (left-hand column of Figure 6.6). While in principle many data models can (and should) be used to perform the abductive inference described above, here we will examine the implied cross-sectional Ising model of the three core panic disorder symptoms: 1) Recurrent Panic Attacks (PA), 2) Persistent Concern (PC) following a panic attack and 3) Avoidance (Av) behavior following a panic attack (American Psychiatric Association, 2013). If our formal theory of panic disorder is an accurate representation of the target system that gives rise to panic disorder, the implied Ising model derived from this theory should be in agreement with a corresponding Ising model derived from empirical data.

Critically, obtaining an implied data model requires not only a formal theory from which we can simulate data, but also a formalized process by which variables are "measured" from those data. The Panic Model is used to generate

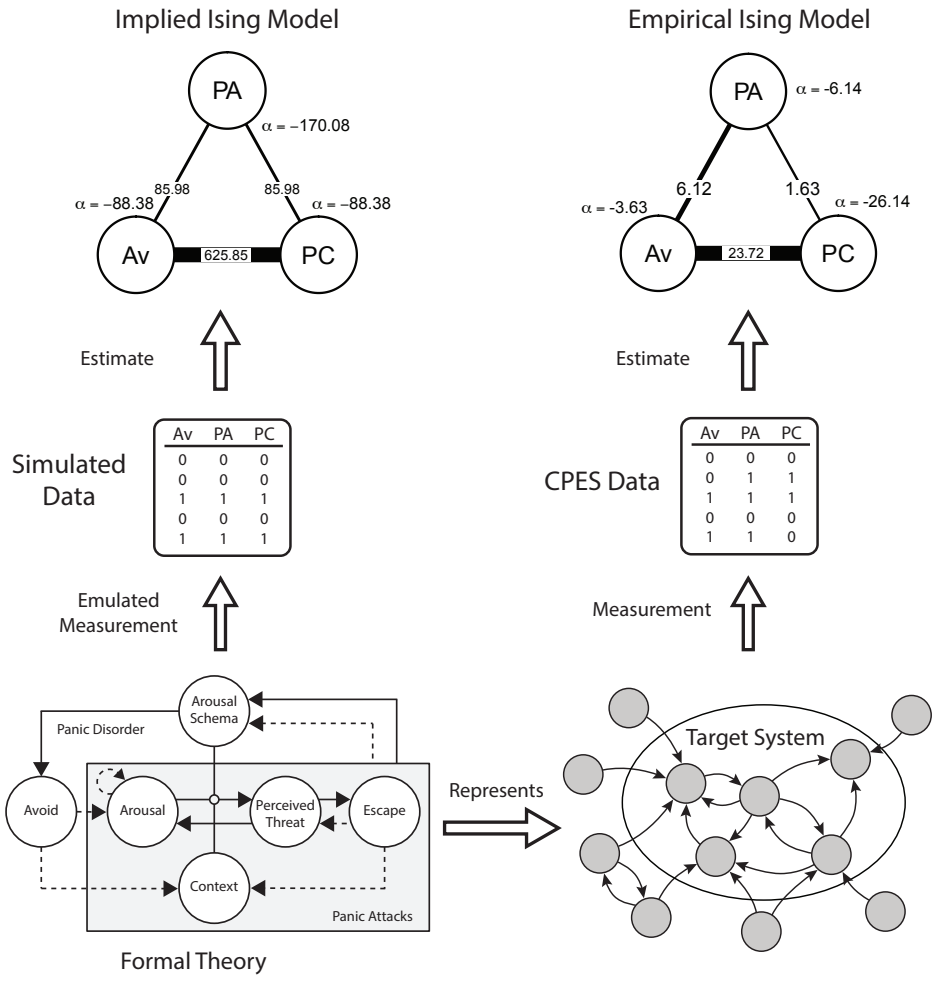


Figure 6.6: Illustration of the third route to formal theories. We take the Panic Model discussed in Section 6.2 as our formal theory, representing the unknown target system that gives rise to panic disorder. To obtain an implied data model from this theory, we first formalize how the components of the theory produce the data of interest, emulating the measurement process. With this in place, we can simulate data from the model in the form of cross-sectional binary symptom variables. We obtain the theory-implied Ising Model by estimating it from these simulated data (top-left corner). To estimate the empirical Ising Model (top-right corner) we make use of empirical measurements of binary symptom variables from the CPES dataset.

intra-individual time series data for multiple individuals (as described in Appendix 6.A) and so we need to define how cross-sectional symptom variables can be extracted from those time series. We specify that Recurrent Panic Attacks (PA = 1) are present for an individual in our simulated data if there are more than three panic attacks in the one month observation period. Persistent Concern is

determined using the average levels of jointly experienced arousal and perceived threat (i.e. anxiety) following a panic attack. If an individual has a panic attack, *and* their average anxiety following a panic attack exceeds a threshold determined by "healthy" simulations (i.e. those without panic attacks), they are classified as having Persistent Concern (PC = 1). Avoidance is defined similarly, with this symptom present if an individual has a panic attack, and their average levels of avoidant behavior following that attack are higher than we would expect to see in the healthy sample. A more detailed account of how we generated these data can be found in Appendix 6.C. This simulated cross-sectional data was used to estimate the implied Ising Model (top left-hand corner, Figure 6.6).[6]

We obtained the corresponding empirical Ising model (right-hand column of Figure 6.6) using the publicly available Collaborative Psychiatric Epidemiology Surveys (CPES) 2001-2003 (Alegria, Jackson, Kessler, & Takeuchi, 2007). The CPES is a nationally representative survey of mental disorders and correlates in the United States, with a total sample size of over twenty thousand participants (of which $n = 11367$ are used in the current analysis; for details see Appendix 6.C). The CPES combines more than 140 items relating to panic attacks and panic disorder, with a diagnostic manual describing how these items can be re-coded into binary symptom variables reflecting Recurrent Panic Attacks, Persistent Concern and Avoidance. PA is present if the participant reported more than three lifetime panic attacks. PC is present if, following an attack, the participant experienced a month or more of persistent concern or worry. Av is present if the participant reports either a month of avoidance behavior following an attack, or a general avoidance of activating situations in the past year.

6.3.3.2 Theory Development: Comparing Model-Implied and Empirical Data Models

As seen in Figure 6.6, there is a similar pattern of conditional dependencies in the implied and empirical data models. In both, all pairwise dependencies are positive, and all thresholds are negative. There is also a similar ordering of conditional dependencies in terms of their magnitude. Within each model, the conditional relationships of PA with Av and PA with PC are of the same order of magnitude, and the conditional relationship between Av and PC is an order of magnitude greater. However, we also see some differences between the models. First, the absolute value of pairwise dependencies and thresholds are much greater in the implied Ising Model (Figure 6.6 (a)) than the empirical Ising Model (Figure 6.6 (b)). Second, we see that the relationships in the implied model are perfectly symmetric, with exactly the same thresholds for Av and PC, and precisely the same weights relating PA to both. In the empirical network, these weights and thresholds are much smaller.

The bivariate contingency tables of all symptom-symptom relationships clarify the nature of these relationships (see Figure 6.7). In both the implied and empirical data models only a small proportion of individuals experience Recurrent Panic Attacks (Empirical 4.3%, Simulated 3.72%). Crucially, in the simulated

[6] Again the parameter estimates are somewhat unstable due to near-deterministic relationships.

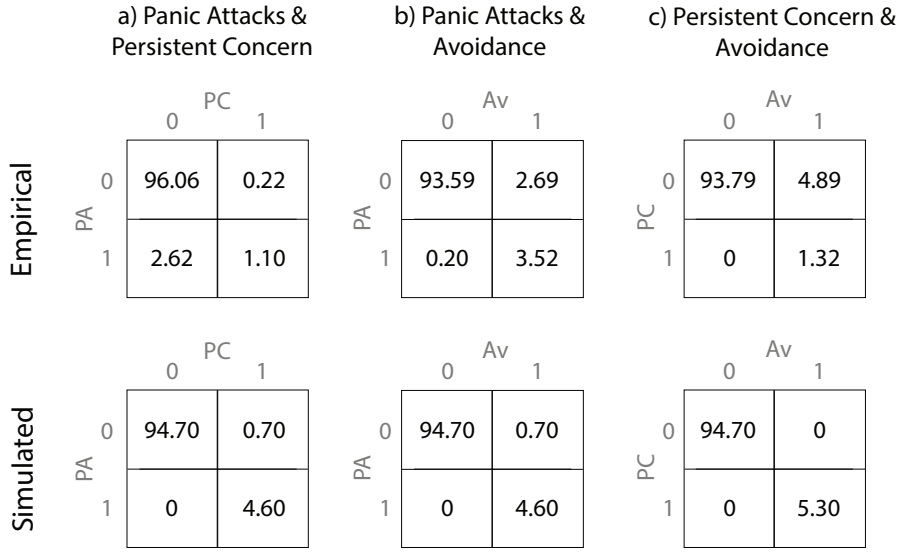


Figure 6.7: Contingency tables showing percentages for each pair of symptom variables (one per column) for the empirical data (top row) and simulated data (bottom row). The CPES contingency tables are based on $n_{CPES} = 11367$ observations. The simulated dataset contains $n_{sim} = 1000$ observations.

dataset, the symptom relationships are almost deterministic: If one symptom is present, so too are all others, and vice versa for the absence of symptoms (apart from three individuals who experience less than three panic attacks in the time window). This is because there is a deterministic relationship between the components underlying these symptoms in the Panic Model: All participants who experience one panic attack have Persistent Concern and Avoidance behavior after those attacks. In contrast, there are non-deterministic relationships in the empirical data. For example, it is actually more common to have Recurrent Panic Attacks without Persistent Concern than with Persistent Concern (column (a)). Similarly, more individuals experience Avoidance without Persistent Concern, than with Persistent Concern (column (c)) Conversely, there are no individuals who experience Persistent Concern but not Avoidance.

The discrepancies between the implied and the empirical data model could arise at any step in the process from formal theory/target system to implied and empirical data model illustrated in Figure 6.6. It could be the case that a discrepancy is due to inaccuracies in how we emulate the measurement process. For example, perhaps Persistent Concern and Avoidance co-occur equally, but the former suffers from a greater degree of recall bias than the latter (for an example of differential symptom recall bias in depressed patients, see Ben-Zeev & Young, 2010). There are also different time scales at which the simulated and empirical symptoms are defined. The simulated symptoms are defined over a month period whereas the CPES items are defined over lifetime prevalence. Due to the deterministic nature of the Panic Model, we believe a month period is a good ap-

proximation for lifetime experience of panic symptoms in this case. Nonetheless, it is a discrepancy in measurement that could lead to discrepancies between the implied and empirical data models. It could also be that the discrepancy is due to estimation issues. However, due to the large sample sizes and simple models used, we suspect it is unlikely that sampling variance is a problem in this instance. For present purposes, we will assume here that discrepancies are to due to inaccuracies in how the theory represents the target system, and as such we can use these discrepancies to directly evaluate our theory.[7]

On a global level there is a good match between the empirical and implied models: The theory implies positive symptom-symptom dependencies, which we also observe in the empirical data. However, the implied model over-estimates the strength of these relationships. This is largely explained by the deterministic causal effects in the theory. In the simulated data, everybody who experiences panic attacks also develops Persistent Concern and, in turn, Avoidance. As seen in Figure 6.7, this is inconsistent with empirical data, identifying a serious shortcoming in the theory. To improve the model, we must include some mechanism by which individuals can experience a panic attack without developing the remaining symptoms of panic disorder. In the empirical data, there is a near deterministic effect of Persistent Concern on Avoidance, suggesting that once Persistent Concern develops, Avoidance will follow; an observation that is consistent with the formal theory. However, inconsistent with the formal theory, the empirical data suggests a relatively low probability of Persistent Concern following panic attacks, suggesting that this is where the theory must be revised if it is to better account for the observation that some individuals experience recurrent panic attacks without developing the full panic disorder syndrome.

This is just one discrepancy in these data that can inform model development and more insights may be gained by focusing on others. Many more insights can be gained by considering different data models based on different data. For example, experimental data on the relation between Arousal and Perceived Threat may allow us to refine the specification of the feedback between those two variables. In general this route offers a great deal of flexibility in theory development. Although the theory is likely to be complex, dynamic and non-linear, the form of the data models used to learn about that theory need not be. Instead, by starting with an initial theory, the researcher can use any data about the phenomena of interest to further develop that theory. In the following section we will provide a full account of this third route by discussing the full process of theory development from establishing the phenomenon, to making novel predictions with a well-developed formal theory.

[7] In practice, inaccurate conceptualizations of how measurements represent the target system will be problematic for any approach to theory development or indeed any scientific endeavor, as evidenced by the growing attention on measurement in psychopathology literature (e.g., Flake & Fried, 2019). Our proposed approach to formal theory development forces us to be specific about the measurement process, just as we are forced to explicate the theory itself. Although we focus on the formalized theory itself here, we consider the formalization of measurement to be a significant advantage of this approach, and one that warrants further development.

6.4 An Abductive Approach to Formal Theory Construction

In Section 6.3.3 we illustrated a clear approach to use empirical data to develop an existing formal theory. However, our description of this approach so far has omitted several critical steps, including how to generate an initial formal theory and how to test that theory. In this section, we propose a three-stage process of formal theory construction, with an emphasis on the role data models play at each stage (see Figure 6.8). First, in the theory generation stage, we establish the phenomenon to be explained, generate an initial verbal theory, and formalize that theory. Second, in the theory development stage, the theory is developed beyond this initial proposal by adapting it such that it is consistent with as many empirical findings as possible. Finally, in the theory evaluation stage, the theory is subjected to strong tests within a hypothetico-deductive framework. The approach to theory construction proposed here places considerable emphasis on the theory's ability to explain phenomena, especially during the generation and development of the theory. Accordingly, the framework we have proposed is a largely abductive approach (Haig, 2005).

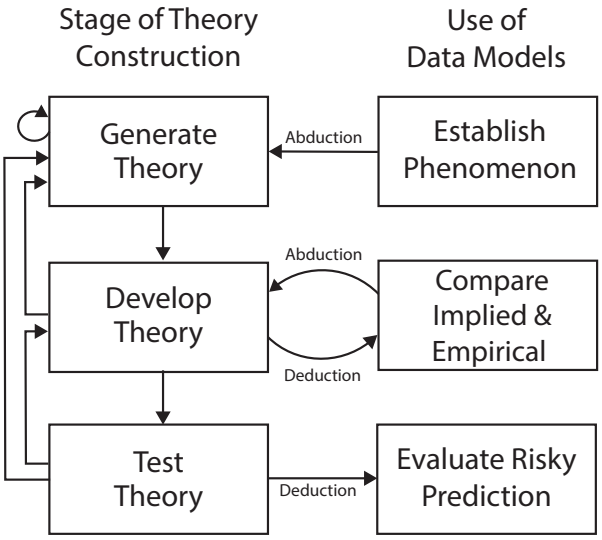


Figure 6.8: Flowchart depicting the process of developing a formal theory with the abductive approach put forward in this section. In the theory generation step we first establish the phenomenon (Section 6.4.1.1) and then generate an initial verbal theory (Section 6.4.1.2) which is subsequently formalized (Section 6.4.1.3). In the second step (Section 6.4.2) the theory is validated by testing whether it is consistent with existing empirical findings that are not part of the core phenomenon. If the formal theory is not consistent with some findings, it is adapted accordingly. If these adaptations lead to a "degenerative" theory (Meehl, 1990) we return to the first step; otherwise we continue to the final step, in which we test the formal theory using risky predictions (Section 6.4.3). If many tests are successful, we tentatively accept the theory. If not, the theory must either be adapted (step two) or a new theory generated (step one).

6.4.1 Generating Theory

6.4.1.1 Establishing the Phenomenon

The goal of a formal theory is to explain phenomena. Accordingly, the first step of theory development is to specify the set of phenomena to be explained. Establishing phenomena is a core aim of science and a full treatment of how best to achieve this aim is beyond the scope of this paper (for a possible way to organize this process see Haig, 2005). However, we suspect that the most appropriate phenomena for initial theory development will often include things that researchers would not think to subject to empirical analysis, as the most robust phenomena may simply be taken for granted as features of the real world. For example, in the case of panic disorder, the core phenomena to be explained are simply the observations that some people experience panic attacks and recurrent attacks tend to co-occur with persistent worry or concern about those attacks and avoidance of situations where such attacks may occur. These are empirical phenomena so robust that they are typically not empirically examined but instead simply assumed to be a feature of the real world.

6.4.1.2 Generate Initial Verbal Theory

Once the phenomena to be explained have been established, how do we go about generating an initial theory to explain them? A brief survey of well-known scientific theories reveals that this initial step into theory is often unstructured and highly creative. For example, in the 19th century August Kekulè dreamt of a snake seizing its own tail, leading him to generate the theory of the benzene ring, a major breakthrough in chemistry (Read, 1995). In the early 20th century, Alfred Wegener noticed that the coastlines of continents fit together similar to puzzle pieces, and consequently developed the theory of continental drift (Wegener, 1966), which formed the basis for the modern theory of plate tectonics (Mauger, Tarbuck, & Lutgens, 1996). In the late 20th century, Howard Gardner explained that he developed his theory of multiple intelligences in the 1980s using "subjective factor analysis" (Walters & Gardner, 1986, p. 176). Although more codified approaches to theory development exist (e.g., Grounded Theory; Strauss & Corbin, 1994), we are unaware of any evidence to suggest that any one approach to theory generation is superior to any other.

Nonetheless, the nature of theories does provide some guidance for how they might initially be generated. Theories achieve their aim of explaining phenomena by representing a target system. Accordingly, generating an initial theory will require that we specify the components thought to compose the target system. This process entails dividing the domain of interest into its constituent components (i.e. "partitioning") and selecting those components one believes must be included in the theory (i.e. "abstraction"), thereby producing a system of components that will be the object of the theory (cf. Elliott-Graves, 2014). For researchers adopting a "network perspective", the target system is typically presumed to comprise cognitive, emotional, behavioral, or physiological components, especially those identified in diagnostic criteria for mental disorders

(Borsboom, 2017). For example, as we have seen in Section 6.2, the target system could consist of Arousal, Perceived Threat, Avoidance and other symptoms of panic attacks or panic disorder. Having identified the relevant components we next specify the posited relations among them. For example, specifying that Perceived Threat leads to Arousal. Within the domain of the network approach, this second step will typically entail specifying causal relations among symptoms or momentary experiences (e.g., thoughts, emotions, and behavior).

Notably, in psychiatry, we do not necessarily need to rely on creative insight about the components and relations among them in order to generate an initial theory. There are already a plethora of verbal theories about mental disorders. If the initial verbal theory is well supported and specific, it will lend itself well to formalization and subsequent theory development. However, even poor verbal theories can be a useful starting point to developing a successful formal theory (Wimsatt, 1987; Smaldino, 2017).

6.4.1.3 Formalize Initial Theory

Once a verbal theory has been specified, the next step is to formalize the theory. To do so, we first need to choose a formal framework. A common formal framework is the use of difference or differential equations, which model how variables change across discrete time steps and continuous time, respectively (e.g., Strogatz, 2015). Specifically, the relations between components is specified by defining the rate of change of each component as a function of all other components and itself. The Panic Model, which we used as an example throughout the paper, uses this formal framework. Another common framework is Agent based Modeling (ABM), in which autonomous agents interact with each other using a set of specified rules (e.g., Grimm & Railsback, 2005). Here, each agent has local rules on how to interact with other agents. Both frameworks can be implemented in essentially any computer programming language and both are likely to be relevant to psychiatric and psychological research as a whole.

Having chosen a formal framework, the next step is to specify the relations between each component in the language of that framework. This process of formalizing relations is an exercise in being specific. Mathematics and computational programming languages require theorists to specify the precise nature of the relationship between variables. Requiring this level of specificity is one advantage of computational modeling, as it has the effect of immediately clarifying what remains unknown about the target system of interest, thereby guiding future research. However, this also means that theorists will often be in the position of needing to explicate relationships when the precise nature of those relationships is uncertain. We believe that, even in the face of this uncertainty, it is better to specify a precise relationship and be wrong than to leave the relationship ambiguously defined, as it is in a verbal theory. Nonetheless, we suspect that theorists will be on firmer foundation for subsequent theory development the more that they are able to draw on empirical data and other resources to inform this initial formal theory. There are several sources of information that can guide the formalization process.

First, empirical research can inform specification of components and the relations among them. For example, one could use the finding that sleep quality predicts next-day affect, but daytime affect does not predict next-night sleep (de Wild-Hartmann et al., 2013) to constrain the set of plausible relationships between those two variables in the formal theory. There could also be empirical data on the rate of change of variables, for example, Siegle, Steinhauer, Thase, Stenger, and Carter (2002) and Siegle, Steinhauer, Carter, Ramel, and Thase (2003) have shown that depressed individuals exhibit longer sustained physiological reactions to negative stimuli than healthy individuals, a finding which is echoed in self-report measures of negative affect (Houben et al., 2015).

Second, we can possibly derive reasonable scales for variables and relationships between variables from basic psychological science. For example, classical results from psychophysics show that increasing the intensity of stimuli in almost all cases leads to a nonlinear response in perception (e.g., Fechner, Howes, & Boring, 1966): When increasing the volume of music to a very high level, individuals cannot hear an additional increase.

Third, in many cases we can use definitions, basic logic, or common sense to choose formalizations. For example, by definition emotions should change at a time scale of minutes (Houben et al., 2015), while mood should only change at a time scale of hours or days (Larsen, 2000). And we can choose scales of some variables using common sense, for example one cannot sleep less than 0 and more than 24 hours a day, and heart rate should be somewhere between 50 and 180.

Fourth, we could use an existing formal model of another target system, which we expect to have a similar structure as the target system giving rise to the phenomenon of interest. This approach is called "analogical modeling". For example, Cramer et al. (2016) formulated a model for interactions between symptoms of Major Depression using the Ising model, which was originally formulated to model magnetism on an atomic level (Ising, 1925). Similarly, Fukano and Gunji formulated a model for interactions among core components of panic attacks using a Lotka-Volterra model originally formulated to represent predator-prey relationships (Fukano & Gunji, 2012).

Fifth, it is also important to note that there are methods by which we can potentially estimate the parameters for a formal theory from empirical data.[8] These approaches require considerable development of the formal theory (e.g., the form of a differential equation), suitable data (typically intensive longitudinal data), and a clear measurement model relating observed variables to theory components (as we did in Section 6.3.3). Accordingly, this approach already requires considerable progress in generating a formal theory and may be limited by practical considerations. Nonetheless, it remains a valuable resource that, if successfully carried out, would likely strengthen subsequent efforts at theory development.

[8]For example, if the theory is formalized in a system of differential equations, the parameters of such equations can in principle be estimated from time series data using, amongst others, Kalman filter techniques and state-space approaches (e.g., Einicke, 2019; Kulikov & Kulikova, 2013; Durbin & Koopman, 2012). For implementations of these estimation methods see Ou et al. (2019); Carpenter et al. (2017); King, Nguyen, and Ionides (2015)

The aim of this initial stage is to generate a formal theory able to explain a set of core phenomena. As we have emphasized throughout this paper, formal theories precisely determine the behavior implied by their theory. Accordingly, explanation in this context means that the theory has demonstrated its ability to produce the behavior of interest. For example, a theory of panic attacks must be able to produce sudden surges of arousal and perceived threat; a theory of depression must be able to produce sustained periods of low mood; and a theory of borderline personality disorder must be able to produce affective instability. We would note that there are very few theories in psychiatry that have reached this stage of not merely positing, but demonstrating, that the theory can explain the phenomena of interest. Accordingly, completing this stage of theory construction would constitute a significant advance in psychiatric theories. Once a theory has reached this stage, it is ready for the next stage of theory construction.

6.4.2 Developing Theory

The formal theory produced in the first stage of theory construction will have demonstrated its ability to explain the core phenomena of interest. However, the fact that the formal theory provides *some* explanation does not mean it is the *correct* explanation. In other words, demonstrating an ability to explain the phenomena of interest is a critical first step, but does not guarantee that the formal theory is a good representation of the target system. To achieve this aim, we propose a stage of theory development in which the theory is refined by deducing implied data models and comparing them to empirical data models. If the two data models align we take this as evidence that the current formal theory is adequately representing the target system; if there are discrepancies between the two data models, we analyze the nature of those discrepancies, consider the best explanation for how they arose, and adapt the formal theory to be able to account for these discrepancies.

This process of further developing a theory through comparisons between implied and empirical data models is exactly the process that we have illustrated already in Section 6.3.3. In that illustration, we derived the implied Ising model for the three variables Panic Attacks, Persistent Concern, and Avoidance, and we compared it to the empirical Ising model for the same variables. We discussed possible explanations for this discrepancy and corresponding adaptations of the Panic Model, for example including a mechanism by which individuals can experience Panic Attacks without experiencing Persistent Concern and Avoidance. If we were to continue in this stage of theory development, we would iterate this process, adapting the Panic Model to include such a mechanism, deriving the implied Ising model from this adapted Panic Model, and determining whether it better accounts for the empirical Ising model (i.e. the discrepancy between implied and empirical Ising models is smaller).

This stage of theory development can make use of many different types of data such as physiological, psychological and behavioral measurements from individuals, cross-sectional data such as clinical interviews and questionnaire data, or experimental data. Depending on the empirical phenomenon we would like to

account for, different kinds of data and data models will be appropriate. In general, however, more complex data models tend be more powerful tools to tease apart competing theories. For example, a great many formal theories might be consistent with a set of means, but it is likely that fewer are consistent with the means *and* the conditional relationships between the variables, captured, for example, by a GGM or Ising model. In other words, there are a more constrained number of possible formal theories that may account for more complex data models, thereby doing more to guide theory development.

There are two important considerations when working through this stage. First, it is important to consider how much trust to place in the empirical data at hand. Discrepancies between the empirical data model and theory-implied model may be due to shortcomings of the formal theory, but they may also be due to poor measurement, insufficient samples, or poor estimation of the parameters of the data model. How do we know when to adapt the formal theory in the face of some discrepancy? For some guidance on this question, we can draw on the large literature on model evaluation and model comparison. A straightforward way to decide whether to adapt the theory would be to derive an implied model of the adapted theory, and then compare the likelihood of the empirical data given the initial theory and the adapted theory. In order to decide whether to accept the adaptation we can use, for instance, a likelihood ratio test or a Bayes factor. This procedure ensures that we only make adaptations to our theory if we are certain enough that they actually lead to a better representation of the target system, and not only the idiosyncratic features of the empirical data at hand. That is, whether we accept an adaptation of the formal theory depends both on how large the improvement is, and how certain we are about it (i.e. how large the sample size is).

Second, it is important to consider the danger of making too many ad hoc revisions to the model that account only for idiosyncratic features of a given data model or, worse, yield new implications that are inconsistent with other empirical findings. For some guidance on this question, we can draw on the literature on theory evaluation from the philosophy of science literature (Meehl, 1990; Lakatos, 1976), which would suggest that the theory development phase has two possible outcomes. If the theory is adapted almost every time it is tested against empirical data, if those adaptations are making the theory increasingly unwieldy, and if additional changes are increasingly difficult to make without causing the theory to be inconsistent with earlier tested empirical findings, the theory can be considered to be "degenerative" (Meehl, 1990; Lakatos, 1976). In such a situation, the initial theory was inappropriate and we return to the first step to generate a different initial formal theory. On the other hand if modifications to a theory expand, rather than contract, its ability to account for other empirical data beyond those it was originally introduced to explain, then we can have greater confidence in those modifications and, in turn, the formal theory. Ultimately, theorists must strive for a balance between the simplicity of the model and its consistency with empirical data models.

The aim of the theory development stage is a formal theory that not only explains the core phenomena of interest, but is also consistent with a range of

empirical data models. Our confidence in such a theory will grow the more data models and the more complex the data models that are consistent with theory, especially if the theory is able to achieve this consistency with minimal ad hoc adjustments. In other words, if a theory has achieved these aims, we can be increasingly confident that it is a good representation of the target system and can prepare to subject the theory to more rigorous testing.

6.4.3 Testing Theory

Well developed formal theories allow one to derive unambiguous hypotheses (i.e. predictions), which in turn provide the opportunity to conduct strong tests of the theory. Notably, much of psychiatric research begins at this stage of hypothesis testing, typically through null hypothesis testing. However, the theories from which these hypotheses are derived are often unclear and, as we have argued in this paper, the process by which hypotheses are derived from these theories is opaque and likely prone to error. As a result, it is often unclear whether these hypothesis tests are an appropriate test of the theory and difficult to know how the results of such tests can further inform theory construction. This is perhaps not surprising as the hypothetico-deductive framework in which much of this research is conducted has very little to say about where these theories come from or how they should be developed (Haig, 2005). In contrast, in the framework proposed here, we have detailed a process by which formal theories can be generated and developed, equipping us to generate precise hypotheses which, in turn, allows us to better test the theory. Accordingly, while this framework is primarily abductive in nature with a focus on a theory's ability to explain phenomena, we also believe this framework substantially strengthens hypothesis testing as a tool for evaluating theories.

Importantly, the theory testing stage calls for strong tests of a theory: risky predictions (Meehl, 1990) that render the theory vulnerable to refutation. Strong tests have at least two key features. First, strong tests entail the prediction of observations that, absent the theory, we would not otherwise expect. For example, the panic disorder model we have discussed throughout this paper predicts that the time to recover from an induction of arousal-related bodily sensations should indicate vulnerability to panic attacks and, thus, should prospectively predict the onset of panic attacks (for details, see Robinaugh, Haslbeck, et al., 2019). To our knowledge, this is not a prediction that has arisen in the context of any other theory and has never been tested. A study testing and finding support for this prediction would lend more credence to this theory than a study testing a prediction we would otherwise expect (e.g., that recurrent panic attacks will be correlated with avoidance behavior).

Second, strong tests entail precise predictions. That is, a prediction that goes beyond merely positing a refutation of the null hypothesis (e.g., a statistically significant association) or even a directional prediction (e.g., a positive association) to instead make precise point predictions about what should be observed. For example, a very well-developed theory of panic disorder would be able to predict the precise value of perceived threat (or interval of perceived threat val-

ues) which are likely to result from a particular arousal-inducing manipulation (e.g., by breathing CO2 enriched air; Roberson-Nay et al., 2017). In other words, just as in the theory development stage, the theory testing stage calls for us to deduce the precise data models implied by our theory and to compare those implied data models to empirical data models. There is, thus, a fine distinction between theory testing and theory development in this framework. In the theory development stage, these comparisons are carried out in the spirit of improving upon and refining the theory. In the theory testing stage, these comparisons are carried out with the aim of subjecting the theory to refutation. A discrepancy between theory-implied and empirical data models in the development stage calls for refinement of the theory. A discrepancy between the theory-implied and empirical data models in the testing stage calls for the theorist to deeply consider the appropriateness of the theory.

Importantly, we are not proposing a "naive falsificationism" approach to theory testing in which a failed test requires abandonment of the theory (Meehl, 1990). A discrepancy between theory implied and empirical data models can provide an opportunity to improve upon a theory, returning us to the stage of theory development. Nonetheless, a risky test should entail risk and repeated failures at this stage should push the theorist toward the generation of a new competing theory. For that reason, we believe that these risky tests should be engaged in only when the theorist is sufficiently confident in the theory that they would be willing to stake its survival on the outcome of the test. To that end, we would make several recommendations. First, as we have stressed throughout this paper, formal theories will strengthen confidence in the predictions being tested by ensuring that they have been correctly deduced. Indeed, the level of specificity required for predictions to constitute a strong test of the theory all but requires that the theory be formalized. Second, the stage of theory development should be used to not only improve upon the theory, but also the assumptions about the instruments, measurements, and analyses that may also be responsible for any discrepancies between the theory-implied and empirical data models. The approach we have argued for in producing implied data models is helpful in this regard, as it forces the theory to formalize not only the theory, but also the measurement of variables; what we have termed "emulated measurement" (see Section 6.3.3). Strengthening confidence in these "emulated measurements" will strengthen the test of the theory, as such issues cannot so readily be blamed should the test fail. Third, research at this stage must be confirmatory in the strictest sense of the term (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). These studies should be preregistered, ideally with model simulations showing the precise theory, measurement, and analysis that will be used in the study.

If a theory fails a strong test, the decision of how to proceed depends upon what Meehl referred to as the "money in the bank" principle (Meehl, 1990): If a theory has a track record of success, it would be unwise to discard the theory in the face of a single, or even several, failed tests. For Meehl, money in the bank was accumulated by passing risky tests. A theory that has passed many such tests should be retained more readily than a theory with no such record. We would

argue for a broader conceptualization that draws on a wider range of criteria for theory appraisal, with particular emphasis on explanatory breadth. A formal theory that can explain a range of phenomena should be retained more readily than a theory that accounts for only a narrow set of phenomena. Nonetheless, we believe that any failure of a strong test should be taken as a serious challenge to the theory that, at a minimum, warrants careful consideration about how to proceed.

If a theory passes a strong test, it is corroborated, with the strength of corroboration proportional to the strength of the test. Notably, because strong tests all but require the evaluation of predictions made by the theory, a theory that has passed several such tests will have demonstrated a strong capacity for supporting prediction. Accordingly, a theory that has moved from generation, through development, and testing will emerge well equipped to support not only the explanation, but also the prediction and control of mental disorders.

6.5 Conclusions

In this paper, we have argued that psychiatry needs formal theories and we have examined how data models can best inform the development of such theories. We focused especially on the network approach to psychopathology and considered three possible routes by which conditional dependence networks may inform formal theories about how mental disorders operate as complex systems. We found that these data models were not themselves capable of representing the structure we presume will be needed for a theory of mental disorders. Perhaps more surprisingly, we also found that we were unable to draw clear and reliable inferences from data models about the underlying system. Together, these findings suggest that merely gathering data models alone is unlikely to readily inform a well-developed formal theory. Instead, we found that the most promising use of empirical data models for theory development was to compare them to "implied data models" derived from an initial formal theory. In this approach, formal theories play an active role in their own development, with initial formalized theories being refined over time through ongoing comparison of implied and empirical data models.

Importantly, our analysis is not a critique of the specific data models we examined here, nor is it a dismissal of their value. Quite the opposite. We believe these data models provide rich and valuable information about the relationships among components of a system. However, our analysis strongly suggests that the network approach to psychopathology cannot survive on these data models alone. Formal theory is needed if the network approach is to move toward the explanation, prediction, and control of mental disorders. Indeed, there is growing recognition that formal theories are needed if we are to avoid problems associated with conflicting empirical results (i.e. the "Replication Crisis"; Collaboration et al., 2015) and move toward an accumulation of knowledge in scientific research (e.g., Muthukrishna & Henrich, 2019; Szollosi & Donkin, 2019; Yarkoni, 2019; Ioannidis, 2014). Accordingly, as a field psychiatry must grapple not only with

methods for the collection and analysis of data, but also methods for the generation and development of formal theories

The research framework we proposed in Section 6.4 is intended to be a first step toward such a method of theory construction. At the heart of this approach is the use of formalized initial theories to start a cycle of theory development in which empirical data informs ongoing theory development and these improved theories inform subsequent empirical research. Critically, this research framework is not intended to suggest that all researchers must develop expertise in computational modeling. Data and the detection of robust empirical phenomena are central to psychiatric research in our proposed framework. However, our framework does suggest that, as a field, psychiatry must do more to develop expertise in computational modeling within its ranks. We suspect that it will only be through ongoing collaboration among theorists and empirical researchers that we will be able to leverage the empirical literature to produce genuine advances in our ability to explain, predict, and control psychopathology.

Appendix 6.A Simulated Data from the Panic Model

In this appendix we describe in more detail how the simulated datasets, presented in both Sections 6.3.2 and 6.3.3 are obtained.

Data is simulated from the Panic Model, the full specification of which is given by (Robinaugh, Haslbeck, et al., 2019), using the statistical programming language *R*. We use the Panic Model to generate time-series data of 1000 individuals, on a single minute time scale, for 12 weeks, using Euler's method with a step size of .001. This yields a total of $n_t = 12,0960$ repeated measurements per person. Each individual starts with a different initial value of arousal schema, drawn from a normal distribution with $\mu = 0.25$ and $\sigma = 0.0225$. The parameters of this distribution were chosen to roughly generate a representative number of panic disorder sufferers (for more details see Robinaugh, Haslbeck, et al., 2019). Otherwise each individual obtains the same parameter values and the same starting values on all processes, with the stochastic noise terms drawn using a different random seed for each individual. The mapping from this raw data to the variables used in the network models of Section 6.3.2 is described in the main text. Code to reproduce this data-generation scheme can be found in the reproducibility archive of this paper.[9]

Appendix 6.B Additional Details: The Panic Model and Statistical Dependencies

In this appendix we describe in more detail the patterns of statistical dependencies produced by the three data models fitted to data simulated from the Panic Model in Section 6.3.2. While in the main text we discuss the statistical dependencies between between Arousal and Perceived Threat, and Arousal Schema and Avoidance, here we focus only on the former. Key to the Arousal-Perceived Threat dependencies is the positive feedback loop between Arousal and Perceived Threat in the Panic Model (as described in Section 6.2). If Arousal and Perceived Threat become sufficiently elevated, this "vicious cycle" leads to runaway positive feedback, with a pronounced spike in both Arousal and Perceived Threat (i.e. a panic attack). This spike initiates a process of homeostatic feedback that brings Arousal down and suppresses Arousal below its baseline for a period of time after this panic attack, a period which we will refer to as a *recovery period*. The panic attack itself lasts about 30 minutes. However, the recovery period lasts for 2-3 hours (see Figure 6.9 panel (a)).

[9] https://osf.io/bnteg/

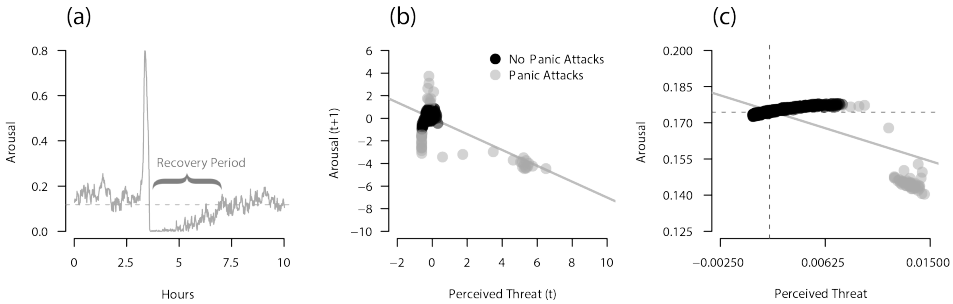


Figure 6.9: Panel (a) depicts Arousal during a panic attack, showing the short sharp peak of arousal levels, followed by a longer recovery period of low arousal, before the system returns to the usual resting state. The dotted line indicates the mean level of arousal over the observation window (0 - 10 hrs). Panel (b) depicts the state-space plot of Perceived Threat and Arousal at the next measurement occasion, as captured by the emulated ESM study and VAR model. Grey points indicate an observation window of 90 min in which either part of a panic attack or the following recovery period is captured. The solid grey line reflects the marginal lagged relationship. Panel (c) depicts the cross-sectional marginal relationship between the mean of Arousal and mean of Perceived Threat, as analyzed in the GGM model. Grey dots indicate individuals who suffer from panic attacks, and black dots represent "healthy" individuals. The solid grey line shows the negative marginal relationship. The dotted grey lines indicate the median of both variables, by which the binarized values used in the Ising model analysis are defined.

In the VAR model in Figure 6.5 (b) in the main text, we observed a strong negative conditional relationship between Perceived Threat at time t and Arousal at time $t+1$, conditioning on all other variables at time t. The distribution of these lagged variables is shown in Figure 6.9 (b), with the grey line representing the also negative marginal relationship. This strong negative cross-lagged relationship is a direct consequence of the recovery period of Arousal: High values of Perceived Threat are closely followed by a long period of low Arousal values. This can be seen in Figure 6.9 (b), where observations over windows in which a panic attack and recovery period occur are shaded in red. By averaging arousal values over a window of 90 minutes, the strong positive causal effects operating locally in time (i.e. over a very short time-interval) are not directly captured, but instead the VAR(1) model describes correctly describes the negative relationship between the *means* of each variable over this window.

In the GGM in Figure 6.5 (c) in the main text, we saw a positive linear relationship between Arousal and Perceived Threat in the estimated GGM. This dependency indicates that high mean levels of Arousal are associated with high mean levels of Perceived Threat, *conditional* on all other variables. We stress the conditional nature of this relationship, because the *marginal* relationship between the two variables is in fact negative as can be seen in Figure 6.9 (c). This negative marginal relationship comes about by combining two groups of individuals that have different mean values on both variables. Individuals who experience panic attacks (grey points) have high average Perceived Threat, but low average Arousal, due to the long recovery period of Arousal after a panic attack. On the other hand, individuals who do *not* experience panic attacks have higher average values of Arousal, and lower average values of Perceived Threat. When inspecting

the two groups separately, we see that there is a positive linear relationship between mean Arousal and Perceived Threat in the group without panic attacks; the group with panic attacks is too small to determine a relationship. Since Escape and Avoidance behavior only occur after Panic attacks, conditioning on those two variables amounts to conditioning on whether an individual had panic attacks. This conditional relationship is then driven mostly by the positive relationship in the (much larger) group of individuals who have no panic attacks, indicated by the black dots in Figure 6.9 (c).

Finally, we can explain the weak positive relationship between Arousal and Perceived Threat in the Ising model (Figure 6.5 (d) in the main text): The levels of these variables are defined by a median split of their mean values, depicted as dotted lines in Figure 6.9 (c). Unlike in the GGM, there is a positive marginal relationship between these binarized variables, as the majority of individuals without panic attacks (denoted by the black points) end up in the low Perceived Threat and low Arousal groups (lower left quadrant Figure 6.9 (c)) or high Perceived Threat and high Arousal groups (upper right quadrant). How then do we end up with a weakly positive conditional relationship between these two binary variables? Similarly to the GGM above, it turns out that conditioning on variables such as Escape behaviour and Avoidance almost entirely separates individuals into either the low Arousal and low Perceived Threat category (for low Escape values) or the high Arousal and high Perceived Threat category (for high Escape value). This means that, once we have conditioned on other variables which have direct and indirect causal connections to Arousal and Perceived Threat, there is very little additional information which Arousal can add to predicting Perceived Threat levels (and vice versa). This produces the weak positive conditional relationship between Arousal and Perceived Threat, as well as the stronger positive connections between Avoidance and Perceived Threat.

Appendix 6.C Details Empirical vs Simulated Ising Model

In this appendix we describe in more detail how the theory-implied and empirical Ising Models presented in Section 6.3.3 are obtained.

6.C.1 Simulated Data and Implied Ising Model

To obtain the theory-implied Ising Model we use the raw time series data generated from the Panic Model and described in Appendix 6.A

To create the binary symptom variables in Section 6.3.3 we transformed the raw time-series data of each individual as follows. First, we define Anxiety at a given time point as the geometric mean of the Arousal and Perceived Threat components at that point in time. Second, we define a panic attack as short, sharp peak of Arousal and Perceived Threat. We code a panic attack to be present in the time series data if Anxiety takes on a value greater than 0.5. The duration of a panic attack is the length of time Anxiety variable stays above this threshold,

and so we define a single panic attack as a sequence of consecutive time points in which Anxiety stays over this threshold. This allows to define our first binary symptom variable, Recurrent Panic Attacks:

1. Recurrent Panic Attacks (PA) : PA is present if the individual experience more than three panic attacks over the observation window.

We define recurrent as more than three over the observation window for consistency with how this symptom is defined in the CPES dataset, detailed below.

Next, we can define the symptom Persistent Concern (PC), again using the time series of Anxiety. This symptom is typically described as experiencing a heightened level of anxiety following a panic attack (American Psychiatric Association, 2013). To define this, for each individual who experiences a panic attack, we calculate the mean level of Anxiety in a window of 1000 minutes (16.67 hours) following the end of each panic attack. If another panic attack occurs in that window, we instead take the mean level of Anxiety between the end of one panic attack and before the beginning of the next. This gives us a vector of mean Anxiety levels per person, one for each panic attack experienced. Next, we must define what we consider to be a "heightened" level of anxiety. We do this by obtaining the distribution of mean Anxiety levels for healthy individuals, that is, those members of our sample who never experience a panic attack. We consider mean Anxiety levels following an attack to be "heightened" if they are greater than the 90th percentile of mean Anxiety levels in the healthy population. This gives us our second binary symptom variable.

2 Persistent Concern (PC): PC is present if, following at least one panic attack, higher average levels of Anxiety are present than in the healthy population, as defined by the 90th percentile of average Anxiety in the healthy population.

Finally we take a similar approach to defining the symptom Avoidance (Av), typically described as engaging in a heightened level of avoidance behaviour following a panic attack. For this symptom, we use the time series of the Avoid component. For each individual who experiences a panic attack, we calculate the mean level of Avoid in a window of 1000 minutes (16.67 hours) following the end of each panic attack, or before the beginning of the next attack, whichever is shorter. Heightened avoidance behaviour is defined relative to the 90th percentile of Avoid levels in the healthy population. This gives us our third binary symptom variable.

3 Avoidance (Av): Av is present if, following at least one panic attack, higher average levels of Avoid are present than in the healthy population, as defined by the 90th percentile of average Avoid in the healthy population.

The Ising model of these three symptom variables is fit using the *EstimateIsing* function from the *IsingSampler* package (Epskamp, 2015), that is, using a non-regularized pseudolikelihood method.

6.C.2 Empirical Symptom Data

To test the empirical predictions of the Panic Model, we made use of the publicly available Collaborative Psychiatric Epidemiology Surveys (CPES) 2001-2003 (Alegria et al., 2007). The CPES is a nationally representative survey of mental disorders and correlates in the United States. The CPES is attractive to use for testing the Panic Model, first because of the large sample size (20,013 participants) ensuring reliable estimates of empirical dependencies, and second, because approximately 146 items in the survey assess either panic attack or panic disorder experiences, characteristics, and diagnoses, typically in terms of lifetime prevalence.

To define our three panic disorder symptoms, we first use the diagnostic manual of the CPES to define whether individuals have ever experienced a panic attack based on responses to 18 items. There are three criteria which must be met for the individual to be classed as having experienced at least one lifetime panic attack. These are shown in Table 6.1. In coding the presence or absence of a panic attack, individuals must positively report at least four out of the thirteen characteristics of a panic attack, according to the second criteria in Table 6.1. Missing values were taken as a failure to report that characteristic.

Criterion	Description	Item number(s)
A	A discrete period of intense fear or discomfort	SC20 or SC20a
B (four or more)	Palplitations, pounding heart	PD1a
	Sweating	PD1e
	Trembling or shaking	PD1f
	Sensation of shortness of breath or smothering	PD1b
	Feeling of choking	PD1h
	Chest pain or discomfort	PD1i
	Nausea or abdominal distress	PD1c
	Feeling dizzy, unsteady, lightheaded or faint	PD1d or PD1m
	Derealization or depersonalization	PD1k or PD1l
	Fear of losing control or going crazy	PD1j
	Fear of dying	PD1n
	Paresthesias (numbing or tingling sensations)	PD1p
	Chills or hot flushes	PD1o
C	Symptoms developed abruptly and reached a peak within 10 minutes	PD3

Table 6.1: Description of the three criteria (A, B and C) necessary to code an individual as having one lifetime panic attack based on items from the CPES survey, based on the CPES diagnostic manual

With this definition of a panic attack in place, we define the three binary symptoms of panic disorder, following the definitions laid out in the diagonistic manual for Panic Disorder.

1. Recurrent Panic Attacks (PA). PA is present if participant reports more than three lifetime occurrences of an unexpected, short, sharp attack of fear or panic (item PD4 and all three criteria in Table 6.1), more than one of which is out of the blue (PD17a)

2. Persistent Concern (PC). PC is present if participants reports that following an attack, they experienced a month or more of at least one of: a) persistent concern about having another attack (PD13a), or b) worry about the implications or consequences of having an attack (PD13b)

3. Avoidance (Av). Av is present if participant reports at least one of a) following an attack, changing everyday activities for a month or more (PD13c), b) following an attack, avoiding situations due to fear of having an attack for a month or more (PD13d), or c) in the past 12 months, avoiding situations that might cause physical sensation (PD42).

In coding this, if two out of three PA criteria were present, and the third was missing, we assigned a positive value to the PA item. The empirical Ising model was fit using the same procedure as the theory-implied Ising model.

References

Aalen, O. O. (1987). Dynamic modelling and causality. *Scandinavian Actuarial Journal*, *1987*(3-4), 177–190.

Aalen, O. O., Borgan, Ø., Keiding, N., & Thormann, J. (1980). Interaction between life history events. nonparametric analysis for prospective and retrospective data in the presence of censoring. *Scandinavian Journal of Statistics*, 161–171.

Aalen, O. O., Gran, J., Røysland, K., Stensrud, M., & Strohmaier, S. (2018). Feedback and mediation in causal inference illustrated by stochastic process models. *Scandinavian Journal of Statistics*, *45*, 62–86.

Aalen, O. O., Røysland, K., Gran, J., Kouyos, R., & Lange, T. (2016). Can we believe the DAGs? A comment on the relationship between causal DAGs and mechanisms. *Statistical methods in medical research*, *25*(5), 2294–2314.

Aalen, O. O., Røysland, K., Gran, J., & Ledergerber, B. (2012). Causality, mediation and time: a dynamic viewpoint. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *175*(4), 831–861.

Abadir, K. M., & Magnus, J. R. (2005). *Matrix Algebra* (Vol. 1). Cambridge University Press.

Achen, C. H. (2005). Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science*, *22*(4), 327–339.

Alegria, M., Jackson, J. S., Kessler, R. C., & Takeuchi, D. (2007). *Collaborative Psychiatric Epidemiology Surveys (CPES), 2001-2003 [United States]*. Inter-university Consortium for Political and Social Research.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5th ed.). Washington, DC.

Andersson, S. A., Madigan, D., Perlman, M. D., et al. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, *25*(2), 505–541.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455.

Armour, C., Fried, E. I., Deserno, M. K., Tsai, J., & Pietrzak, R. H. (2017). A network analysis of DSM-5 posttraumatic stress disorder symptoms and correlates in us military veterans. *Journal of Anxiety Disorders*, *45*, 49–59.

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 359–388.

Atkinson, K. E. (1989). *An introduction to numerical analysis*. New York, NY: John Wiley & Sons.

Bailer-Jones, D. M. (2009). *Scientific models in philosophy of science*. Pittsburgh, PA: University of Pittsburgh Press.

Bak, M., Drukker, M., Hasmi, L., & van Os, J. (2016). An n=1 clinical network analysis of symptoms and treatment in psychosis. *PloS one*, *11*(9), e0162811.

Bastiaansen, J. A., Kunkels, Y. K., Blaauw, F., Boker, S. M., Ceulemans, E., Chen, M., ... et al. (2019, Mar). *Time to get personal? the impact of researchers' choices on the selection of treatment targets using the experience sampling methodology.* PsyArXiv. Retrieved from psyarxiv.com/c8vp7 doi: 10.31234/osf.io/c8vp7

Ben-Zeev, D., & Young, M. A. (2010). Accuracy of hospitalized depressed patients' and healthy controls' retrospective symptom reports: An experience sampling study. *The Journal of Nervous and Mental Disease*, *198*(4), 280–285.

Bernat, D. H., August, G. J., Hektner, J. M., & Bloomquist, M. L. (2007). The early risers preventive intervention: Testing for six-year outcomes and mediational processes. *Journal of abnormal child psychology*, *35*(4), 605–617.

Bisconti, T., Bergeman, C. S., & Boker, S. M. (2004). Emotional well-being in recently bereaved widows: A dynamical system approach. *Journal of Gerontology, Series B: Psychological Sciences and Social Sciences*, *59*, 158-167.

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *The Philosophical Review*, *97*(3), 303–352.

Boker, S. M. (2002). Consequences of continuity: The hunt for intrinsic properties within parameters of dynamics in psychological processes. *Multivariate Behavioral Research*, *37*(3), 405–422.

Boker, S. M., Deboeck, P., Edler, C., & Keel, P. (2010). Generalized local linear approximation of derivatives from time series. In S.-M. Chow & E. Ferrar (Eds.), *Statistical methods for modeling human dynamics: An interdisciplinary dialogue* (p. 179-212). Boca Raton, FL: Taylor & Francis.

Boker, S. M., & McArdle, J. J. (1995). Statistical vector field analysis applied to mixed crosssectional and longitudinal data. *Experimental Aging Research*, *21*(1), 77-93.

Boker, S. M., Montpetit, M. A., Hunter, M. D., & Bergeman, C. S. (2010). Modeling resilience with differential equations. In P. Molenaar & K. Newell (Eds.), *Learning and development: Individual pathways of change* (p. 183-206). Washington, DC: American Psychological Association.

Boker, S. M., Neale, M., & Rausch, J. (2004). Latent differential equation modeling with multivariate multi-occasion indicators. In K. van Montfort, J. H. Oud, & A. Satorra (Eds.), *Recent developments on structural equation models* (pp. 151–174). Dordrecht, the Netherlands: Kluwer Academic.

Boker, S. M., & Nesselroade, J. R. (2002). A method for modeling the intrinsic dynamics of intraindividual variability: Recovering parameters of simulated oscillators in multi-wave panel data. *Multivariate Behavioral Research*, *37*, 127–160.

Boker, S. M., Staples, A. D., & Hu, Y. (2016). Dynamics of change and change in dynamics. *Journal for person-oriented research*, *2*(1-2), 34–55.

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. NY: New York: The Guilford Press.

Bollen, K. A. (1987). Total, direct, and indirect effects in structural equation models. *Sociological methodology*, 37–69.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Oxford, England, John Wiley & Sons.

Bongers, S., & Mooij, J. M. (2018). From random differential equations to structural causal models: The stochastic case. *arXiv preprint arXiv:1803.08784*.

Borgatti, S. P. (2005). Centrality and network flow. *Social networks*, *27*(1), 55–71.

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5–13.

Borsboom, D., & Cramer, A. O. (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*, 91–121.

Borsboom, D., Cramer, A. O., Schmittmann, V. D., Epskamp, S., & Waldorp, L. J. (2011). The small world of psychopathology. *PloS one*, *6*(11), e27407.

Borsboom, D., van der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S., ... others (2012). What kind of causal modelling approach does personality research need? *European Journal of Personality*, *26*, 372–390.

Bos, F. M., Snippe, E., de Vos, S., Hartmann, J. A., Simons, C. J., van der Krieke, L., ... Wichers, M. (2017). Can we jump from cross-sectional to dynamic interpretations of networks implications for the network perspective in psychiatry. *Psychotherapy and Psychosomatics*, *86*(3), 175–177.

Boschloo, L., Schoevers, R. A., van Borkulo, C. D., Borsboom, D., & Oldehinkel, A. J. (2016). The network structure of psychopathology in a community sample of preadolescents. *Journal of Abnormal Psychology*, *125*(4), 599–606.

Bramsen, R. H., Lasgaard, M., Koss, M. P., Shevlin, M., Elklit, A., & Banner, J. (2013). Testing a multiple mediator model of the effect of childhood sexual abuse on adolescent sexual victimization. *American Journal of Orthopsychiatry*, *83*(1), 47–54.

Bringmann, L. F., Elmer, T., Epskamp, S., Krause, R. W., Schoch, D., Wichers, M., ... Snippe, E. (2019). What do centrality measures measure in psychological networks? *The journal of abnormal psychology*.

Bringmann, L. F., Lemmens, L., Huibers, M., Borsboom, D., & Tuerlinckx, F. (2015). Revealing the dynamic network structure of the beck depression inventory-ii. *Psychological medicine*, *45*(4), 747–757.

Bringmann, L. F., Pe, M. L., Vissers, N., Ceulemans, E., Borsboom, D., Vanpaemel, W., ... Kuppens, P. (2016). Assessing temporal emotion dynamics using networks. *Assessment*, *23*(4), 425–435.

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., ... Tuerlinckx, F. (2013). A Network Approach to Psychopathology: New Insights into Clinical Longitudinal Data. *PloS One*, *8*(4), e60188. doi: 10.1371/journal.pone.0060188

Browne, M. W., & Nesselroade, J. R. (2005). Representing psychological processes with dynamic factor models: Some promising uses and extensions of ARMA time series models. In A. Maydue-Olivares & J. J. McArdle (Eds.), *Psychometrics: A festschrift to Roderick P. McDonald* (p. 415-452). Mahwah, NJ: Lawrence Erlbaum Associates.

Bulteel, K., Tuerlinckx, F., Brose, A., & Ceulemans, E. (2016). Using raw var regression coefficients to build networks can be misleading. *Multivariate behavioral research*, *51*(2-3), 330–344.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

Cartwright, N. (1983). *How the laws of physics lie.* Oxford University Press.

Cartwright, N. (1999). Causal diversity and the markov condition. *Synthese*, *121*(1-2), 3–27.

Cartwright, N. (2007). *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge University Press.

Chasalow, S. (2012). combinat: combinatorics utilities [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=combinat` (R package version 0.0-8)

Chow, S.-M. (2019). Practical tools and guidelines for exploring and fitting linear and nonlinear dynamical systems models. *Multivariate Behavioral Research*, 1–29.

Chow, S.-M., Ferrer, E., & Hsieh, F. (2011). *Statistical methods for modeling human dynamics: An interdisciplinary dialogue*. New York: Routledge.

Chow, S.-M., Ferrer, E., & Nesselroade, J. R. (2007). An unscented kalman filter approach to the estimation of nonlinear dynamical systems models. *Multivariate Behavioral Research*, *42*(2), 283–321.

Chow, S.-M., Ou, L., Ciptadi, A., Prince, E. B., You, D., Hunter, M. D., … Messinger, D. S. (2018). Representing sudden shifts in intensive dyadic interaction data using differential equation models with regime switching. *Psychometrika*, *83*(2), 476–510.

Chow, S.-M., Ram, N., Boker, S., Fujita, F., Clore, G., & Nesselroade, J. (2005). Capturing weekly fluctuation in emotion using a latent differential structural approach. *Emotion*, *5*(2), 208–225.

Christian, C., Perko, V., Vanzhula, I., Tregarthen, J., Forbush, K., & Levinson, C. (2019). Eating disorder core symptoms and symptom pathways across developmental stages: A network analysis. *Journal of abnormal psychology*.

Clark, D. M. (1986). A cognitive approach to panic. *Behaviour Research and Therapy*, *24*(4), 461–470.

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology*, *112*(4), 558–557.

Coleman, J. S. (1968). The mathematical study of change. *Methodology in social research*, 428–478.

Collaboration, O. S., et al. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Contreras, A., Nieto, I., Valiente, C., Espinosa, R., & Vazquez, C. (2019). The study of psychopathology from the network analysis perspective: A systematic review. *Psychotherapy and Psychosomatics*, 1–13.

Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L. J., & Cramer, A. O. (2015). State of the aRt personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality*, *54*, 13–29.

Cox, D. R., & Wermuth, N. (1996). *Multivariate dependencies: Models, analysis and interpretation*. Chapman and Hall/CRC.

Cramer, A. O., van Borkulo, C. D., Giltay, E. J., van der Maas, H. L., Kendler, K. S., Scheffer, M., & Borsboom, D. (2016). Major depression as a complex dynamic system. *PLoS One*, *11*(12), e0167490.

Cramer, A. O., Waldorp, L. J., van der Maas, H. L. J., & Borsboom, D. (2010). Comorbidity: A network perspective. *Behavioral and Brain Sciences*, *33*(2-3), 137–150. doi: 10.1017/S0140525X09991567

Dablander, F., & Hinne, M. (2019). Node centrality measures are a poor substitute for causal inference. *Scientific reports*, *9*(1), 6846.

Dablander, F., Ryan, O., & Haslbeck, J. M. (2019, Jan). *Choosing between AR(1) and VAR(1) models in typical psychological applications.* PsyArXiv. Retrieved from psyarxiv.com/qgewy doi: 10.31234/osf.io/qgewy

Dahlhaus, R., & Eichler, M. (2003). Causality and graphical models in time series analysis. *Oxford Statistical Science Series*, 115–137.

Dalege, J., Borsboom, D., van Harreveld, F., van den Berg, H., Conner, M., & van der Maas, H. L. (2016). Toward a formalized account of attitudes: The Causal Attitude Network (CAN) model. *Psychological Review*, *123*(1), 2–22.

David, S. J., Marshall, A. J., Evanovich, E. K., & Mumma, G. H. (2018). Intraindividual dynamic network analysis–implications for clinical assessment. *Journal of psychopathology and behavioral assessment*, *40*(2), 235–248.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–31.

Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review*, *70*(2), 161–189.

Dawid, A. P. (2010). Beware of the DAG! In *Causality: Objectives and Assessment* (pp. 59–86).

Deboeck, P. R., & Preacher, K. J. (2016). No need to be discrete: A method for continuous time mediation analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 61–75.

De Haan-Rietdijk, S., Gottman, J. M., Bergeman, C. S., & Hamaker, E. L. (2016). Get over it! a multilevel threshold autoregressive model for state-dependent affect regulation. *Psychometrika*, *81*(1), 217–241.

De Haan-Rietdijk, S., Voelkle, M. C., Keijsers, L., & Hamaker, E. (2017). Discrete- versus continuous-time modeling of unequally spaced ESM data. *Frontiers in Psychology*, *8*, 1849.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, 157–175.

Denollet, J., & De Vries, J. (2006). Positive and negative affect within the realm of depression, stress and fatigue: The two-factor distress model of the global mood scale (gms). *Journal of affective disorders*, *91*(2), 171–180.

Deserno, M. K., Borsboom, D., Begeer, S., & Geurts, H. M. (2017). Multicausal systems ask for multicausal approaches: A network perspective on subjective well-being in individuals with autism spectrum disorder. *Autism*, *21*(8), 960–971.

de Wild-Hartmann, J. A., Wichers, M., van Bemmel, A. L., Derom, C., Thiery, E., Jacobs, N., ... Simons, C. J. (2013). Day-to-day associations between subjective sleep and affect in regard to future depressionin a female population-based sample. *The British Journal of Psychiatry*, *202*(6), 407–412.

Didelez, V. (2000). *Graphical models for event history analysis based on local independence*. Logos Berlin.

Didelez, V. (2007). Graphical models for composable finite Markov processes. *Scandinavian Journal of Statistics*, *34*(1), 169–185.

Didelez, V. (2008). Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(1), 245–264.

Didelez, V. (2019). Defining causal mediation with a longitudinal mediator and a survival outcome. *Lifetime data analysis*, *25*(4), 593–610.

Dormann, C., & Griffin, M. A. (2015). Optimal time lags in panel studies. *Psychological methods*, *20*(4), 489.

Driver, C. C., Oud, J. H., & Voelkle, M. C. (2017). Continuous time structural equation modeling with R package ctsem. *Journal of Statistical Software*, *77*(5), 1–35. doi: 10.18637/jss.v077.i05

Driver, C. C., & Voelkle, M. C. (2018). Understanding the time course of interventions with continuous time dynamic models. In K. L. Montfort, J. H. Oud, & M. C. Voelkle (Eds.), *Continuous time modeling in the behavioral and related sciences* (p. 179-203). New York: Springer.

Durbin, J., & Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford university press.

Ebbinghaus, H. (1913/2013). Memory: A contribution to experimental psychology. *Annals of Neurosciences*, *20*(4), 155.

Eberhardt, F. (2013). Experimental indistinguishability of causal structures. *Philosophy of Science*, *80*(5), 684–696.

Eichler, M., & Didelez, V. (2010). On Granger causality and the effect of interventions in time series. *Lifetime Data Analysis*, *16*(1), 3–32.

Einicke, G. A. (2019). *Smoothing, filtering and prediction: Estimating the past, present and future*. New York, NY: Prime Publishing.

Elliott-Graves, A. (2014). *The role of target systems in scientific practice* (Unpublished doctoral dissertation). University of Pennsylvania, Philadelphia, Pennsylvania.

Epskamp, S. (2015). IsingSampler: Sampling methods and distribution functions for the Ising model [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=IsingSampler (R package version 0.2)

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*(1), 195–212.

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, *48*(4), 1–18. Retrieved from http://www.jstatsoft.org/v48/i04/

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods*.

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network pschometrics: Combining network and latent variable models. *Psychometrika*, *82*(4), 904–927.

Epskamp, S., van Borkulo, C. D., van der Veen, D. C., Servaas, M. N., Isvoranu, A.-M., Riese, H., & Cramer, A. O. (2018). Personalized network modeling in psychopathology: The importance of contemporaneous and temporal connections. *Clinical Psychological Science*, *6*(3), 416–427.

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2018). The Gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research*, *53*(4), 453–480.

Epstein, J. M. (2008). Why model? *Journal of Artificial Societies and Social Simulation*, *11*(4), 12.

Fabio Di Narzo, A., Aznarte, J. L., & Stigler, M. (2009). tsDyn: Time series analysis based on dynamical systems theory [Computer software manual]. Retrieved from https://cran.r-project.org/package=tsDyn/vignettes/tsDyn.pdf (R package version 0.7)

Fechner, G. T., Howes, D. H., & Boring, E. G. (1966). *Elements of Psychophysics* (Vol. 1). Holt, Rinehart and Winston New York.

Fisher, A. J., & Boswell, J. F. (2016). Enhancing the personalization of psychotherapy with dynamic assessment and modeling. *Assessment*, *23*(4), 496–506.

Fisher, A. J., Reeves, J. W., Lawyer, G., Medaglia, J. D., & Rubel, J. A. (2017). Exploring the Idiographic Dynamics of Mood and Anxiety via Network Analysis. *Journal of Abnormal Psychology*, *126*(8), 1044–1056. doi: 0.1037/abn0000311

Fisher, M. (2001). *Modeling negative autoregression in continuous time.* (http://www.markfisher.net/ mefisher/papers/continuous_ar.pdf)

Flake, J. K., & Fried, E. I. (2019, Jan). *Measurement schmeasurement: Questionable measurement practices and how to avoid them.* PsyArXiv. Retrieved from psyarxiv.com/hs7wm doi: 10.31234/osf.io/hs7wm

Fonseca-Pedrero, E., Ortuño, J., Debbané, M., Chan, R. C., Cicero, D., Zhang, L. C., ... Fried, E. I. (2018). The network structure of schizotypal personality traits. *Schizophrenia Bulletin*, *44*(2), S468–S479.

Forré, P., & Mooij, J. M. (2018). Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. *arXiv preprint arXiv:1807.03024*.

Forré, P., & Mooij, J. M. (2019). Causal calculus in the presence of cycles, latent confounders and selection bias. *arXiv preprint arXiv:1901.00433*.

Freedman, H. I. (1980). *Deterministic mathematical models in population ecology* (Vol. 57). Marcel Dekker Incorporated.

Freedman, R. R., Ianni, P., Ettedgui, E., & Puthezhath, N. (1985). Ambulatory monitoring of panic disorder. *Archives of General Psychiatry*, *42*(3), 244–248.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35–41.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, *1*(3), 215–239.

Fried, E. I., Bockting, C., Arjadi, R., Borsboom, D., Amshoff, M., Cramer, A. O., ... Stroebe, M. (2015). From loss to loneliness: The relationship between bereavement and depressive symptoms. *Journal of Abnormal Psychology*, *124*(2), 256–265.

Fried, E. I., Boschloo, L., van Borkulo, C. D., Schoevers, R. A., Romeijn, J.-W., Wichers, M., ... Borsboom, D. (2015). Commentary:"consistent superiority of selective serotonin reuptake inhibitors over placebo in reducing depressed mood in patients with major depression". *Frontiers in Psychiatry*, *6*, 117.

Fried, E. I., & Cramer, A. O. (2017). Moving forward: challenges and directions for psychopathological network theory and methodology. *Perspectives on Psychological Science*, *12*(6), 999–1020.

Fried, E. I., Epskamp, S., Nesse, R. M., Tuerlinckx, F., & Borsboom, D. (2016). What are 'good' depression symptoms? comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis. *Journal of Affective Disorders*, *189*, 314–320.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1) (No. 10). New York, NY: Springer Series in Statistics.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441.

Fukano, T., & Gunji, Y.-P. (2012). Mathematical models of panic disorder. *Nonlinear Dynamics, Psychology, and Life Sciences*, *16*(4), 457–470.

Galles, D., & Pearl, J. (1995). Testing identifiability of causal effects. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (pp. 185–195).

Gardner, C., & Kleinman, A. (2019). Medicine and the mind — the consequences of psychiatry's identity crisis. *New England Journal of Medicine*, *381*(18), 1697-1699. Retrieved from https://doi.org/10.1056/NEJMp1910603 doi: 10.1056/NEJMp1910603

Gault-Sherman, M. (2012). It's a two-way street: The bidirectional relationship between parenting and delinquency. *Journal of Youth and Adolescence*, *41*, 121-145.

Glauber, R. J. (1963). Time-dependent statistics of the Ising model. *Journal of Mathematical Physics*, *4*(2), 294–307.

Glymour, M. M. (2006). Using causal diagrams to understand common problems in social epidemiology. *Methods in Social Epidemiology*, 393–428.

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PloS One*, *12*(6), e0174035.

Gollob, H. F., & Reichardt, C. S. (1987). Taking account of time lags in causal models. *Child Development*, *58*, 80–92.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 424–438.

Greenland, S., & Robins, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *International journal of epidemiology*, *15*(3), 413–419.

Grimm, V., & Railsback, S. F. (2005). *Individual-based modeling and e ecology*. Princeton University press.

Groen, R. N., Snippe, E., Bringmann, L. F., Simons, C. J., Hartmann, J. A., Bos, E. H., & Wichers, M. (2019). Capturing the risk of persisting depressive symptoms: A dynamic network investigation of patients' daily symptom experiences. *Psychiatry Research*, *271*, 640–648. doi: 10.1016/j.psychres.2018.12.054

Grömping, U. (2006). Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software*, *17*(1), 1–27.

Haig, B. D. (2005). An abductive theory of scientific method. *Psychological Methods*, *10*(4), 371–388.

Haig, B. D. (2008). Precis of 'an abductive theory of scientific method'. *Journal of Clinical Psychology*, *64*(9), 1019–1022.

Haig, B. D. (2014). *Investigating the psychological world: Scientific method in the behavioral sciences*. Cambridge, MA: MIT press.

Hamaker, E. L. (2012). Why researchers should think "within-person": A paradigmatic rationale. In M. R. Meehl & T. S. Conner (Eds.), (pp. 43–61). New York, NY: Guilford.

Hamaker, E. L., & Dolan, C. V. (2009). Idiographic data analysis: Quantitative methods—from simple to advanced. In *Dynamic process methodology in the social and developmental sciences* (pp. 191–216). Springer.

Hamaker, E. L., Dolan, C. V., & Molenaar, P. C. M. (2005). Statistical modeling of the individual: Rationale and application of multivariate time series analysis. *Multivariate Behavioral Research*, *40*(2), 207-233. doi: 10.1207/s15327906mbr4002_3

Hamaker, E. L., & Grasman, R. (2012). Regime switching state-space models applied to psychological processes: Handling missing data and making inferences. *Psychometrika*, *77*(2), 400–422.

Hamaker, E. L., Grasman, R. P., & Kamphuis, J. H. (2010). Regime-switching models to study psychological process. In P. C. M. Molenaar & K. M. Newell (Eds.), *Individual pathways of change: Statistical models for analyzing learning and development* (p. 155-168). Washington, DC: American Psychological Association.

Hamaker, E. L., Grasman, R. P., & Kamphuis, J. H. (2016). Modeling bas dysregulation in bipolar disorder: Illustrating the potential of time series analysis. *Assessment*, *23*(4), 436–446.

Hamaker, E. L., & Grasman, R. P. P. P. (2015). To center or not to center? investigating inertia with a multilevel autoregressive model. *Frontiers in Psychology*, *5*, 1492. doi: 10.3389/fpsyg.2014.01492

Hamaker, E. L., Kuiper, R., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*(1), 102-116. doi: 10.1037/a0038889

Hamaker, E. L., & Wichers, M. (2017). No time like the present: Discovering the hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science*, *26*(1), 10–15.

Hamaker, E. L., Zhang, Z., & van der Maas, H. L. (2009). Using threshold autoregressive models to study dyadic interactions. *Psychometrika*, *74*(4), 727–745.

Hamerle, A., Nagl, W., & Singer, H. (1991). Problems with the estimation of stochastic differential equations using structural equations models. *Journal of Mathematical Sociology*, *16*(3), 201–220.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society*, 357–384.

Hamilton, J. D. (1994). *Time series analysis* (Vol. 2). Princeton, NJ: Princeton University Press.

Haslbeck, J. M., Bringmann, L. F., & Waldorp, L. J. (2017). How to estimate time-varying vector autoregressive models? a comparison of two methods. *arXiv preprint arXiv:1711.05204*.

Haslbeck, J. M., Epskamp, S., Marsman, M., & Waldorp, L. J. (2018). Interpreting the Ising model: The input matters. *arXiv preprint arXiv:1811.02916*.

Haslbeck, J. M., & Fried, E. I. (2017). How predictable are symptoms in psychopathological networks? A reanalysis of 18 published datasets. *Psychological Medicine*, *47*(16), 2767–2776.

Haslbeck, J. M., & Ryan, O. (2019, Sep). *Recovering bistable systems from psychological time series.* PsyArXiv. Retrieved from psyarxiv.com/kcv3s doi: 10.31234/osf.io/kcv3s

Haslbeck, J. M., & Waldorp, L. J. (2018). How well do network models predict observations? on the importance of predictability in network models. *Behavior Research Methods*, *50*(2), 853–861.

Hayduk, L. A. (2009). Finite feedback cycling in structural equation models. *Structural Equation Modeling*, *16*(4), 658–675.

Heeren, A., & McNally, R. J. (2016). An integrative network approach to social anxiety disorder: The complex dynamic interplay among attentional bias for threat, attentional control, and symptoms. *Journal of Anxiety Disorders*, *42*, 95–104.

Hernan, M. A., & Robins, J. M. (2019). *Causal inference*. CRC Boca Raton, FL.

Higgs, P. W. (1964). Broken symmetries and the masses of gauge bosons. *Physical Review Letters*, *13*(16), 508–509.

Hirsch, M. W., Smale, S., & Devaney, R. L. (2012). *Differential equations, dynamical systems, and an introduction to chaos*. Academic press.

Hjelmeland, H., & Loa Knizek, B. (2018). The emperor's new clothes? A critical look at the interpersonal theory of suicide. *Death Studies*, 1–11. doi: 10.1080/07481187.2018.1527796

Ho, D. D., Neumann, A. U., Perelson, A. S., Chen, W., Leonard, J. M., & Markowitz, M. (1995). Rapid turnover of plasma virions and cd4 lymphocytes in hiv-1 infection. *Nature*, *373*(6510), 123–126.

Hoge, E. A., Bui, E., Marques, L., Metcalf, C. A., Morris, L. K., Robinaugh, D. J., ... Simon, N. M. (2013). Randomized controlled trial of mindfulness meditation for generalized anxiety disorder: effects on anxiety and stress reactivity. *The Journal of clinical psychiatry*, *74*(8), 786–792.

Hoorelbeke, K., Marchetti, I., De Schryver, M., & Koster, E. H. (2016). The interplay between cognitive risk and resilience factors in remitted depression: a network analysis. *Journal of Affective Disorders*, *195*, 96–104.

Horn, E. E., Strachan, E., & Turkheimer, E. (2015). Psychological distress and recurrent herpetic disease: A dynamic study of lesion recurrence and viral shedding episodes in adults. *Multivariate behavioral research*, *50*(1), 134–135.

Hosenfeld, B., Bos, E. H., Wardenaar, K. J., Conradi, H. J., van der Maas, H. L., Visser, I., & de Jonge, P. (2015). Major depressive disorder as a nonlinear dynamic system: bimodality in the frequency distribution of depressive symptoms over time. *BMC Psychiatry*, *15*(222).

Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin*, *141*(4), 901–930.

Ichii, K. (1991). Measuring mutual causation: Effects of suicide news on suicides in Japan. *Social Science Research*, *20*, 188-195.

Ioannidis, J. P. (2014). How to make more published research true. *PLOS Medicine*, *11*(10). doi: e1001747

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, *31*(1), 253–258.

Isvoranu, A.-M., van Borkulo, C. D., Boyette, L.-L., Wigman, J. T., Vinkers, C. H., Borsboom, D., & Investigators, G. (2016). A network approach to psychosis: pathways between childhood trauma and psychotic symptoms. *Schizophrenia Bulletin*, *43*(1), 187–196.

Johnston, J., & DiNardo, J. (1972). Econometric methods. *New York*, *19*(7), 22.

Jones, P. J. (2018). networktools: Tools for identifying important nodes in networks [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=networktools` (R package version 1.2.0)

Jones, P. J., Mair, P., Riemann, B. C., Mugno, B. L., & McNally, R. J. (2018). A network perspective on comorbid depression in adolescents with obsessive-compulsive disorder. *Journal of Anxiety Disorders*, *53*, 1–8.

Kaiser, T., & Laireiter, A.-R. (2018). Daily dynamic assessment and modelling of intersession processes in ambulatory psychotherapy: A proof of concept study. *Psychotherapy Research*, 1–12.

Kalisch, M., & Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, *8*(Mar), 613–636.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., & Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, *47*(11), 1–26. Retrieved from `http://www.jstatsoft.org/v47/i11/`

Kalisch, R., Cramer, A. O., Binder, H., Fritz, J., Leertouwer, I., Lunansky, G., ... Van Harmelen, A.-L. (2019). Deconstructing and reconstructing resilience: a dynamic network approach. *Perspectives on Psychological Science*, *14*(5), 765–777.

Kellen, D. (2019). A model hierarchy for psychological science. *Computational Brain & Behavior*, *2*(3-4), 160–165.

Kendler, K. S. (2019). From many to one to many—the search for causes of psychiatric illness. *JAMA Psychiatry*. doi: 10.1001/jamapsychiatry.2019.1200

Kim, C.-J., & Nelson, C. R. (1999). *State-space models with regime switching: Classical and Gibbs-sampling approaches with applications*. Cambridge, MA: The MIT Press. doi: 10.2307/2669796

King, A. A., Nguyen, D., & Ionides, E. L. (2015). Statistical inference for partially observed Markov processes via the R package pomp. *arXiv preprint arXiv:1509.00503*.

Knefel, M., Tran, U. S., & Lueger-Schuster, B. (2016). The association of posttraumatic stress disorder, complex posttraumatic stress disorder, and borderline personality disorder from a network analytical perspective. *Journal of Anxiety Disorders*, *43*, 70–78.

Kossakowski, J., Groot, P., Haslbeck, J. M., Borsboom, D., & Wichers, M. (2017). Data from 'critical slowing down as a personalized early warning signal for depression'. *Journal of Open Psychology Data*, *5*(1).

Koval, P., Kuppens, P., Allen, N. B., & Sheeber, L. (2012). Getting stuck in depression: The roles of rumination and emotional inertia. *Cognition and Emotion*, *26*, 1412-1427.

Kraemer, N., Schaefer, J., & Boulesteix, A.-L. (2009). Regularized estimation of large-scale gene regulatory networks using gaussian graphical models. *BMC Bioinformatics*, *10*(384).

Kroeze, R., van der Veen, D. C., Servaas, M. N., Bastiaansen, J. A., Oude Voshaar, R., Borsboom, D., & Riese, H. (2017). Personalized feedback on symptom dynamics of psychopathology: A proof-of-principle study. *Journal for Person-Oriented Research*, *3*(1), 1–11.

Kuiper, R. M., & Ryan, O. (2018). Drawing conclusions from cross-lagged relationships: Re-considering the role of the time-interval. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(5), 809–823.

Kulikov, G. Y., & Kulikova, M. V. (2013). Accurate numerical implementation of the continuous-discrete extended Kalman filter. *IEEE Transactions on Automatic Control*, *59*(1), 273–279.

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychological Science*, *21*(7), 984–991.

Kuppens, P., Sheeber, L. B., Yap, M. B. H., Whittle, S., Simmons, J., & Allen, N. B. (2012). Emotional inertia prospectively predicts the onset of depression in adolescence. *Emotion*, *12*, 283-289.

Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In S. Harding (Ed.), *Can theories be refuted?* (pp. 205–259). Dordrecht: Springer.

Larsen, R. J. (2000). Toward a science of mood regulation. *Psychological Inquiry*, *11*(3), 129–141.

Lauritzen, S. L. (1996). *Graphical Models* (Vol. 17). Clarendon Press.

Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, 1350–1360.

Levina, E., Rothman, A., Zhu, J., et al. (2008). Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics*, *2*(1), 245–263.

Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. SAGE publications.

Liu, S., Kuppens, P., & Bringmann, L. F. (2019). On the use of empirical bayes estimates as measures of individual traits. *Assessment*, 1073191119885019.

Lodewyckx, T., Tuerlinckx, F., Kuppens, P., Allen, N. B., & Sheeber, L. (2011). A hierarchical state space approach to affective dynamics. *Journal of mathematical psychology*, *55*(1), 68–83.

Lunansky, G., van Borkulo, C. D., & Borsboom, D. (2019, May). *Personality, resilience, and psychopathology: A model for the interaction between slow and fast network processes in the context of mental health.* PsyArXiv. Retrieved from psyarxiv.com/mznbw doi: 10.31234/osf.io/mznbw

Lutz, W., Ehrlich, T., Rubel, J., Hallwachs, N., Röttger, M.-A., Jorasz, C., ... Tschitsaz-Stucki, A. (2013). The ups and downs of psychotherapy: Sudden gains and sudden losses identified with session reports. *Psychotherapy Research*, *23*(1), 14–24.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin*, *114*(1), 185–199.

Malinsky, D., & Danks, D. (2018). Causal discovery algorithms: A practical guide. *Philosophy Compass*, *13*(1), e12470.

Marks, R. J. I. (2012). *Introduction to Shannon Sampling and Interpolation Theory*. Springer Science & Business Media.

Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., van Bork, R., Waldorp, L., ... Maris, G. (2018). An introduction to network psychometrics: Relating ising network models to item response theory models. *Multivariate Behavioral Research*, *53*(1), 15–35.

Mauger, R., Tarbuck, E. J., & Lutgens, F. K. (1996). *Earth: An introduction to physical geology.* Prentice-Hall.

Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods*, *12*(1), 23–44.

McNally, R. J. (2016). Can network analysis transform psychopathology? *Behaviour Research and Therapy*, *86*, 95–104.

McNally, R. J., Robinaugh, D. J., Wu, G. W., Wang, L., Deserno, M. K., & Borsboom, D. (2015). Mental disorders as causal systems: A network approach to posttraumatic stress disorder. *Clinical Psychological Science*, *3*(6), 836–849.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141.

Meier, B. P., & Robinson, M. D. (2004). Why the sunny side is up: Associations between affect and vertical position. *Psychological science*, *15*(4), 243–247.

Moberly, N. J., & Watkins, E. R. (2008). Ruminative self-focus and negative affect: An experience sampling study. *Journal of Abnormal Psychology*, *117*, 314-323.

Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, *2*(4), 201–218.

Molenaar, P. C. (2008). Consequences of the ergodic theorems for classical test theory, factor analysis, and the analysis of developmental processes. *Handbook of Cognitive Aging*, 90–104.

Molenaar, P. C., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current directions in psychological science*, *18*(2), 112–117.

Moler, C., & Van Loan, C. (2003). Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM review*, *45*(1), 3–49.

Mooij, J. M., Janzing, D., Heskes, T., & Schölkopf, B. (2011). On causal discovery with cyclic additive noise models. In *Advances in Neural Information Processing Systems* (pp. 639–647).

Mooij, J. M., Janzing, D., & Schölkopf, B. (2013). From ordinary differential equations to structural causal models: the deterministic case. *arXiv preprint arXiv:1304.7920*.

Murphy, K. P. (2012). *Machine Learning: a Probabilistic Perspective*. MIT press.

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229.

Nelson, B., McGorry, P. D., Wichers, M., Wigman, J. T., & Hartmann, J. A. (2017). Moving from static to dynamic models of the onset of mental disorder: a review. *JAMA Psychiatry*, *74*(5), 528–534.

Newman, M. (2018). *Networks*. Oxford university press.

Newton, I. (1687). *Philosophiæ naturalis principia mathematica (mathematical principles of natural philosophy)*. London.

Nguyen, J., & Frigg, R. (2017). Mathematics is not the only language in the book of nature. *Synthese*, 1–22.

Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social networks*, *32*(3), 245–251.

Oravecz, Z., & Tuerlinckx, F. (2011). The linear mixed model and the hierarchical ornstein–uhlenbeck model: Some equivalences and differences. *British Journal of Mathematical and Statistical Psychology*, *64*(1), 134–160.

Oravecz, Z., Tuerlinckx, F., & Vandekerckhove, J. (2009). A hierarchical Ornstein-Uhlenbeck model for continuous repeated measurement data. *Psychometrika*, *74*, 395–418.

Oravecz, Z., Tuerlinckx, F., & Vandekerckhove, J. (2011). A hierarchical latent stochastic difference equation model for affective dynamics. *Psychological Methods*, *16*, 468–490.

Oravecz, Z., Tuerlinckx, F., & Vandekerckhove, J. (2016). Bayesian data analysis with the bivariate hierarchical ornstein-uhlenbeck process model. *Multivariate behavioral research*, *51*(1), 106–119.

Ou, L., Hunter, M. D., & Chow, S.-M. (2019). dynr: Dynamic modeling in R [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=dynr (R package version 0.1.14-9)

Oud, J. H. (2007). Continuous time modeling of reciprocal relationships in the cross-lagged panel design. In S. M. Boker & M. J. Wenger (Eds.), *Data analytic techniques for dynamic systems in the social and behavioral sciences* (p. 87-129). Mahwah, NJ: Lawrence Erlbaum Associates.

Oud, J. H., & Delsing, M. J. M. H. (2010). Continuous time modeling of panel data by means of SEM. In K. van Montfort, J. H. Oud, & A. Satorra (Eds.), *Longitudinal research with latent variables* (pp. 201–244). New York, NY: Springer.

Oud, J. H., & Jansen, R. A. (2000). Continuous time state space modeling of panel data by means of SEM. *Psychometrika*, *65*(2), 199–215.

Oud, J. H., van Leeuwe, J., & Jansen, R. (1993). Kalman filtering in discrete and continuous time based on longitudinal lisrel models. *Advances in longitudinal and multivariate analysis in the behavioral sciences, ITS, Nijmegen, Netherlands*, 3–26.

Oud, J. H., Voelkle, M. C., & Driver, C. C. (2018). Sem based carma time series modeling for arbitrary n. *Multivariate behavioral research*, *53*(1), 36–56.

Papoulis, A., & Pillai, S. U. (2002). *Probability, Random Variables, and Stochastic Processes*. Tata McGraw-Hill Education.

Pe, M. L., Kircanski, K., Thompson, R. J., Bringmann, L. F., Tuerlinckx, F., Mestdagh, M., ... others (2015). Emotion-Network Density in Major Depressive Disorder. *Clinical Psychological Science*, *3*(2), 292–300. doi: 10.1177/2167702614540645

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J., & Verma, T. (1991). A theory of inferred causation. In J. Allen, R. Fikes, & E. Sandewall (Eds.), *Principles of knowledge representation and reasoning: Proceedings of the second international conference* (pp. 441–452). San Mateo, CA.

Pero, F. (2015). *Whither structuralism for scientific representation?* (Unpublished doctoral dissertation). Univeristy of Florence.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Raykov, T., & Marcoulides, G. A. (2001). Can there be infinitely many models equivalent to a given covariance structure model? *Structural Equation Modeling*, *8*(1), 142–149.

Read, J. (1995). *From alchemy to chemistry*. Courier Corporation.

Reichardt, C. S. (2011). Commentary: Are three waves of data sufficient for assessing mediation? *Multivariate Behavioral Research*, *46*(5), 842–851.

Richards, A., French, C. C., Johnson, W., Naparstek, J., & Williams, J. (1992). Effects of mood manipulation and anxiety on performance of an emotional stroop task. *British Journal of Psychology*, *83*(4), 479–491.

Richards, A., & Whittaker, T. M. (1990). Effects of anxiety and mood manipulation in autobiographical memory. *British Journal of Clinical Psychology*, *29*(2), 145–153.

Richardson, T. S., & Robins, J. M. (2013). Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, *128*(30), 2013.

Roberson-Nay, R., Gorlin, E. I., Beadel, J. R., Cash, T., Vrana, S., & Teachman, B. A. (2017). Temporal stability of multiple response systems to 7.5% carbon dioxide challenge. *Biological psychology*, *124*, 111–118.

Robinaugh, D. J., Haslbeck, J. M., Waldorp, L. J., Kossakowski, J. J., Fried, E. I., Millner, A. J., ... Borsboom, D. (2019). *Advancing the network theory of mental disorders:a computational model of panic disorder.* PsyArXiv. Retrieved from psyarxiv.com/km37w doi: 10.31234/osf.io/km37w

Robinaugh, D. J., Hoekstra, R. H., Toner, E. R., & Borsboom, D. (2019). The network approach to psychopathology: a review of the literature 2008–2018 and an agenda for future research. *Psychological Medicine*, 1–14. doi: 10.1017/S0033291719003404

Robinaugh, D. J., LeBlanc, N. J., Vuletich, H. A., & McNally, R. J. (2014). Network analysis of persistent complex bereavement disorder in conjugally bereaved adults. *Journal of abnormal psychology*, *123*(3), 510–522.

Robinaugh, D. J., Millner, A. J., & McNally, R. J. (2016). Identifying highly influential nodes in the complicated grief network. *Journal of Abnormal Psychology*, *125*(6), 747–757.

Robins, J. M. (1999). Association, causation, and marginal structural models. *Synthese*, *121*(1-2), 151–179.

Robins, J. M. (2003). Semantics of causal dag models and the identification of direct and indirect effects. In P. Green, N. Hjort, & S. Richardson (Eds.), *Highly structured stochastic systems* (p. 70-81). New York, NY: Oxford University Press.

Robins, J. M., & Richardson, T. S. (2010). Alternative graphical causal models and the identification of direct effects. *Causality and psychopathology: Finding the determinants of disorders and their cures*, 103–158.

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42.

Rovine, M. J., & Walls, T. A. (2006). Multilevel autoregressive modeling of interindividual differences in the stability of a process. In T. A. Walls & J. L. Schafer (Eds.), *Models for intensive longitudinal data* (pp. 124–147). New York, NY: Oxford University Press.

Rubel, J. A., Fisher, A. J., Husen, K., & Lutz, W. (2018). Translating person-specific network models into personalized treatments: Development and demonstration of the dynamic assessment treatment algorithm for individual networks (data-in). *Psychotherapy and psychosomatics*, *87*(4), 249–252.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688–701.

Ryan, O., Bringmann, L. F., & Schuurman, N. (2019, Oct). *The challenge of generating causal hypotheses using network models.* PsyArXiv. Retrieved from psyarxiv.com/ryg69 doi: 10.31234/osf.io/ryg69

Ryan, O., & Hamaker, E. L. (2019). *Time to intervene: A continuous-time approach to network analysis and centrality.* (Manuscript in Preparation)

Ryan, O., Kuiper, R. M., & Hamaker, E. L. (2018). A continuous time approach to intensive longitudinal data: What, why and how? In K. L. Montfort, J. H. Oud, & M. C. Voelkle (Eds.), *Continuous time modeling in the behavioral and related sciences* (pp. 29–57). New York: Springer.

Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., ... Sugihara, G. (2009). Early-warning signals for critical transitions. *Nature*, *461*(7260), 53.

Scheffer, M., Bolhuis, J. E., Borsboom, D., Buchman, T. G., Gijzel, S. M., Goulson, D., ... others (2018). Quantifying resilience of humans and other animals. *Proceedings of the National Academy of Sciences*, *115*(47), 11883–11890.

Schmittmann, V. D., Cramer, A. O., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New ideas in psychology*, *31*(1), 43–53.

Schuurman, N. K., Ferrer, E., de Boer-Sonnenschein, M., & Hamaker, E. L. (2016). How to compare cross-lagged associations in a multilevel autoregressive model. *Psychological Methods*, *21*(2), 206-221.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Schweder, T. (1970). Composable Markov processes. *Journal of Applied Probability*, *7*(2), 400–410.

Scutari, M. (2010). Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, *35*(3), 1–22. doi: 10.18637/jss.v035.i03

Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, *7*(Oct), 2003–2030.

Shojaie, A., & Michailidis, G. (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, *97*(3), 519–538.

Siegle, G. J., Steinhauer, S. R., Carter, C. S., Ramel, W., & Thase, M. E. (2003). Do the seconds turn into hours? relationships between sustained pupil dilation in response to emotional information and self-reported rumination. *Cognitive Therapy and Research*, *27*(3), 365–382.

Siegle, G. J., Steinhauer, S. R., Thase, M. E., Stenger, V. A., & Carter, C. S. (2002). Can't shake that feeling: event-related fMRI assessment of sustained amygdala activity in response to emotional information in depressed individuals. *Biological Psychiatry*, *51*(9), 693–707.

Smaldino, P. E. (2017). Models are stupid, and we need more of them. In R. Vallacher, S. Read, & A. Nowak (Eds.), *Computational social psychology* (pp. 311–331). New York, NY: Routledge.

Smith, E. R., & Conrey, F. R. (2007). Agent-based modeling: A new approach for theory building in social psychology. *Personality and Social Psychology Review*, *11*(1), 87–104.

Snippe, E., Viechtbauer, W., Geschwind, N., Klippel, A., De Jonge, P., & Wichers, M. (2017). The impact of treatments for depression on the dynamic network structure of mental states: Two randomized controlled trials. *Scientific Reports*, *7*, 46523.

Spector, P. E., & Brannick, M. T. (2011). Methodological urban legends: The misuse of statistical control variables. *Organizational Research Methods*, *14*(2), 287–305.

Spirtes, P. (1995). Directed cyclic graphical representations of feedback models. In *Proceedings of the eleventh conference on uncertainty in artificial intelligence* (pp. 491–498).

Spirtes, P. (2010). Introduction to causal inference. *Journal of Machine Learning Research*, *11*(May), 1643–1662.

Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G., & Richardson, T. (2000). *Causation, Prediction, and Search*. MIT press.

Spirtes, P., Meek, C., & Richardson, T. (1995). Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (pp. 499–506).

Spirtes, P., & Zhang, K. (2016). Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, *3*.

Spitzer, R. L., Kroenke, K., & Williams, J. B. (1980). *Diagnostic and statistical manual of mental disorders*. Washington, DC: American Psychiatric Association.

Steele, J. S., & Ferrer, E. (2011). Latent differential equation modeling of self-regulatory and coregulatory affective processes. *Multivariate Behavioral Research*, *46*(6), 956–984.

Steele, R. J., & Raftery, A. E. (2010). Performance of bayesian model selection criteria for gaussian mixture models. *Frontiers of Statistical Decision Making and Bayesian Analysis*, *2*, 113–130.

Stiles, W. B., Leach, C., Barkham, M., Lucock, M., Iveson, S., Shapiro, D. A., ... Hardy, G. E. (2003). Early sudden gains in psychotherapy under routine clinic conditions: Practice-based evidence. *Journal of Consulting and Clinical Psychology*, *71*(1), 14–21.

Strauss, A., & Corbin, J. (1994). Grounded theory methodology. *Handbook of qualitative research*, *17*, 273–85.

Strogatz, S. H. (2015). *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. Colorado, USA: Westview press.

Stutz, C., & Williams, B. (1999). Obituary: Ernst Ising. *Physics Today*, *52*, 106–108.

Suárez, M., & Pero, F. (2019). The representational semantic conception. *Philosophy of Science*, *86*(2), 344–365.

Suls, J., Green, P., & Hillis, S. (1998). Emotional reactivity to everyday problems, affective inertia, and neuroticism. *Personality and Social Psychology Bulletin*, *24*(2), 127–136.

Suppes, P. (1962). Models of data. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, methodology and the philosophy of science: Proceedings of the 1960 international congress.* CA, Stanford University Press.

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese*, *87*(3), 449–508.

Szollosi, A., & Donkin, C. (2019). Neglected sources of flexibility in psychological theories: From replicability to good explanations. *Computational Brain & Behavior*, 1–3.

Tomarken, A. J., & Waller, N. G. (2003). Potential problems with "well fitting" models. *Journal of Abnormal Psychology*, *112*(4), 578–598.

Tong, H., & Lim, K. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *42*(3), 245–268.

Tzelgov, J., & Henik, A. (1991). Suppression situations in psychological research: Definitions, implications, and applications. *Psychological Bulletin*, *109*(3), 524–536.

Uhler, C., Raskutti, G., Bühlmann, P., & Yu, B. (2013). Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 436–463.

van Borkulo, C., Boschloo, L., Borsboom, D., Penninx, B. W., Waldorp, L. J., & Schoevers, R. A. (2015). Association of symptom network structure with the course of depression. *JAMA Psychiatry*, *72*(12), 1219–1226.

Van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A., & Waldorp, L. J. (2014). A new method for constructing networks from binary data. *Scientific Reports*, *4*, 5918.

van de Leemput, I. A., Wichers, M., Cramer, A. O., Borsboom, D., Tuerlinckx, F., Kuppens, P., ... others (2014). Critical slowing down as early warning for the onset and termination of depression. *Proceedings of the National Academy of Sciences*, *111*(1), 87–92.

Van Der Maas, H. L., Dolan, C. V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychological Review*, *113*(4), 842–861.

VanderWeele, T. J. (2015). *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

VanderWeele, T. J., & Robins, J. M. (2007). Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect. *American Journal of Epidemiology*, *166*(9), 1096–1104.

VanderWeele, T. J., & Tchetgen Tchetgen, E. J. (2017). Mediation analysis with time varying exposures and mediators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*(3), 917–938.

van Elteren, C., & Quax, R. (2019). The dynamic importance of nodes is poorly predicted by static topological features. *arXiv preprint arXiv:1904.06654*.

van Montfort, K., Oud, J. H., & Voelkle, M. C. (2018). *Continuous time modeling in the behavioral and related sciences*. Cham: Springer.

Van Orden, K. A., Witte, T. K., Cukrowicz, K. C., Braithwaite, S. R., Selby, E. A., & Joiner Jr, T. E. (2010). The interpersonal theory of suicide. *Psychological Review*, *117*(2), 575.

van Rooijen, G., Isvoranu, A.-M., Meijer, C. J., van Borkulo, C. D., Ruhé, H. G., de Haan, L., et al. (2017). A symptom network structure of the psychosis spectrum. *Schizophrenia research*, *189*, 75–83.

Visser, I., & Speekenbrink, M. (2010). depmixS4: An R package for hidden markov models. *Journal of Statistical Software*, *36*(7), 1–21. Retrieved from `http://www.jstatsoft.org/v36/i07/`

Voelkle, M. C., & Oud, J. H. (2013). Continuous time modelling with individually varying time intervals for oscillating and non-oscillating processes. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 103–126.

Voelkle, M. C., Oud, J. H., Davidov, E., & Schmidt, P. (2012). An SEM approach to continuous time modeling of panel data: relating authoritarianism and anomia. *Psychological Methods*, *17*, 176-192.

Volterra, V. (1931). Variations and fluctuations of the number of individuals in animal species living together. In R. N. Chapman (Ed.), *Animal ecology* (pp. 409–448). New York, NY: McGraw-Hill.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638.

Walker, M. (2017). *Why we sleep: Unlocking the power of sleep and dreams*. New York: Simon and Schuster.

Walters, J. M., & Gardner, H. (1986). The theory of multiple intelligences: Some issues and answers. *Practical intelligence: Nature and origins of competence in the everyday world*, 163–182.

Warren, K. (2002). Thresholds and the abstinence violation effect: A nonlinear dynamical model of the behaviors of intellectually disabled sex offenders. *Journal of Interpersonal Violence*, *17*(11), 1198–1217.

Watkins, M. W., Lei, P.-W., & Canivez, G. L. (2007). Psychometric intelligence and achievement: A cross-lagged panel analysis. *Intelligence*, *35*, 59-68.

Wegener, A. (1966). *The origin of continents and oceans*. New York, NY: Dover Publications.

Wermuth, N., & Lauritzen, S. L. (1983). Graphical and recursive models for contingency tables. *Biometrika*, 537–552.

Wichers, M. (2014). The dynamic nature of depression: a new micro-level perspective of mental disorder that meets current challenges. *Psychological Medicine*, *44*(7), 1349–1360.

Wichers, M., Schreuder, M. J., Goekoop, R., & Groen, R. N. (2019). Can we predict the direction of sudden shifts in symptoms? transdiagnostic implications from a complex systems perspective on psychopathology. *Psychological medicine*, *49*(3), 380–387.

Wichers, M., Wigman, J., & Myin-Germeys, I. (2015). Micro-level affect dynamics in psychopathology viewed from complex dynamical system theory. *Emotion Review*, *7*(4), 362–367.

Wimsatt, W. C. (1987). False models as means to truer theories. *Neutral models in biology*, 23–55.

Wolfram Research, Inc. (2019). *Mathematica*, *Version 12.0.* (Champaign, IL)

Woodward, J. F. (2011). Data and phenomena: a restatement and defense. *Synthese*, *182*(1), 165–179.

Yarkoni, T. (2019, Nov). *The generalizability crisis.* PsyArXiv. Retrieved from psyarxiv.com/jqw35 doi: 10.31234/osf.io/jqw35

Nederlandse Samenvatting

Psychologische fenomenen zijn het best te begrijpen als complexe, dynamische processen. Dit perspectief heeft de laatste jaren enorm aan populariteit gewonnen, bijvoorbeeld in het onderzoek naar psychiatrische stoornissen. Onderzoekers in de klinische psychologie bestuderen deze processen momenteel meestal met simpele statistische modellen die gebruik maken van transversale (*cross-sectional*) data of intensieve longitudinale data. Deze modellen zijn slechts deels – of zelfs helemaal niet – in staat om belangrijke aspecten van de complexe dynamiek in kaart te brengen. Dit proefschrift onderzoekt hoe de kaders van *dynamische systeemtheorie* en *interventionele causale inferentie* gebruikt kunnen worden om deze huidige praktijk te verbeteren.

In hoofdstuk 2 wordt het nut van het gebruik van statistische netwerkmodellen om causale hypotheses te genereren onderzocht. Er wordt beargumenteerd dat het genereren van hypotheses enorm afhankelijk is van de specificaties van de 'beoogde causale structuur'. In dit hoofdstuk wordt aangetoond dat het genereren van hypotheses op basis van netwerkmodellen zelfs in de meest ideale situatie zeer lastig is. Om onderzoekers te helpen bij het exploratief genereren van hypotheses wordt hier ook een software applicatie geïntroduceerd. Deze applicatie wordt verder toegelicht met een empirisch voorbeeld. De discussie van dit hoofdstuk benadrukt de noodzaak om de rol van tijd in acht te nemen bij het leren over dynamische processen. Derhalve hebben de overige hoofdstukken voornamelijk een focus op statistische modellen voor intensieve longitudinale data.

Hoofdstuk 3 is gericht op de analyse van intensieve longitudinale data met differentiaalvergelijkingsmodellen. Deze *continuous-time* (CT) modellen vermijden de praktische en conceptuele beperkingen van gebruikelijke discrete tijdreeksmodellen met betrekking tot het tijdsinterval probleem. Dit hoofdstuk is toegespitst op een hele simpele differentiaalvergelijking, de CT versie van het VAR(1) model (een veelvoorkomend discreet tijdreeksmodel). Aan de hand van dit model wordt een brede introductie over de concepten van dynamische systeemtheorie gegeven, en de interpretatie van CT modellen wordt aan de hand van een empirisch voorbeeld uitgewerkt.

In hoofdstuk 4 wordt de CT benadering van dynamische netwerkanalyse geïntroduceerd, gebaseerd op het CT-VAR(1) model. Er wordt aangetoond dat huidige netwerkbenaderingen kunnen leiden tot misleidende conclusies met betrekking tot causale relaties tussen variabelen en incorrecte suggesties voor interventies, omdat deze afhankelijk zijn van het geanalyseerde tijdsinterval. Het CT-VAR(1) model heeft deze afhankelijkheid niet, doordat het een netwerk van moment-tot-moment relaties schat. Dit resulteert in betere eigenschappen met

betrekking tot conclusies over causale relaties. Hiernaast worden nieuwe *centrality* statistieken specifiek voor CT netwerken ontwikkeld. Geïnspireerd door de interventionele causale inferentie literatuur, stellen deze statistieken onderzoekers in staat om variabelen voor verschillende interventies (acuut of continu) optimaal te identificeren.

De focus van hoofdstuk 5 is bistabiele systemen, een specifiek type dynamisch systeem dat veel aandacht heeft gekregen in de psychologie literatuur. De capaciteit van verschillende gangbare statistische modellen om de eigenschappen van zulke systemen te achterhalen wordt onderzocht, waarbij wordt uitgegaan van twee scenario's: ideale data en meer realistische (met langere tijdsintervallen tussen metingen) data. Met de ideale data wordt aangetoond dat sommige statistische modellen gebruikt kunnen worden om de globale dynamiek (de meer stabiele relaties tussen variabelen) te achterhalen maar dat het moeilijk is om dit correct te doen voor de microdynamiek (moment-tot-moment relaties). Voor de realistische data geldt dat de globale dynamiek nog steeds gevonden kan worden maar dat dit helemaal niet mogelijk is voor microdynamiek. Deze resultaten benadrukken a) hoe moeilijk het is om inferenties maken op basis van statistische modellen zonder een sterke theorie, en b) de fundamentele rol van de frequentie van de dataverzameling bij het statistisch modelleren van tijdreeksen.

In hoofdstuk 6 wordt betoogd dat formele theorieën van kritiek belang zijn om onderzoek te doen naar psychiatrische stoornissen. Formele theorieën – geoperationaliseerd als een set differentiaalvergelijkingen – zijn gebruikelijk in disciplines die dynamische systeemtheorie toepassen, maar ontbreken bijna altijd in de klinische psychologie. Eerst wordt een kort overzicht gegeven van de wetenschapsfilosofische literatuur om het belang van formele theorieën te benadrukken. Daarna worden drie manieren onderzocht om formele theorieën te construeren op basis van statische modellen: a) het gebruik van statistische modellen als formele theorieën, b) het gebruik van statistische modellen om formele theorieën van af te leiden, en c) het gebruik van statistische modellen om bestaande formele theorieën te verbeteren. De derde benadering blijkt het meeste veelbelovende te zijn maar staat ook het verst af van de huidige praktijk. Tenslotte wordt een kader voorgesteld dat beschrijft hoe empirisch onderzoek het beste kan worden ingezet om formele theorieën over psychiatrische stoornissen te genereren, testen, en verbeteren.

Dit proefschrift eindigt met een duidelijke boodschap voor de klinische psychologie, de psychiatrie en de methodologen die binnen deze disciplines werken: als we hopen een beter begrip te krijgen over de complexe dynamische processen die ten grondslag liggen aan psychopathologie, moeten we de huidige onderzoekspraktijken met betrekking tot de ontwikkeling van formele theorie radicaal heroriënteren.

About the Author

Oisín Ryan was born on November 23rd 1991 in Kilkenny, Ireland. In 2013 he obtained his BSc. in Psychology from the University of Limerick with first-class honors. During his bachelor program he worked as a research assistant for Dr. Timothy D. Ritchie and visited Utrecht University in the Netherlands for six months as part of an Erasmus program. In September 2013 he returned to Utrecht University and enrolled in the research masters Methodology and Statistics for the Behavioral, Biomedical and Social Sciences, graduating cum laude in 2015.

Awarded a talent grant from the Netherlands Organization for Scientific Research (NWO), he began his PhD project in September 2015 under the supervision of Prof. dr. Ellen Hamaker at the Department of Methodology and Statistics in Utrecht. He has given a variety of invited presentations and workshops to different groups, and has presented his research at several international conferences, including the International Meeting of the Psychometrics Society (IMPS), the Association for Psychological Science (APS) annual convention, and the Conference on Complex Systems (CCS). In 2019 he spent one month as a visiting scholar with Dr. Donald J. Robinaugh at the Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston.

As of January 2020, Oisín holds a post-doctoral position at Utrecht University, allowing him to continue his research on dynamical systems modeling in clinical psychology and psychiatry.

Publications & Working Papers

Note: (*) denotes joint first authorship

Ryan, O. & Hamaker, E. L. (under review). Time to intervene: A Continuous-Time approach to network analysis and centrality.

Ryan, O., Bringmann, L. F., & Schuurman, N. K. (under review). The challenge of generating causal hypotheses using network models. Pre-print DOI: 10.31234/osf.io/ryg69

Haslbeck, J. M. B.*, **Ryan, O.***, Robinaugh, D.*, Waldorp, L. J., & Borsboom, D. (under review). Modeling psychopathology: From data models to formal theories. Pre-print DOI: 10.31234/osf.io/jgm7f

Haslbeck, J. M. B.* & **Ryan, O.*** (under review). Recovering within-person dynamics from psychological time series. Pre-print DOI: 10.31234/osf.io/dymhw

Dablander, F.*, **Ryan, O.***, & Haslbeck, J. M. B.* (under review). Choosing between AR(1) and VAR(1) models in typical psychological applications. Pre-print DOI: 10.31234/osf.io/qgewy

Haslbeck, J. M. B.*, **Ryan, O.***, & Dablander, F.* (under review). The sum of all fears: Comparing networks based on symptom sum-scores. Pre-print DOI: 10.31234/osf.io/3nxu9

Robinaugh, D. J., Haslbeck, J. M. B., **Ryan, O.**, Fried, E. I., & Waldorp, L. J. (under review) Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science*. Pre-print available from `https://psyarxiv.com/ugz7y`.

Groen, R. N., **Ryan, O**., Wigman, J. T. W., Riese, H., Penninx, B. W. J. H., Giltay, E. J., Wichers, M., & Hartman, C. A. (in press) Comorbidity between depression and anxiety: Assessing the role of bridge mental states in dynamic psychological networks. *BMC Medicine*.

Bastiaansen, J. A., Kunkels, Y. K., Blaauw, F., Boker, S. M., Ceulemans, E., Chen, M., Chow, S. M, de Jonge, P., Emerencia, A. C., Epskamp, S., Fisher, A. J., Hamaker, E.L., Kuppens, P., Lutz, W., Meyer, M. J., Moulder, R., Oravecz, Z., Riese, H., Rubel, J., **Ryan, O.**, Servaas, M. N., Sjobeck, G., Snipper, E., Trull,

T. J., Tschacher, W., van der Veen, D. C., Wichers, M., Wood, P. K., Woods, W. C., Wright, A. G. C., Albers, C. J. & Bringmann, L. F. (2020). Time to get personal? The impact of researchers' choices on the selection of treatment targets using the experience sampling methodology. *Journal of Psychosomatic Research*. https://doi.org/10.1016/j.jpsychores.2020.110211

Kuiper, R. M. & **Ryan, O.** (2019). Meta-analysis of lagged regression models: A continuous-time approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(3), 396–413. DOI: 10.1080/10705511.2019.1652613

Hamaker, E. L. & **Ryan, O.** (2019). A squared standard error is not a measure of individual differences. *Proceedings of the National Academy of Sciences*, *116*(14), 6544-6545. DOI: 10.1073/pnas.1818033116

Ryan, O., Kuiper, R. M., & Hamaker, E. L. (2018). A continuous-time approach to intensive longitudinal data: what, why and how? In K. v. Montfort, J. H. L. Oud, & M. C. Voelkle (Eds.), *Continuous time modeling in the behavioral and related sciences* (pp. 27–54). New York, NY: Springer.

Kuiper, R. M., & **Ryan, O.** (2018). Drawing conclusions from cross-lagged relationships: Re-considering the role of the time-interval. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(5), 809–823. DOI: 10.1080/10705511.2018.1431046

van de Schoot, R., Winter, S. D., **Ryan, O.**, Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217–239.

Hepper, E. G., Wildschut, T., Sedikides, C., Ritchie, T. D., Yung, Y. F., Hansen, N., ... **Ryan, O.,** & Stephan, E., & Vingerhoets, A . J. J. (2014). Pancultural nostalgia: prototypical conceptions across cultures. *Emotion*, *14*(4), 733–747.

Acknowledgements

Ellen. Since most people only ever do one PhD, I guess everyone feels that their experience was unique, whether it was or not. But I'm pretty sure mine was, and a large part of that was down to you. Thank you for sharing your knowledge, for your support, your openness, your honesty, your humor and your endless supply of (in)appropriate idioms. Thanks for giving me plenty of rope, and for stepping in every now and then to ensure I didn't hang myself with it. I'm not sure how this all would have turned out if I didn't have you to act as my mentor, confidant, "work mom", and critic. I'm glad we don't have to deal with that counterfactual.

Jonas. If I were to make a highlight reel of the last few years of my PhD, you would feature in a striking amount of it. In Utrecht and Amsterdam, Boston, Greece and Bordeaux; In our apartments, offices and Airbnbs, spas, bars, and too many restaurants. Our dinners, debates, discussions and arguments all helped keep me motivated, focused, and having fun. I'm proud of the work we did together, and I feel tremendously lucky to have you as a friend. My hope is to enjoy many more years of friendship and collaboration: May we never tire of ranting at each other, never agree on how to use semicolons, and never stop surreptitiously editing each other's punctuation.

Noémi and Laura, if I were to extend the uncomfortable "work family" metaphor, you two would be my work big-sisters. Laura, thanks for your willingness to discuss any topic at any time, and do it all with good humor. Noémi, thanks for your willingness to argue with me about any topic at any time. I learned a lot from both your empathy and your general intolerance of nonsense.

Don, the month I spent working with you and Jonas in Boston was probably the best experience I've had in academia. Thanks for putting up with us demanding your time, shouting so loudly that your boss installed a white noise machine outside your office, and for the marathon skype sessions that have followed every few weeks since we left the US. They have been a true inspiration.

Fabian, I'm very happy that we ended up working with each other so much in the last couple of years. I hope to write many more lengthy, informative, and ultimately unnecessary appendices with you. I consider myself a fan of your work, and I'm delighted that I can count you as a good friend too.

Thanks to everyone at the M&S department, especially Fayette, Thomas, Kees, Jolien, Anne, Erik-Jan and Ayoub, for all the lunches, coffees and walks around the uithof. All of these were essential distractions and stress-release valves which kept me sane in the office. Thanks especially to Erik-Jan and Ayoub in the last couple of years for sharing the office with me, feeding my caffeine addiction, dealing with my grumpy behaviour, and engaging me in general tomfoolery when it was needed most. Thanks to Rebecca for always being willing to help with the

most obscure mathematical questions that came to mind, and thanks to Gerko for all the random pieces of advice over the years. Thanks to Flip and Chantal for the cookies and impromptu Dutch lessons, and to Kevin for putting up with my constant requests for office equipment. Thanks to Irene Klugkist, who has somehow conspired to play an important role in many big moments in my life in recent years: You gave me my first job in the Netherlands, helping me be able to move here in the first place, you introduced me to Roline for the first time (although it was not the smoothest first meeting), and now you've been a member of my reading committee. I'm grateful for all of your help.

Thanks to everyone outside of work for distracting me, giving me many great memories and experiences, and making me feel at home in the past years. Thanks to Thomas, Millitza and Fayette for all the trips, the marathon board game sessions, and the willingness to engage in all sorts of random children's pastimes. Thanks to Joris for all the coffees, visits to the uithof sheep, and your boundless enthusiasm for new adventures. Thanks to EJ and Lara, Marloes and Vincent, for all the dinners, games, bottles of wine, late nights and shared recipes. Especially thanks to all the Kamphuises for making me feel part of the family.

Thanks to all my friends in Ireland, especially Ali, Eoin, Declan, Luke, the Steves, Kate, Niamh, Cillian, Úna, Zoe and Sally. Thanks for all the meet-ups, whatsapp calls, facebook messages and occasional reminders of the most embarrassing moments of my past. Thanks in particular to Luke for designing the cover of this book, and Ali for your never-ending enthusiasm for planning shared holidays and visits to London or Utrecht.

Thanks to all my family for putting up with my too-infrequent visits home and for ferrying me all over the place when I do arrive back. Thanks to my parents for all the support in moving and setting up a new life in the Netherlands; Colin, Pia and the kids for having myself and Roline stay for as long as we want every Christmas; thanks to Niamh for "checking in to see if I'm still alive" on a frequent basis, and thanks to Paudie for always reminding me how similar we are, and giving me the peace of mind that everything is taken care of back home.

Finally, thanks to Roline. I have threatened many ways of messing with you in my acknowledgements, all of which you have successfully vetoed: This paragraph is longer than one word, it won't feature a proposal of any kind, and there's no need to search for cryptic messages spelled out across the first letter of each preceding line (turns out that's much more difficult to pull off than you'd think). You were there for everything the last few years: The joy, the despair, frustrations and elations, the bouts of self-doubt, stress, mania and lethargy. You're in the top two people who suffered most in my bad moments during this PhD, and I'm genuinely not sure if I'm number one on that list or you. Through all that you built us an amazing life together in Utrecht. I promise not to do another PhD, I promise not to use the doctor title during petty disagreements, and I promise not to argue with you about what constitutes a "real job" anymore. I couldn't have done this without you. You're my rock, my best friend, and the love of my life. Thanks.