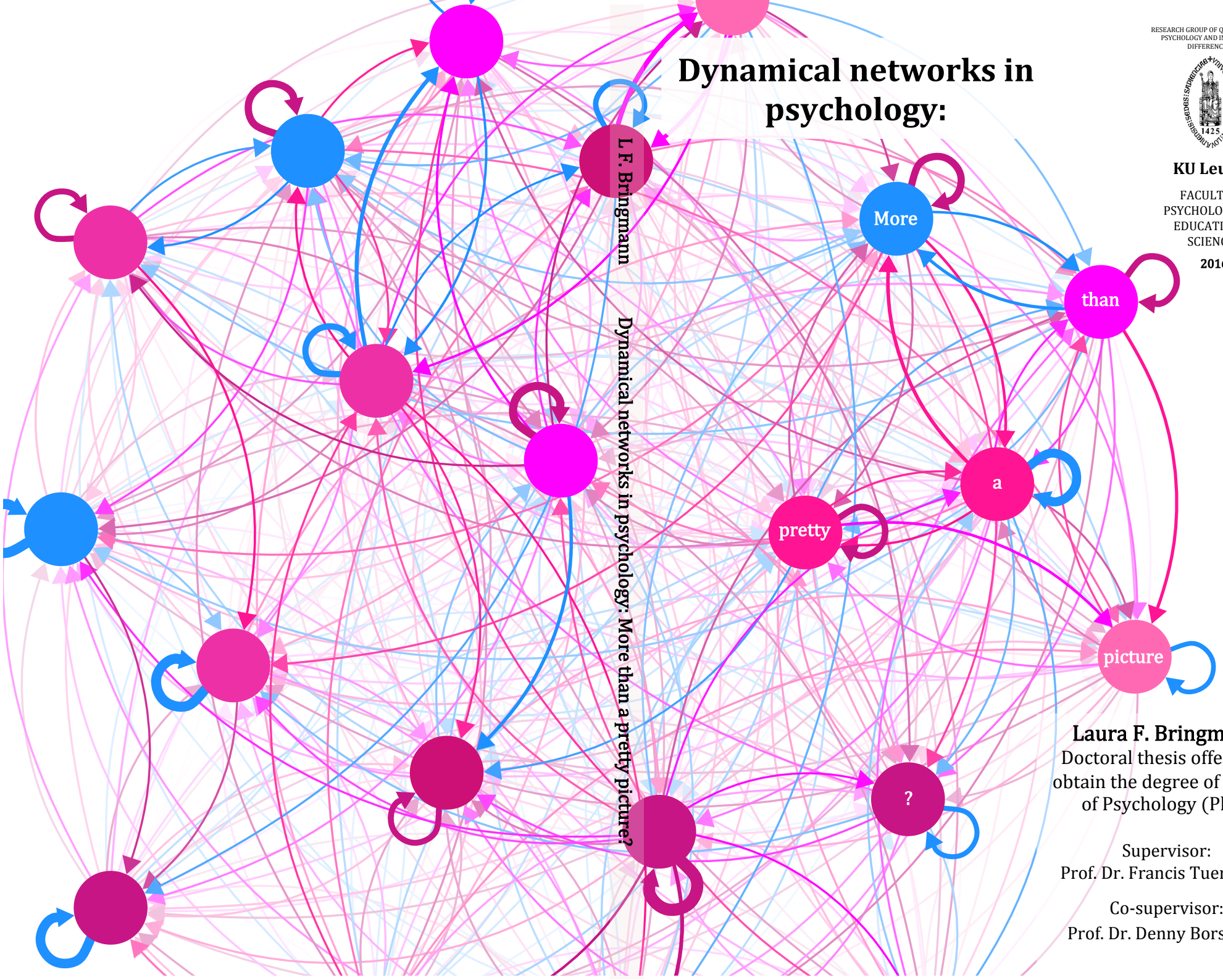
Dynamical networks in psychology: More than a pretty picture?

Thesis · October 2016

DOI: 10.13140/RG.2.2.28223.10404

1 author:

Laura Bringmann
University of Groningen
16 PUBLICATIONS **119** CITATIONS



Dynamical networks in psychology:

RESEARCH GROUP OF QUANTITATIVE PSYCHOLOGY AND INDIVIDUAL DIFFERENCES

KU Leuven

FACULTY OF PSYCHOLOGY AND EDUCATIONAL SCIENCES

2016

Laura F. Bringmann

Doctoral thesis offered to obtain the degree of Doctor of Psychology (PhD)

Supervisor:
Prof. Dr. Francis Tuerlinckx

Co-supervisor:
Prof. Dr. Denny Borsboom

KU LEUVEN

FACULTY OF PSYCHOLOGY AND
EDUCATIONAL SCIENCES

Research Group of Quantitative Psychology and
Individual Differences

Dynamical networks in psychology:

More than a pretty picture?

Laura F. Bringmann

Doctoral thesis offered to obtain the degree of
Doctor of Psychology (PhD)

Supervisor: Prof. Dr. Francis Tuerlinckx

Co-supervisor: Prof. Dr. Denny Borsboom

2016

Summary

In this thesis, we provide different perspectives on dynamical networks in psychology. The main technique used here to infer networks is the multilevel vector autoregressive (VAR) model. In a VAR model, the structure of the time-dependency within and between variables is explicitly modeled through a set of regression equations. Using a multilevel extension of a VAR model allows one to study the dynamics both within an individual as well as at group level.

The multilevel VAR model is further introduced in Chapter 2. In this study, longitudinal emotion data from individuals with residual depressive symptoms were examined. Besides visualization of the inferred networks, we also show how network structures can be further studied with network analyses, such as centrality techniques.

Chapter 3 focuses on individual networks estimated with a multilevel VAR model. In this chapter, the main goal is to study connectivity of individual emotion networks and their relation to neuroticism. The results suggest that individuals with high levels of neuroticism have a denser emotion network compared with their less neurotic peers.

In Chapter 4, we estimate the network of symptom dynamics that characterizes the Beck Depression Inventory-II (BDI-II), based on repeated administrations of the questionnaire to a group of depressed individuals who participated in a treatment study. Since the BDI-II symptoms decreased during treatment, the means changed, indicating changing dynamics. To account for this change in dynamics a linear trend was included in the multilevel VAR model. Beyond visualization, we conduct several network analyses, such as centrality and cluster analyses.

Chapter 5 lays the foundation for studying time-varying networks in psychology. Networks are likely to change over time, due to for example therapy (see Chapter 4). Up until now there has been no easy way to detect changing dynamics. With a time varying autoregressive (TV-AR) model, changes in means and temporal dynamics can be easily identified and modeled, and therefore the model has significant potential for studying changing dynamics in psychology.

Chapter 6 concerns psychological networks based on fMRI data. We use a new data driven technique, ancestral graphs (AGs), and compare it with a standard hypothesis driven method, Structural Equation Modeling (SEM). In contrast to VAR models, network analysis in both SEM and AG is based on the replication of the condition-specific trials and not on time-dependencies in time series data. As AGs can test explicitly the assumption of missing regions (nodes) in the network, it leads in general to more accurate network structures than the SEM method. Although currently mainly used in fMRI research, AGs could also be a promising solution for estimating networks in other fields of psychology, such as emotion research.

In Chapter 7, a more general theoretical perspective on psychological science is taken. Network techniques are highly interdisciplinary and analyses done in physics seem to translate to other fields, such as social or psychological science. Still, in measurement debates, physical measurement is seen as largely disconnected from psychological measurement. We argue instead that there are interesting parallels and connections between the two. In the last chapter, the discussion, a critical examination of the general topic of the thesis is presented, ultimately answering the question: Dynamical networks in psychology – more than a pretty picture?

Samenvatting

In deze thesis werpen we vanuit verschillende perspectieven een blik op dynamische netwerken in de psychologie. De techniek die in deze thesis het meest wordt gebruikt om netwerken te schatten is het multivariate autoregressieve (VAR) model. In een VAR model wordt de tijdsafhankelijkheid binnen en tussen variabelen expliciet gemodelleerd met behulp van een set van regressievergelijkingen. Door middel van een multilevel extensie van het VAR model kan niet alleen de intra-individuele dynamiek bestudeerd worden, maar ook de dynamiek op groepslevel.

Het multilevel VAR model wordt verder geïntroduceerd in hoofdstuk 2. Hier analyseren we met behulp van het multilevel VAR model longitudinale emotiedata van personen met residuele depressieve klachten. Naast het visualiseren van het verkregen netwerk worden netwerkanalyses zoals centraliteitsanalyses uitgevoerd.

In hoofdstuk 3 ligt de focus op individuele netwerken die ook weer verkregen zijn met een multilevel VAR model. Het hoofddoel van dit hoofdstuk is het bestuderen van connectiviteit van individuele emotienetwerken en de relatie tot neuroticisme: personen die hoog scoren op neuroticisme hebben een dichter verbonden emotienetwerk in vergelijking met personen die lager scoren op neuroticisme.

In hoofdstuk 4 schatten we een netwerk dat de dynamiek van de symptomen van de Beck Depression Inventory-II (BDI-II) representeert. Data van de BDI-II symptomen werden verkregen in een longitudinale studie waar personen met een depressie tijdens alle therapiesessies die ze ondergingen deze vragenlijst invulden. Omdat de symptomen afnamen tijdens de behandeling, veranderde de gemiddelde symptoomscore en daarmee de dynamiek. Voor deze verandering in dynamiek is gecorrigeerd door een lineaire trend op te nemen in het multilevel VAR model. Naast netwerkvisualisatie zijn er verscheidene netwerkanalyses uitgevoerd zoals centraliteits- en clusteranalyses.

Hoofdstuk 5 effent de weg voor het schatten van psychologische netwerken die veranderen gedurende de tijd. Het is aannemelijk dat netwerken gedurende de tijd veranderen als gevolg van bijvoorbeeld therapie (zoals in hoofdstuk 4). In dit hoofdstuk stellen we een eenvoudige manier om veranderende dynamiek te detecteren voor: het tijds-variërende autoregressieve (TV-AR) model. Met dit model kunnen veranderingen in gemiddeldes en temporele dynamiek gemakkelijk worden geïdentificeerd en gemodelleerd.

Hoofdstuk 6 betreft psychologische netwerken gebaseerd op fMRI data. We maken hier gebruik van een nieuwe datagedreven techniek, ancestral graphs (AGs), en vergelijken deze met een standaard hypothesegedreven methode, Structural Equation Modelling (SEM). In tegenstelling tot VAR modellen zijn netwerkanalyses in zowel AG als SEM gebaseerd op de replicatie van conditiespecifieke tests en niet op de tijdsafhankelijkheid in de tijdreeksdata. Omdat met AGs expliciet getest kan worden of er gebieden in het netwerk missen, leidt deze techniek over het algemeen tot meer accurate netwerk structuren dan de SEM methode. Hoewel deze methode momenteel vooral in fMRI onderzoek gebruikt wordt, kan de AG techniek ook nuttig zijn voor het schatten van netwerken in ander psychologisch onderzoek.

In hoofdstuk 7 bekijken we psychologisch onderzoek in de brede theoretische zin. Het veld waarin netwerken worden onderzocht is zeer interdisciplinair. Analyses die uitgevoerd worden om variabelen te meten in de fysica worden ook gebruikt in psychologisch onderzoek. Echter, in het theoretisch debat rond meetproblemen worden meetmethoden in de fysica meestal gescheiden bestudeerd van die in de psychologie. Wij laten in dit hoofdstuk zien dat er interessante parallellen en connecties tussen de twee onderzoeksgebieden zijn. In het laatste hoofdstuk, de discussie, worden de aangestipte onderwerpen van deze thesis kritisch onderzocht, uiteindelijk leidend tot het beantwoorden van de hoofdvraag: Dynamische netwerken in de psychologie – meer dan een mooi plaatje?

To my best friend, my favourite co-author and love of my life:
Markus I. Eronen

Contents

1 Introduction

1.1 Networks everywhere

Take a piece of paper. Draw a few points and some lines between these points. Congratulations, you have drawn yourself a network. More formally, networks can be seen as simplified representations capturing how elements in a system are interconnected. So in essence, everything that can be represented as dots (i.e., nodes or vertices) with lines (i.e., edges, ties or links) between the dots amounts to a network. Once you have taken a network perspective you will find networks everywhere. Well known networks are the *internet* and the *World Wide Web* (Newman, 2010). Whereas the internet has a clear physical structure (computers are linked by physical cables), the Web is a more abstract network with webpages being the nodes and the *hyperlinks* on the webpages, linking the webpage to other pages, the edges (van Steen, 2010). Besides information networks, an essential and biological network is our *brain* where white matter tracts, or bundles of axons, connect pairs of brain regions (Rubinov & Sporns, 2010). Finally, one of the oldest fields in network research is social networks, in which every individual is a node, and links between individuals are determined by, for example, their friendship or co-authorship.

As networks are so broadly defined, it should not come as a surprise that the research field of networks is very interdisciplinary, crossing all disciplines of science. However, what is striking is that networks of different scientific fields have similar properties, which are often described and analyzed using *graph theory*, the branch of mathematics that studies networks (van Steen, 2010). An example of such a network property is the phenomenon of *hubs* (Newman, 2010). Studying hubs in a network is essentially asking the question about the importance of a node in a network. Importance or centrality can be measured in several ways, of which degree centrality is the most common one. The degree of a node is calculated by counting the number of edges attached to it, for example by counting the number of friends an individual has in a friendship network. In case of directed networks (such as the World Wide Web), degree can be split up into incoming edges (in-degree), the information a node receives, and outgoing edges (out-degree), the information a node sends out to other nodes in the network. Hubs then are nodes with an unusually high centrality degree and thus are seen as very central or important in the network (Newman, 2010). In social networks, for example, there are often

only a few central individuals that have a lot of acquaintances, in the World Wide Web network there are a few pages that are linked by an exceptionally large number of other pages, and in the brain there are a few brain regions that are involved in a large number of brain processes (Newman, 2010; Sporns, 2011). Another example of a network property is the *small world effect* (Watts & Strogatz, 1998). This is the phenomenon that even when a lot of nodes are not directly connected, most nodes can be reached in just a few steps. Thus, in the context of social networks: strangers can become acquaintances by just a few steps in a friendship network (Newman, 2010; Travers & Milgram, 1969).

Importantly, through revealing the structure and properties of real-world networks, new insights can evolve that have real life consequences. A lively example is criminal networks. To fight crime, a hierarchical paradigm has long been prominent (Klerks, 2001). This was under the assumption that organized crime has a pyramid structure, and thus targeting the leader of the criminal organization would result in a disruption of the whole (criminal) pyramid. Taking a network perspective by focusing on the links between criminals, however, turned out to be far more fruitful to disrupt criminal networks than traditional law enforcement (Borgatti, Mehra, Brass, & Labianca, 2009; Duijn, Kashirin, & Sloot, 2014).

1.2 Networks in psychology

Networks have not been an unfamiliar topic in psychological research. Techniques like neural networks have been used in neuropsychological studies and social networks have been present in social psychology for decades (see for example; Bronfenbrenner, 1986; Mason, Conrey, & Smith, 2007; Posner & Rothbart, 2007; Rubinov & Sporns, 2010). Only recently, however, the network approach has found its way to psychopathology, emotion research and personality research (Borsboom, Cramer, Schmittmann, Epskamp, & Waldorp, 2011; Bringmann, Vissers, et al., 2013; Costantini et al., 2015).

The foundations for this recent network perspective in psychology can be found in the conceptualization of psychological disorders. It is a well-observed fact that specific symptoms often co-occur and are thus highly inter-correlated. For example, depressive symptoms are commonly more highly correlated with each other than with symptoms of schizophrenia (Cramer, Borsboom, Aggen, & Kendler, 2012). The question then is why specific symptoms tend to co-occur. In the literature, the traditional answer has been the common cause or latent variable approach. According to this framework, symptoms of, for example, depression hang strongly together because they have a common underlying cause, the disorder depression. Thus, the unobserved latent variable, the disorder depression, causes observable symptoms such as *loss of interest*, *depressed mood* and *suicidal ideation*. A further implication of this approach is that symptoms are *only* related because they share the same underlying cause or factor and thus are mere indicators of the disorder (i.e, depression; Bollen & Lennox, 1991; Cramer,

Waldorp, van der Maas, & Borsboom, 2010; Reise & Waller, 2009; Schmittmann et al., 2013).

The network perspective, however, gives a new answer to the co-occurrence of symptoms. Taking a network approach, symptoms do not hang together because they are caused by the same latent variable or disorder, but because symptoms are (part of) the disorder (Cramer et al., 2010). Thus, there is a mereological relation between symptoms and disorders, which makes it unlikely that the disorder and symptoms can be separated from each other as is possible with many medical diseases (Borsboom & Cramer, 2013). You can, for example, have human immunodeficiency virus (HIV) without having symptoms (Paltiel et al., 2005), but it is implausible that you can diagnose somebody with depression without this person having any depression symptoms. Following this reasoning, the latent variable or common cause approach might be plausible for medical diseases, as a medical condition such as HIV causes symptoms such as fatigue and fever, but seems illogical in explaining the covariance between symptoms of depression. Instead, the interrelatedness of symptoms occurs simply because symptoms are directly influencing each other: if I am having sleeping problems I am also more likely to experience loss of interest and sadness, finally resulting in a full-blown depressive disorder. Thus, disorders are networks of symptoms that directly (causally) influence each other, leading to the co-occurrence of symptoms. Consequently, if we want to get a better understanding of psychological disorders, we should refocus on the relation between the symptoms by elucidating the patterns of interactions among symptoms, in other words: we need a network approach (Fried, Nesse, Zivin, Guille, & Sen, 2014; Fried, Epskamp, Nesse, Tuerlinckx, & Borsboom, 2016). This conceptualization can then be generalized to other psychological phenomena, resulting in networks of, for instance, emotions and personality traits (Cramer, van der Sluis, et al., 2012a; Pe et al., 2015).

1.3 Constructing networks in psychology

Once you decide what the nodes in your network should represent (e.g., symptoms), there are several ways to construct edges between the nodes. One way, resembling the construction of social networks, is to simply ask individuals or clients about the causal relationships between their symptoms (Frewen, Allen, Lanius, & Neufeld, 2012; Frewen, Schmittmann, Bringmann, & Borsboom, 2013). However, just as individuals have difficulties in describing correctly their friendship network (Newman, Barabási, & Watts, 2006, p. 12), also this kind of *perceived causal relation networks* are likely to be prone to error. Still, such networks could be potentially useful as a starting point for therapy.

Alternatively, edges can be derived indirectly through the association between items of a questionnaire (e.g., the Beck Depression Inventory) answered by one or multiple individuals at one (cross-sectional) or several (longitudinal) time points. The edges in this case represent, for instance, the correlation (*association networks*) or partial-correlation (*concentration networks*) between nodes or

items of symptoms. Especially constructing concentration networks has been a very popular approach of estimating networks in psychology, mostly because only cross-sectional data is required (see, for example, Boschloo et al., 2015; Cramer, Borsboom, et al., 2012; Costantini et al., 2015; Fried et al., 2015; McNally et al., 2015; Rhemtulla et al., 2016; Robinaugh, LeBlanc, Vuletich, & McNally, 2014; van Borkulo et al., 2015).

Although widely used, psychological networks based on cross-sectional data are not unproblematic (Bos & Wanders, 2016; Borsboom, Kievit, Cervone, & Hood, 2009; Hamaker, 2012; Kievit, Frankenhuis, Waldorp, & Borsboom, 2013; Molenaar, 2004; Robinson, 1950). The problem is that results found at the population level do not automatically generalize to the person level (Hamaker, 2012). This is a general problem, and thus not specific to network studies. Cross-sectional data, as it has only one time point, cannot tell us about the processes happening within an individual, but only across individuals. However, it is often wrongly assumed that patterns (such as how symptoms co-occur) found through cross-sectional data are descriptive of individual processes (Kievit et al., 2013).

Take for example the relationship between migraine and sleep. If we did a cross-sectional study, we might very well find that there is a positive relationship between sleep and migraine (Schmitz & Skinner, 1993). In this case, we could wrongly conclude that individuals who sleep more are more likely to have migraines. However, the relationship found at the group level simply comes about because in general individuals who often have migraine attacks tend to sleep more in order to prevent them. In contrast, *within* each individual of the group we would actually find a negative relationship; sleeping less is related to more migraines. In fact, in order to be able to generalize from the group to the individual level the process under study must be *ergodic* (Molenaar & Campbell, 2009). Ergodicity is a mathematical notion that in the context of psychology implies that all within-person moments (e.g., mean and variance) of a process are the same across individuals and over time (Molenaar, 2015). For example, all subjects should have the same mean and co-variance between variables over time. Thus, ergodicity is a very strong requirement and unlikely to hold in most if not all psychological datasets (Hamaker, 2012).

This limitation of cross-sectional data has been widely acknowledged in psychological research and calls have been made for more and longer intensive longitudinal data for deriving networks (Borsboom & Cramer, 2013; Fried et al., 2015; van Borkulo, Borsboom, & Schoevers, 2016). Fortunately, in the past three decades, we have witnessed a spectacular growth of intensive longitudinal data (aan het Rot, Hogenelst, & Schoevers, 2012; Bolger & Laurenceau, 2013; Trull & Ebner-Priemer, 2013). Recently, for instance, an experience sampling study within a single individual resulted in over 1400 measurements of his momentary states, such as mood (Wichers, Groot, & Psychosystems, 2016). This inflation of intensive longitudinal studies is partly a result of an increasing recognition of the necessity of these kinds of data for the study of individual processes, and partly a result of expanding technological

possibilities of gathering data through mobile devices like smart-phones. With this development, networks representing individual psychological processes have become within reach.

1.4 Temporal dynamical networks

The availability of longitudinal studies makes it possible to study all kinds of inter- and intra-individual differences. For instance, in developmental studies longitudinal data is often used to study change in the mean of a process, such as the increase of short-term memory capacity from childhood to adulthood. Although studying these kinds of gross underlying trends is without a doubt a crucial part of psychological research, the network approach takes a different perspective. Instead of studying just the mean levels of, for example, symptoms, this approach highlights the interaction between variables over time, the temporal dynamics. When studying temporal dynamics, the focus is on how variables within individuals influence themselves or each other over time, resulting in temporal dynamical networks.

Vector autoregressive models

Many approaches for studying temporal dynamics are available. What most of these approaches have in common is that they are based on a form of vector autoregressive (VAR) modeling, a family of statistical techniques in which the structure of the time-dependency within and between variables is explicitly modeled through a set of regression equations. Although most VAR techniques model time as a discrete process, there are also alternatives modeling psychological processes as evolving continuously over time (Oravecz, Tuerlinckx, & Vandekerckhove, 2011; Voelkle & Oud, 2013). In addition, VAR techniques can be applied in various frameworks, such as such as the Bayesian (e.g., Pole, West, & Harrison, 1994; Schuurman, Grasman, & Hamaker, in press) and the structural equation framework (e.g., Hamaker, Dolan, & Molenaar, 2003; Voelkle, Oud, Davidov, & Schmidt, 2012). As the basic VAR model is not explained elsewhere in the thesis, I will discuss this model below in more detail. After that, the two most important extensions considered in the upcoming chapters, a multilevel VAR and a time-varying VAR (TV-VAR) model, are shortly explained.

The standard discrete VAR model is a multivariate regression model and has as input time series data of only one individual (or dyad). Consider for example the smallest network possible, a bivariate VAR model with lag 1:

$$y_{1,t} = \beta_{10} + \beta_{11} y_{1,t-1} + \beta_{12} y_{2,t-1} + \varepsilon_{1,t} \tag{1.1}$$

$$y_{2,t} = \beta_{20} + \beta_{21} y_{1,t-1} + \beta_{22} y_{2,t-1} + \varepsilon_{2,t}. \tag{1.2}$$

In a VAR(1) model there are y_{it} variables (nodes of the network), where $i = 1, 2, \ldots, m$ is the number of variables (in this case $m = 2$) and t is the time index (Brandt & Williams, 2007). Each dependent variable $(y_{1,t}, y_{2,t})$ is regressed on its lagged values ($y_{1,t-1}, y_{2,t-1}$ respectively) through the autoregressive parameters (β_{11} and β_{22}). These parameters capture the strength and direction of the autoregressive effect a variable has on itself from one time point to the next and are also known as the self-loops in the network. As an example, consider Figure 1.1. This is a hypothetical example of positive and negative affect measured daily in a single individual over 100 days. Autoregressive effects, the green solid self-loops in this figure, indicate to what extent each variable is predictive of itself over time. A positive autoregressive effect indicates that current levels of, for instance, NA predict NA levels at the next time point, such as the next day. In addition, a positive autoregressive effect indicates that the process is not very prone to change, such that its values across time will only slowly go back to baseline values (Hamaker & Dolan, 2009). A negative autoregressive effect, on the other hand, indicates a jigsaw pattern in the sense that it predicts fast changing process. That is, high values at a given time point predict low values of NA at the next time point and vice versa.

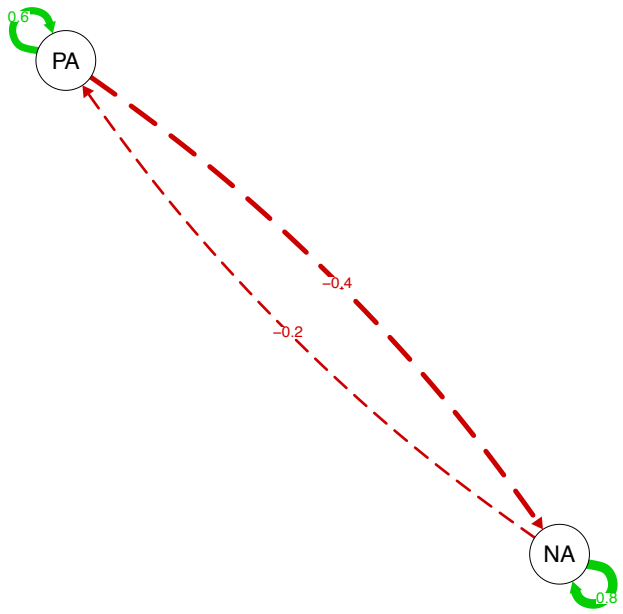


Figure 1.1: *Simulated time series for valence process (Positive Affect (PA) and Negative Affect (NA)) of a single individual.* The numbers indicate the values of the self-loops (β_{11} and β_{22}) and cross-regressive edges (β_{12} and β_{21}).

Additionally, each dependent variable $(y_{1,t}, y_{2,t})$ is regressed on the lagged values of each of the other dependent variables ($y_{2,t-1}, y_{1,t-1}$ respectively) through the cross-regressive (i.e. cross-lagged) parameters β_{12} and β_{21}. These edges are represented by the dashed red lines in Figure 1.1. Cross-regressive effects indicate the direction and strength of the effect a variable has on other variables

from one time point to the next. Considering again the example of NA and PA, NA experienced at one time is likely to be predicted by not only NA at the previous time point (autoregressive effect), but also by PA (cross-regressive effect). For example, if there is a negative cross-lagged effect from NA to PA, when the individual under study experiences an increase in her NA at one time point, she is likely to experience decreased (opposite) PA values at the next time point, whereas a positive cross-lagged effect from NA to PA indicates that if she has a high NA at one time point she is also likely to experience an increase in her PA values at the next time point.

On the other hand, auto- and cross-regressive coefficients that are close to zero indicate that there is no predictive value within or between the variables. In such a case, for example, an individual's NA could not be predicted by her NA itself nor by her PA.

In the VAR model, the β_{i0} denote the intercepts. As networks represent the interaction between and within variables, the intercepts are not included in the network. The innovation terms $\varepsilon_{1,t}$ and $\varepsilon_{2,t}$ (also known as residuals, perturbations, or random shocks) are the part of the current observations $y_{1,t}$ and $y_{2,t}$ that cannot be explained by the previous observations $(y_{1,t-1}, y_{2,t-1})$. The innovations are assumed to follow a white noise process, meaning that all innovation processes have a mean zero and a time invariant covariance matrix. Although serial correlation in the innovation structure is not allowed, innovations are allowed to correlate across equations. Note that equations 1.1 and 1.2 do not have to be estimated simultaneously to obtain correct estimates, but can be estimated with equation-by-equation ordinary least squares (Brandt & Williams, 2007, p.24).

The VAR model specified above can also be rewritten in a more general vector form:

$$\mathbf{y}_t = \boldsymbol{\beta}_0 + \boldsymbol{B}_1 \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t, \tag{1.3}$$

with

$$\mathbf{y}_t = \begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix}, \boldsymbol{\beta}_0 = \begin{bmatrix} \beta_{10} \\ \beta_{20} \end{bmatrix}, \boldsymbol{B}_1 = \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix} \text{ and } \boldsymbol{\varepsilon}_t = \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}.$$

Multilevel VAR models

To estimate a VAR based network, many time points (at least over 50) need to be available. Additionally, a VAR model allows one to study only the intra-individual dynamics, whereas in order to generalize results, inter-individual differences are also of importance.

Therefore, to infer dynamical networks, it is fruitful to combine the standard VAR model with a multilevel model. In a multilevel VAR model, temporal dynamics can be modeled not only within an individual, but also at group level, estimating both average or population and individual networks. In

the multilevel approach, individuals are assumed to come from a population, most often a multivariate normal distribution (the average of such a population being the fixed effect), but at the same time the individuals are allowed to differ from one another (such deviations from the population mean are the random effects). By using a multilevel VAR approach to estimate networks, having less time points per individual can be compensated by having more individuals, and in this way reliable statistical inferences can be made at both the population and individual level; this is known as *borrowing strength*.

Changing networks: TV-VAR models

Although models based on VAR or multilevel VAR techniques are often used for inferring dynamical networks, these models are limited by assuming dynamics to be invariant over time. This is because one main assumption of both VAR and its extension, multilevel VAR, is stationarity. In general terms, stationarity means that the statistical properties of the data under study do not change over time, and thus dynamical features such as the interaction between variables of a network are assumed to be invariant over time (Chatfield, 2003). However, like biological and sociological network structures (Ahmed & Xing, 2009; Newman et al., 2006; Rosvall & Bergstrom, 2010), networks in psychology are likely to evolve and thus change over time. Imagine, for example, a network of symptoms of a depressed individual. This network might consist of strongly interrelated symptoms, meaning that as one symptom gets activated other symptoms are also prone to be activated. In order to disrupt this unwholesome pattern, the individual undergoes therapy. In this case the network is not only expected to change, but change is in fact the purpose of therapy. Thus, instead of a single network, a network film is needed, capturing the changing network dynamics over time. This can be accomplished by using a time-varying VAR (TV-VAR). The defining feature of a TV-VAR model is that the coefficients of a VAR model, and thus the network structure, are now allowed to vary over time, following an unspecified function of time (Dahlhaus, 1997).

1.5 Outline of this thesis

In the upcoming chapters we will provide different perspectives on networks in psychology. Chapters 2 to 4 all deal with networks based on multilevel VAR models. In Chapter 2, the multilevel VAR model for inferring dynamical psychological networks is introduced. In this study, longitudinal emotion data from individuals with residual depressive symptoms were examined. We show how average, individual and inter-individual networks can be constructed and visualized. In addition to the visualization, we also show how the inferred network structures can be further analyzed using network analyses, such as centrality techniques. Furthermore, a validation data set was used, leading to a highly similar network structure as in the original data.

Chapter 3 focuses on individual networks, estimated with a multilevel VAR model. In this chapter, the main goal is to study connectivity of individual emotion networks and their relation to neuroticism. Connectivity or density analysis is a network analysis that indicates how strongly the network is interconnected. The denser a network is, the more strongly variables interact. The results suggest that individuals with high levels of neuroticism had a denser emotion network compared with their less neurotic peers. This effect is especially pronounced for negative emotion networks. Results were validated in a second dataset.

When using a multilevel VAR model, the number of variables that can be included in the network is limited. Therefore, in Chapter 4 we use a moving window approach in order to allow more variables in the network. In this study, we estimate the network of symptom dynamics that characterizes the Beck Depression Inventory-II (BDI-II; containing 21 symptoms), based on repeated administrations of the questionnaire to a group of depressed individuals who participated in a treatment study of an average of 14 weekly assessments. The focus is on the average group effects and not on individual effects, as simulation results indicated that, using a moving window approach, the links of the network at population level can be estimated well, but not the variance components or individual links. Since the BDI-II symptoms decreased during treatment, the means changed, indicating non-stationarity. For this reason, a linear trend was included in the multilevel VAR model. Beyond visualization, we conducted several network analyses, such as centrality and cluster analyses. Results indicated that of the 21 items, the symptom *loss of pleasure* was the most central item in the network. Cluster analyses suggested that the dynamic structure of the BDI-II involves two clusters, which is consistent with earlier psychometric analyses.

Chapter 5 lays the foundation for studying time-varying networks in psychology. Networks are likely to change over time, due to for example therapy (see chapter 4). However, up until now there has been no easy way to detect nonstationarity due to trends and changes in the interactions between variables simultaneously. With a TV-AR model (which is easily extended to a TV-VAR model), changes in means and temporal dynamics can be easily identified and modeled. Notably, no prior knowledge of the processes that drive change in the dynamic structure is necessary. Thus, in this chapter we hope to show that the TV-(V)AR model has significant potential for studying changing dynamics and thus networks in psychology.

In Chapter 6, brain networks in psychology are discussed. The nodes in this case are brain regions of interest (ROIs) involved in motion perception. Not only the kind of data (fMRI data) used here differs from the previous chapters, but also the method to infer networks. We used a new data driven technique, ancestral graphs (AGs), and compared it with a standard hypothesis driven method, Structural Equation Modeling (SEM). In contrast to VAR based models, network analysis in both SEM and AG is based on the replication of the condition-specific trials and not on the time-

dependencies of variables in the time series. As AG can test explicitly the assumption of missing regions (nodes) in the network, it leads in general to more accurate network structures than the SEM method. Although currently mainly used in fMRI research, AG could also be a promising solution in other fields of psychological networks as it is very likely that not all relevant variables in a network have been taken into account, which can lead to spurious relationships in networks when not modeled explicitly.

In Chapter 7, a more general theoretical perspective on psychological science is taken. As is underpinned in network research, network analyses are highly interdisciplinary, and analyses done in physics seem to translate to other fields, such as social or psychological science. Still, in measurement debates, physical measurement is seen as largely disconnected from psychological measurement. We argue instead that there are interesting parallels and connections between the two. More specifically, our novel approach is to study the issue of validity based on the history of measurement in physics, which results in concrete points that are relevant for the validity debate in psychology and thus also for network research. For example, psychologists would benefit from focusing more on the robustness of measurements. Robustness refers here to the idea that if there are several independent ways of measuring something, this increases our confidence in the measurements. This general point can also be seen in practice in the rest of the thesis. We try, for example, to get more confidence in our, still very new, network methods by performing analyses in several datasets (Chapter 2 and 3) or use independent ways to measure network connections and see if result converge (Chapter 4).

In the last chapter, the discussion, a critical examination of the thesis is presented, ultimately answering the question: Dynamical networks in psychology – more than a pretty picture?

As chapters 2 to 7 each contain a full published article, there may be some overlap between the chapters. Data and code of chapters 2 to 5 can be found online on the homepage of the journal where the article was published. Chapters 2 to 7 have been published in the following journals:

Chapter 2: Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., Borsboom, D., & Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLoS ONE, 8*, e60188, 1-13.

Chapter 3: Bringmann, L. F., Pe, M. L., Vissers, N., Ceulemans, E., Borsboom, D., Vanpaemel, W., Tuerlinckx, F., & Kuppens, P. (in press). Assessing temporal emotion dynamics using networks. *Assessment*.

Chapter 4: Bringmann, L. F., Lemmens, L. H. J. M., Huibers, M. J. H., Borsboom, D., & Tuerlinckx, F. (2015). Revealing the dynamic network structure of the Beck Depression Inventory-II. *Psychological Medicine, 45*, 747-757.

Chapter 5: Bringmann, L. F., Hamaker, E. L., Vigo, D. E., Aubert, A., Borsboom, D., & Tuerlinckx, F. (in press). Changing dynamics: Time-varying autoregressive models using generalized additive modeling. *Psychological Methods*.

Chapter 6: Bringmann, L. F., Scholte, H. S., & Waldorp, L. J. (2013). Matching structural, effective, and functional connectivity: A comparison between structural equation modeling and ancestral graphs. *Brain connectivity, 3*, 375-385.

Chapter 7: Bringmann, L. F., & Eronen, M. I. (2016).[1] Heating up the measurement debate: What psychologists can learn from the history of physics. *Theory & Psychology, 26*, 27-43.

[1]Both authors contributed equally to this article

2 A network approach to psychopathology: New insights into longitudinal data

Theoretical considerations and empirical evidence in psychology point towards a network perspective, in which psychological constructs are conceptualized as networks of interacting components instead of measurements of a latent construct, as hypothesized in traditional perspectives (Borsboom, 2008; Borsboom et al., 2011; Cramer et al., 2010; Cramer, Borsboom, et al., 2012; Kendler, 2012; Schmittmann et al., 2013). From this perspective, mental disorders are understood as networks of interacting symptoms (Cramer et al., 2010) that form mechanistic property clusters (Kendler, Zachar, & Craver, 2011): sets of causally intertwined properties that need not share one fundamental underlying cause. By focusing on the interaction between symptoms, the network approach naturally captures the fact that symptoms of psychopathology co-evolve dynamically (Ebner-Priemer, Eid, Kleindienst, Stabenow, & Trull, 2009): if one symptom arises (e.g., insomnia), that symptom can cause other symptoms to arise as well (e.g., concentration problems; Cramer et al., 2010).

Such patterns of symptom interaction are likely to vary across individuals. For instance, some people have a higher degree of emotional variability than others, and such differences are known to be related to personality traits, such as neuroticism (Kuppens, Oravecz, & Tuerlinckx, 2010). Likewise, some people may feature stronger connections between sleep deprivation and affect, such that a night of bad sleep quickly leads to depressed mood, whereas others may be more resilient (see e.g., Meney, Waterhouse, Atkinson, Reilly, & Davenne, 1998). By focusing on patterns of symptom dynamics, the network approach may potentially yield important insights into how the dynamics of psychopathology relate to intra- and inter-individual differences. Despite the fact that the network perspective is highly suggestive in this respect, techniques to actually empirically chart differences in the dynamical structure of individuals' symptom dynamics have so far been lacking. In this paper, we present a methodology suited for this task and we apply this methodology to data of individuals with residual depressive symptoms (Geschwind, Peeters, Drukker, Van Os, & Wichers, 2011) to illustrate its potential use in psychopathology research.

The natural starting point for the study of symptom network dynamics lies in the analysis of symptoms measured over different time points. Such time series data have recently become available

due to the rising popularity of data collection approaches using the Experience Sampling Method (ESM), where data about the experiences and affect of participants in their daily life are collected repeatedly over time (Bolger, Davis, & Rafaeli, 2003; Czsikszentmihalyi & Larson, 1987; Stone & Shiffman, 1994). However current statistical tools for inferring networks from empirical data, as they have been developed and applied mostly in systems biology (see e.g., Hendrickx, Hendriks, Eilers, Smilde, & Hoefsloot, 2011) and neuroscience (see e.g., Gates & Molenaar, 2012; Sporns, 2011), are not optimally suited for data from ESM studies, for several reasons. First, ESM studies do not feature very long time series on a single system (i.e., the number of time points per subject is limited), which hampers the application of typical time series modeling techniques (e.g., J. D. Hamilton, 1994; Shumway & Stoffer, 2010). Secondly, ESM data are hierarchically structured because several persons are measured repeatedly leading to measurements that are clustered within persons (Schwartz & Stone, 1998). This hierarchical structure necessitates the use of separate models for each individual. In combination with the relatively short time series, this leads to unstable results when traditional network models are applied.

In the present article, we demonstrate a statistical method that is tailored to extract network structures from ESM data. We present a multilevel approach to vector autoregressive (VAR) modeling that optimally utilizes the nested structure that typically arises in ESM protocols. This approach is applied to data from an ESM study with a sample of people who feature residual depressive symptoms after a depressive episode (see Geschwind et al., 2011), and validated in a normal sample. This paper presents the first glimpses of the dynamic weighted network architecture of psychopathology, and develops a methodology that yields new possibilities to analyze and understand the structure of disorders.

The outline of the paper is as follows: first, we elaborate on the ESM study used for the analysis and introduce the methodology, the multilevel-VAR method. Second, we explain how a network can be inferred from the data by estimating the average connection strengths between symptoms or variables of interest. Third, we show how the multilevel-VAR method provides information about inter-individual differences in addition to the average network. Fourth, we discuss how network models can be extended with explanatory variables, and how the networks as such can be further analyzed through local and global analyses. In the fifth section, we show how much of the main results can be replicated using an independent dataset that serves as cross-validation. The software code (in *R*; R Core Team, 2012) and data necessary to perform the analyses that result in the main figures reported in this article can be found online.

2.1 Method

Data

We inferred a network structure of six items from an ESM study (Geschwind et al., 2011). The ESM study followed 129 participants with residual depressive symptoms over the course of 12 days, of which the first six days were the baseline period. The following six days took place after 2-3 months, after the participants had been randomly divided into a treatment group (63 participants receiving mindfulness therapy (mean age of 44.6 years and $SD = 9.7$; 79% female)) and a control group (66 participants assigned to a waiting list with a mean age of 43.2 years and $SD = 9.5$; 73% female). Every day subjects were randomly notified by a beeper in each of ten 90-minute time blocks between 7:30 am and 10:30 pm. When signaled, they had to fill out the ESM self-assessment form assessing mood and social context in daily life. This resulted in a maximum of 60 responses per period (baseline or post-baseline). All self-assessments were rated on 7-point Likert scales.

For the purpose of our analysis, we selected a number of items that captured distinctive kinds of mood states. Mood states can be broadly differentiated in terms of their valence (positive/negative) and their degree of arousal (high/low; Barrett, 1998; Reisenzein, 1994; Russell, 1980; Russell, Weiss, & Mendelsohn, 1989; C. A. Smith & Ellsworth, 1985). We included four items that covered different values of the two factors of the mood space. Regarding positive mood, we chose the items 'I feel cheerful' and 'I feel relaxed' to represent high and low arousal respectively. For representing negative mood, we chose the items 'I feel fearful' and 'I feel sad', which capture the subjective experience of high and low arousal respectively (Baas, De Dreu, & Nijstad, 2008; R. Larsen & Diener, 1992; D. Watson & Tellegen, 1985). Furthermore, we included the item 'worry' because worrying is thought to play a significant role in emotion regulation, including the onset and maintenance of negative mood (Borkovec, Ray, & Stober, 1998; Brosschot, Gerin, & Thayer, 2006; Gruber, Eidelman, & Harvey, 2008). The sixth item of the network, 'pleasantness of the event', concerned the environmental context, and assessed the pleasantness of the most important event that happened between the current and the previous response.

Introducing Multilevel-VAR

To overcome the difficulties that accompany the analysis of nested longitudinal data we developed a novel combination of VAR (e.g., J. D. Hamilton, 1994; Pfaff, 2008) and multilevel modeling (e.g., Snijders & Bosker, 2012). A VAR model is a multivariate extension of an autoregressive (AR) model (Shumway & Stoffer, 2010). An AR model is typically applied to a repeatedly measured variable obtained from a single subject. In this way, the time dynamics within an individual are modeled. An

AR model can be considered as a regression model in which a variable at time point t is regressed to a lagged (measured at a previous time point, $t-1$) version of that same variable (Walls & Schafer, 2006). In VAR the time dynamics is modeled for multiple variables. Thus, variables are regressed on a lagged version of the same variable and all other variables of the multivariate system. By combining the VAR model with a multilevel model, time dynamics can be modeled not only within an individual, but also at group level, since the multilevel model allows the VAR coefficients to differ across individuals. Thus, a combination of both models allows for modeling both individual and population dynamics.

The combination of both modeling approaches has, to the best of our knowledge, not yet extensively been studied or applied in the statistical, psychometric or econometric literature. The methods developed in Lodewyckx, Tuerlinckx, Kuppens, Allen, and Sheeber (2011) and Oravecz et al. (2011) have an approach that comes close to what is presented in this paper. However, both methods have a Bayesian and more complex modeling approach and are not easily generalizable to ESM data (Lodewyckx et al., 2011) or can only estimate bivariate symmetric models (Oravecz et al., 2011). Consequently, the specific disadvantages of Lodewyckx et al. (2011) and Oravecz et al. (2011) make them not directly applicable for network inference as we envision it. The modeling approach of Pe and Kuppens (2012) has a similar goal to the method presented in this paper, but makes more approximations (because only bivariate models are used, even though a network of four variables is inferred). Other recent approaches using VAR and/or multilevel can be found in the literature (Funatogawa, Funatogawa, & Ohashi, 2007; Horváth & Wieringa, 2008; Schmid, 2001; Tschacher & Ramseyer, 2009; Tschacher, Zorn, & Ramseyer, 2012). However, in the majority of these studies, the dynamic parameters are not treated as random effects but as mere fixed effects (for an exception, see: Horváth & Wieringa, 2008). In addition, many of these studies do not consider a network approach, nor do they present an accessible way of applying the proposed methodology. In the present paper we present a comprehensive random effects modeling strategy that is optimized to the context of network inference in psychopathology, is implemented in *R* (R Core Team, 2012), and can be easily passed on to network analysis routines.

The population network

In this section we explain how a population network of the six variables (cheerful, relaxed, sad, worry, fear and event) can be inferred with the multilevel-VAR method. The main goal is to estimate the average connection strengths between all variables in the population. These connection strengths can then be represented in a network. To estimate these connection strengths we apply the multilevel-VAR method to the measured values at baseline of the six variables. For an arbitrarily chosen criterion variable (i.e., cheerful, relaxed, sad, worry, fear or event, for $j = 1, 2, ..., 6$, respectively), the model

equation is as follows:

$$\begin{aligned} Y_{pdtj} = \gamma_{0pdj} + \gamma_{1pdj} \cdot cheerful_{p,d,t-1} + \gamma_{2pdj} \cdot sad_{p,d,t-1} + \gamma_{3pdj} \cdot worry_{p,d,t-1} \\ + \gamma_{4pdj} \cdot fear_{p,d,t-1} + \gamma_{5pdj} \cdot event_{p,d,t-1} + \gamma_{6pdj} \cdot relaxed_{p,d,t-1} + \varepsilon_{pdtj}. \end{aligned} \tag{2.1}$$

In our case, Y_{pdtj} represents the measurement for person p ($p = 1,\ 2, \ldots,\ 129$) at day d ($d = 1,\ 2, \ldots, 12$) and time t of the j-th criterion variable. Equation 2.1 represents the multiple regression of a single variable at time point t on all other variables at time point $t-1$. Because there are six variables, there are six such regression equations – one for each variable. At baseline (i.e., at days 1 to 6 before the therapy treatment is applied, such that $d < 7$), the regression coefficients (i.e., intercept and regression weights) are decomposed as follows:

$$\gamma_{kpdj} = \beta_{kj}^{base} + b_{kpj}, \tag{2.2}$$

where β_{kj}^{base} represents the population average effect (fixed effect) at baseline of the lagged variable k (for $k = 0$, this is the intercept) on the criterion variable j, and b_{kpj} is the person-specific deviation (random effect) of this general effect. In the remainder, person-specific effects will always be denoted in Roman letters. In order to illustrate our model, let us consider the regression equation for the variable 'cheerful'. Because we identify all variables explicitly with their names, we only use the j-index to distinguish the regression coefficients, but not to identify the variables (hence, the variables carry only three indices, as compared to 2.1). At baseline ($d = 1, 2,\ \ldots,\ 6$), the model reads (not all predictors are explicitly included in the interest of clarity):

$$\begin{aligned} cheerful_{pdt} = (\beta_{01}^{base} + b_{0p1}^{pre}) \\ + (\beta_{11}^{base} + b_{1p1}) \cdot cheerful_{p,d,t-1} \\ + (\beta_{21}^{base} + b_{2p1}) \cdot sad_{p,d,t-1} + \ldots \\ + (\beta_{61}^{base} + b_{6p1}) \cdot relaxed_{p,d,t-1} + \varepsilon_{pdt1}. \end{aligned}$$

Focusing on the baseline level, we may now construct a 6-by-6 matrix B^{base} with the fixed effects β_{kj}^{base} ($k, j = 1, \ldots, 6$). The matrix B^{base} captures the dependence of the 6-dimensional state (i.e., cheerful, sad, worry, fear, event, and relaxed) of a typical individual (i.e., for which $b_{kpj} = 0$) upon the previous 6-dimensional state (all effects at baseline). A specific element β_{kj}^{base} thus expresses the extent to which variable k at time point $t-1$ is related to variable j at time t, while controlling for all other variables. The elements on the diagonal (i.e., β_{jj}^{base}) are the autoregressive effects (self-loops),

while the off-diagonal elements are the cross-regressive effects (β_{kj}^{base}; connections between different variables). Note that, in general, B^{base} is asymmetric.

The matrix B^{base} can be viewed as an adjacency matrix (Boccaletti, Latora, Moreno, Chavez, & Hwang, 2006) of a weighted network. The matrix B^{base} contains the fixed effects of the multilevel-VAR model and represents the lag 1-links between the nodes (i.e., the variables). Thus the matrix B^{base}can be thought of as the population average of the network structure. Because we are looking at several specific links, we control for multiple testing by controlling the False Discovery Rate (FDR method; Benjamini & Hochberg, 1995)) at 5%. The generated network structure can be visualized through the *R* package *qgraph* (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012). Only connections that surpass the significance threshold are shown in the visual representation.

Since fitting a multilevel-VAR model directly to the multivariate time series of the participants in the sample is computationally challenging, we approach the problem by breaking up the complicated multivariate problem into a series of easier-to-compute univariate models which are integrated in a second step (i.e., by representing them as a network). This approach can be considered as an instance of the so-called pseudo-likelihood method (Arnold & Strauss, 1991; Fieuws & Verbeke, 2006). By using univariate models, most parameters are estimated directly (e.g., all fixed effects β_{kj} and variances of the error terms ε_{pdtj}). However, some parameters of the model such as the correlations between error terms of the different univariate regression models can only be estimated indirectly in our approach. In Appendix 2.A, a more elaborate description of the pseudo-likelihood method is given, and it is shown through simulations that point estimates of most directly and indirectly estimated parameters are on average close to the true values. This indicates that the pseudo-likelihood fitting procedure of the multilevel-VAR approach is a feasible alternative to full likelihood fitting procedures. The modeling is carried out using the *lme4* package in *R* (see *R*-code; Bates, Maechler, & Bolker, 2012).

Individual differences

The multilevel-VAR method provides information about inter-individual differences (random effects) in the network, in addition to the population average network (see Equation 2.2). Through the random effects we can construct networks of individual variability and infer a network for each subject of the ESM study separately (see *R*-code). In this paper, we take a random effect approach to estimate inter-individual differences, and assume that these person-specific parameters b_{kj} are drawn from a multivariate normal distribution with a zero mean vector and an unstructured covariance matrix (see e.g.,Verbeke & Molenberghs, 2000; see Equation 2.2). Other approaches to deal with inter-individual differences are fixed-effects analysis (i.e., constructing a dummy variable for each subject; Baltagi, 2008) and conditional analysis (see e.g., Verbeke, Spiessens, & Lesaffre, 2001). In the multilevel-

VAR method a random-effects approach is taken because it avoids possible problems related to the two previously mentioned approaches. The approach is more parsimonious in terms of number of parameters: instead of having to estimate explicitly all person-specific parameters as in the dummy variable approach, only the variance parameters have to be estimated (Neyman & Scott, 1948), which at the same time avoids problems of inconsistent estimators (Gelman & Hill, 2007). The random-effects approach also allows one to evaluate all effects, in contrast to the conditional analysis approach, in which effects of between-person variables, such as possible therapy effects, cannot be evaluated (Tuerlinckx, Rijmen, Verbeke, & Boeck, 2006).

To construct a network representing individual variability, we take the estimate of the population standard deviation of the person-specific (random) effects $SD(b_{kpj})$. Thus, each connection in the network represents the SD of the random effects for that specific connection. Connections in the network that have a large standard deviation represent a high variation of the value (connection strength) of that specific connection over individuals. In addition, the model in Equations 2.1 and 2.2 allows for constructing a network of a single subject. These $N = 1$ networks are a combination of the individual random effect, which is added to the fixed effect of the relevant link (connection) in the network. For instance, in the individual network of person p at baseline, the link from node k to node j has a value of $\beta_{kj}^{base} + b_{kpj}$ (see Equation 2.2).

Extending the network model with explanatory variables: Local and global network analyses

As is the case in a standard multilevel analysis, explanatory variables that might explain part of the inter-individual variability can be added (called level-2 variables in standard multilevel terminology, see Snijders & Bosker, 2012). In this paper, we present two examples. In the first example, the explanatory variable therapy-intervention is added to the standard model. We compare the network of the therapy group with the non-therapy group by comparing specific links in the networks. A therapy effect on the network structure implies a significant three-way interaction. For example, if there is a therapy effect on the link from sad to cheerful, this means that the interaction between the variables therapy (therapy or control), time (pre or post baseline) and sad (ranging from 1 to 7) is significant in the regression model that applies to the variable cheerful, signifying that the effect of feeling sad on feeling cheerful has changed from pre- to post-therapy.

In the second example, we explain variability in individual networks by relating it to covariates; here, neuroticism functions as an example. We present a global network analysis, in which the overall structure of the network is taken into account; these analyses contrast with local network analyses, which compare specific connections across networks. A representative example of such a network

analysis is a centrality analysis. We will examine whether the structure of the network regarding centrality changes when the degree of neuroticism changes. This question is approached by looking at differences in the network structure of three different groups: low, mid and high neuroticism.

Therapy: Local network analysis

In order to analyze whether therapy had a significant effect on the network structure we added the variable 'therapy-intervention' to the baseline model (see Equation 2.2; see *R*-code). Thus, besides reports measured at baseline (i.e., $d < 7$), we also added post-baseline measurement instances (i.e., $d \geq 7$). The regression coefficients (for which $k = 0$ is the intercept and $k > 0$ are the regression weights) are now equal to:

$$\gamma_{kpdj} = \beta_{kj}^{base} + \beta_{kj}^{post} + \delta_{kj} Therapy_p + b_{kpj}, \tag{2.3}$$

where $d \geq 7$ and the term $Therapy_p$ equals to 0 if person p belongs to the control group, and takes value 1 if the person received mindfulness therapy. As can be seen from the equation, β_{0j}^{post} represents the difference between the intercept at baseline and post-baseline for the control group. In general, Equation 2.3 allows for a difference between the mean of the control and the therapy group, so differences between the two groups post-baseline are accommodated for. A comparison of Equations 2.2 and 2.3 shows that the model assumes person-specific deviations from the regression weights to be the same pre- and post-baseline (i.e., persons who deviate in a particular way from the mean structure during baseline will continue to do so post-baseline). This restriction is made for reasons of parsimony. However, for the intercept, the model allows person-specific deviation of the general intercept to be different pre- and post-baseline (therefore the pre-baseline person-specific deviation will be denoted as b_{0pj}^{pre} and post-baseline as b_{0pj}^{post}).

To illustrate this model, let us consider the regression equations for the variable 'cheerful'. The post-baseline ($d = 7, \ldots 12$) model for the controls becomes

$$\begin{aligned} cheerful_{pdt} = &(\beta_{01}^{base} + \beta_{01}^{post} + b_{0p1}^{post}) \\ &+ (\beta_{11}^{base} + \beta_{11}^{post} + b_{1p1}) \cdot cheerful_{p,d,t-1} \\ &+ (\beta_{21}^{base} + \beta_{21}^{post} + b_{2p1}) \cdot sad_{p,d,t-1} + \ldots \\ &+ (\beta_{61}^{base} + \beta_{61}^{post} + b_{6p1}) \cdot relaxed_{p,d,t-1} + \varepsilon_{pdt1}, \end{aligned}$$

while that for the therapy group equals

$$
\begin{aligned}
cheerful_{pdt} = &(\beta_{01}^{base} + \beta_{01}^{post} + \delta_{0j} Therapy_p + b_{0p1}^{post}) \\
&+ (\beta_{11}^{base} + \beta_{11}^{post} + \delta_{11} Therapy_p + b_{1p1}) \cdot cheerful_{p,d,t-1} \\
&+ (\beta_{21}^{base} + \beta_{21}^{post} + \delta_{21} Therapy_p + b_{2p1}) \cdot sad_{p,d,t-1} + \dots \\
&+ (\beta_{61}^{base} + \beta_{61}^{post} + \delta_{61} Therapy_p + b_{6p1}) \cdot relaxed_{p,d,t-1} + \varepsilon_{pdt1}.
\end{aligned} \tag{2.4}
$$

Analogous to the construction of B^{base}, as described in the previous section, we may construct matrices $B^{post-control}$ and $B^{post-therapy}$, which can be interpreted as network structures that describe the post-intervention behavior of the relevant variables as they apply to control and therapy groups. Through this model we can evaluate the effect of therapy, by looking at the three-way interactions between a predictor variable, the post-baseline indicator and the therapy-indicator (the parameters of interest are δ_{kj} in Equations 2.3 and 2.4). Because we are looking at several specific links, we control for multiple testing. This is done by controlling the False Discovery Rate (FDR method; Benjamini & Hochberg, 1995) at 5%.

Neuroticism: Global network analysis

Important information about a network can be gained by analyzing its global structure, for example by looking at the relative centrality of different nodes. In a centrality analysis, nodes are ordered in terms of the degree to which they occupy a central place in the network. Relevant centrality measures can be constructed in different ways (Opsahl, Agneessens, & Skvoretz, 2010); here, we focus on betweenness centrality. Betweenness centrality takes direct and indirect weighted links between the nodes into account. First, for each pair of nodes x and y (e.g., worry and cheerful), the strongest direct and/or indirect connecting paths from x to y and from y to x are determined. Then for each node, it is calculated to which degree the node lies on the shortest path between two other nodes. The more often a node lies on the shortest path between two other nodes, the more the node can funnel and influence the flow in the network, and the higher its betweenness centrality is (Opsahl et al., 2010). To evaluate whether betweenness centrality of the network changes when the degree of neuroticism changes we added the variable neuroticism to the regression model in the same way as the variable therapy was added (see *R*-code):

$$
\gamma_{kpdj} = \beta_{kj}^{base} + \beta_{kj}^{post} + \delta_{kj} Therapy_p + \eta_{kj} Neuroticism_p + b_{kpj}. \tag{2.5}
$$

In this study, neuroticism was assessed with the NEO-FFI scale of neuroticism (Hoekstra, Ormel, & De Fruyt, 1996). In order to be able to deal with possible nonlinear effects of neuroticism on the

network structure, the continuous neuroticism measure was subdivided into three groups (based on the three quartiles) resulting in a low, middle and high neuroticism group, corresponding respectively to sum scores 12-34, 35-45, and 46-60 on the NEO-FFI scale. The term $Neuroticism_p$ equals to 0 if person p belongs to the low neuroticism group, takes value 1 if the person p belongs to the middle neuroticism group and equals to 2 if the person p belongs to the high neuroticism group. For reasons of parsimony, we let neuroticism interact with the connection strengths of the baseline network only.

For computing betweenness centrality, we used the *R* package *qgraph* (Epskamp et al., 2012). To assess the uncertainty of betweenness centrality, we used a nonparametric bootstrap method to construct the distribution of the betweenness statistic under the null hypothesis that the fitted model is correct (Efron & Tibshirani, 1994). To achieve this, the bootstrap was implemented for the multilevel model, taking time dependency into account. As a result, the latter also implies that the *R*-code cannot be run on a standard computer, due to extra computational difficulty (bootstrapping large multilevel models is much more computationally demanding than bootstrapping, for example, linear models). In total 1000 datasets were bootstrapped.

Finally, the multilevel-VAR model was fitted to each of the 1000 simulated datasets, and from the estimated coefficients, the betweenness at baseline for low, mid and high levels of neuroticism was computed. From the distribution of betweenness scores, we calculated the median and the 50% and 95% bootstrap confidence intervals (see *R*-code for an example of the nonparametric bootstrapping procedure).[1]

Replication of the results: A validation dataset

In order to test if results found with the multilevel-VAR method could be replicated, we compared the main outcomes of the main dataset with a second validation dataset. The validation data we used was from an ESM study of Kuppens (part of the data are published in Koval, Kuppens, Allen, & Sheeber, 2012; Pe, Koval, & Kuppens, 2013; Pe, Raes, et al., 2013). In this ESM study, 97 university students (with a mean age of 19.1 years, $SD = 1.3$; 63% female) were followed over the course of seven days. The participants had to fill out an ESM self-assessment form assessing mood and social context in daily life 11 times a day. This resulted in a maximum of 77 responses. All self-assessments were rated on scale from 0 to 100. From this dataset, we selected the variables that this set had in common with the variables of the main dataset: cheerful, relaxed, sad, worry and fear. Note that 'worry' was assessed slightly differently in the validation study: "How much have you worried since the previous beep" instead of "I am worrying at the moment". Furthermore, the pleasantness of events was not

[1] This part is based on a corrected version of the paper. The nonparametric bootstrap procedure was executed with help of Merijn Mestdagh.

measured in this study. To increase comparability, networks inferred from these five variables were compared with networks inferred from the five corresponding variables of the main dataset.

First, we inferred a population network containing the five variables cheerful, relaxed, sad, worry and fear for both the main dataset and the validation dataset. Then the connection strength of the links of the main network was correlated with the links of the validation network. The higher the correlation, the better the two inferred networks agree. To assess the correlation, we used both Pearsons product moment correlation and Spearmans rank order correlation coefficient. In addition, we assessed to which extent the variances of inter-individual differences are comparable in the two studies. The correlation between the variances of the random effects of the links of both networks was calculated using Spearmans product moment correlation and Pearsons rank order correlation coefficient.

In the validation dataset, there is no therapeutic intervention, so the local network analysis could not be replicated. However, neuroticism was measured in the validation set, and thus we applied the global network analysis to the validation set. Hence, we tested whether the centrality of the network changes in the same way in both datasets when the degree of neuroticism varies. Again, we used only the five variables that both sets have in common. In this ESM study, neuroticism was measured with the Dutch version of the Ten Item Personality Inventory (Gosling, Rentfrow, & Swann, 2003; Hofmans, Kuppens, & Allik, 2008) with a sum score ranging from 1 to 7. Neuroticism was again subdivided into three groups: a low, middle and high neuroticism group, corresponding respectively to sum scores 1-2, 2.5-4.5, and 5-7 on the TIPI scale.

Model assumptions

In order to apply the multilevel-VAR model three assumptions on which the model is built need some further commenting. The first assumption is that we start the clock again at the start of each day as to avoid the day-night problem, which means that we do not use the measurements of yesterday to predict the measurements of today (because a night separates the two days). A night is a relatively large time interval and is psychologically and physiologically qualitatively different from daytime (e.g., Lavie, 2001). Thus, the first measurement of the day was excluded from analysis. With regard to time it is furthermore assumed that the time intervals between two consecutive measurements are approximately equal.We will come back to both aspects when discussing the results.

Stationarity is a second important assumption inherent to the model. In order for a process to be (weakly) stationary, the mean and variance of the series must stay unchanged over time (Box, Jenkins, & Reinsel, 1994). Stationarity was tested with the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test separately for every subject and variable pre and post intervention. The null hypothesis of the KPSS

test is that a time series is stationary (Kwiatkowski, Phillips, Schmidt, & Shin, 1992). Furthermore, a general check was executed to test for a trend, and thus non-stationarity, in the overall data. This was done by comparing the model outlined in the previous section with a model into which a person-specific linear deterministic trend was added (using the beep number as a predictor). For both models, the Bayesian Information Criterion (BIC) was calculated (by summing the separate BICs of the six univariate models). The BIC is a comparative model selection method that takes both the goodness-of-fit of a model and the complexity of the model (as measured by the number of parameters) into account. Models with a large number of parameters are penalized (Schwarz et al., 1978). The model with the lowest BIC is the preferred model.

The specific order of the model is the third assumption. For reasons of parsimony, we present only the results of the baseline models with lag-1 predictors included. However, we also fitted competing models of orders two and three (i.e., with all lags included up to the specified order). In order to keep the problem computationally tractable, we did not allow for random effects on predictors of lags larger than one and in the main dataset we constrained the additional lag effects to be equal at pre- and post-baseline and in control and therapy groups.

2.2 Results

This section is organized as follows. We start by discussing the validity of the stated assumptions (because the validity of the results depends on the veracity of the assumptions). Subsequently, we discuss the population network, individual differences, and the effect of explanatory variables.

Assumptions

Since a measurement is not allowed to predict the following measurement overnight, we deleted the first measurement of each day. Furthermore the data had to be lagged. Together this led to a reduction in the number of reports included in the analysis: The average number of useable data points went down from an average of 49 to an average of 35 reports for each period (baseline and post-baseline). Regarding the assumption of equally spaced time points, the ESM study, having a quasi-random beeping scheme, violates this assumption. However, the extent of the violation is taken to be small, since the variation in between-measurement points is relatively small with an average of 1.5 hours and a standard deviation of 0.54.

Concerning stationarity, the KPSS test indicated that a vast majority of the data was stationary (about 77%). In addition, the BIC indicated that the models without trend were a better fit to the data (BIC= 172896) than the models with linear trend (BIC= 172995). Thus, overall the data are judged to be sufficiently stationary.

Regarding the lag order, the BIC indicated that the order-3 model fitted best and the order-2 model fitted better than the order-1 model. However, the lag-1 coefficients were very similar across the three models. Since the impulse response functions (J. D. Hamilton, 1994; Lütkepohl, 2007) also did not reveal any substantive effects of interest, which could have warranted a more complex analysis, we proceeded with the order-1 results.

The population network

The inferred population network at baseline is presented in Figure 2.1 (i.e., the matrix B^{base}). Each variable is represented by a node in the network and relations between items are represented by the weighted arrows (connection strength) between nodes. The arrow from item k to item j is a visual depiction of the weight β_{kj}, expressing the strength of the relation between item k at time $t-1$ and item j at time t. Arrows can be either red, indicating a negative relationship (i.e., $\beta_{kj} < 0$), or green, indicating a positive relationship (i.e., $\beta_{kj} > 0$). Furthermore, the strength of the relation from item k to item j (i.e., a more extreme value of β_{kj}) is translated into the thickness of the arrows: the thicker the arrow between two nodes, the stronger the relation between the nodes or items. Note that item responses can also be predicted from the previous state of the item itself. These arrows are the self-loops in the network.

In Figure 2.1, only arrows that surpass the threshold for significance (i.e., p-value of the t-statistic is smaller than 0.05) are represented in bold in the network; the non-significant arrows are made transparent. Controlling for multiple testing by controlling the False Discovery Rate (FDR method; Benjamini & Hochberg, 1995) at 5% does not lead to qualitatively or quantitatively different conclusions. From Figure 2.1, a few general insights on the dynamical network structure between the six items can be derived. First, in accordance with a dynamical view on emotions, both the positive and the negative items form a cluster representing self-perpetuating cycles in which the components of negative and positive emotions interact (see also Fredrickson & Joiner, 2002; Zelenski & Larsen, 2000). We find that positive or excitatory connections exist among items of the same valence, while negative or inhibitory relationships exist among clusters of mood states of opposite valence (e.g., cheerful, relaxed and pleasant event on the one hand and sad, worry and fearful on the other hand). This is in line with existing theories in affect research (J. T. Larsen, McGraw, & Cacioppo, 2001; Pe & Kuppens, 2012; Russell & Carroll, 1999; D. Watson, Wiese, Vaidya, & Tellegen, 1999).

A second insight from Figure 2.1 is that the self-loops or autoregressive effects are always positive and they are generally among the strongest connections in the network, indicating that, for instance, the current experience of worry or cheerfulness predicts future feelings of worry or cheerfulness. At a more detailed level, we see that in the baseline model, for example, worry leads to increases in negative

affect by enhancing negative moods and inhibiting positive moods. This lines up well with previous findings (e.g., McLaughlin, Borkovec, & Sibrava, 2007; Moberly & Watkins, 2008; Segerstrom, Tsao, Alden, & Craske, 2000) and supports the validity of our approach.

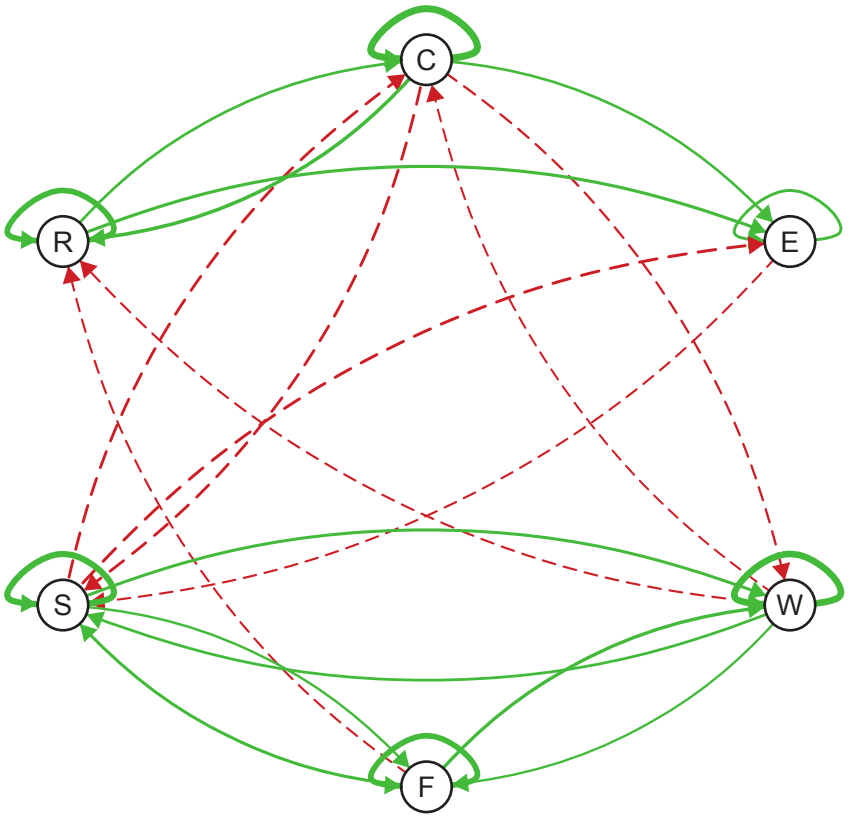


Figure 2.1: *Estimated population network at baseline.* The six items are: C=cheerful, E=pleasant event, W=worry, F=fearful, S=sad and R=relaxed. Solid green arrows correspond to positive connections and red dashed arrows to negative connections. Only arrows that surpass the significance threshold are shown (i.e., for which the p-value of the t-statistic is smaller than 0.05). Arrows can be either red, indicating a negative relationship (i.e., $\beta_{kj} < 0$), or green, indicating a positive relationship (i.e., $\beta_{kj} > 0$). Furthermore, the strength of the relation from item k to item j (i.e., an extremer value for β_{kj}) is translated into the thickness of the arrows: the thicker the arrow between two nodes, the stronger the nodes or items are related. Note that item responses can also be predicted from the previous state of the item itself. These arrows are the self-loops in the network.

Individual Differences

The multilevel-VAR method also provides information about inter-individual differences (random effects) in the network in addition to the population average network (fixed effects). The links with the largest inter-individual differences are shown in Figure 2.2. The arrows in the network now represent the estimated variance of the relevant VAR parameters over individuals. Only arrows containing a $SD(b_{kpj})$ larger than 0.1 are emphasized in Figure 2.2. For example, the pronounced self-loop on the

item ‘worry’ indicates a high individual variability.

This individual variability can also be immediately observed in the networks of individual subjects. Figure 2.3 illustrates the individual networks at baseline for two persons. The network on the left has a quite strong self-loop for the item ‘worry’, which means that when this person worries, he or she tends to worry for a longer time. On the other hand, the network of the participant on the right has a weak self-loop for the item ‘worry’, meaning that when this person starts to worry he or she is likely to worry for only a short time. Thus, not only can we verify which arrows have a high inter-individual variability; we can also immediately indicate what these arrows look like in networks that apply to an individual person.

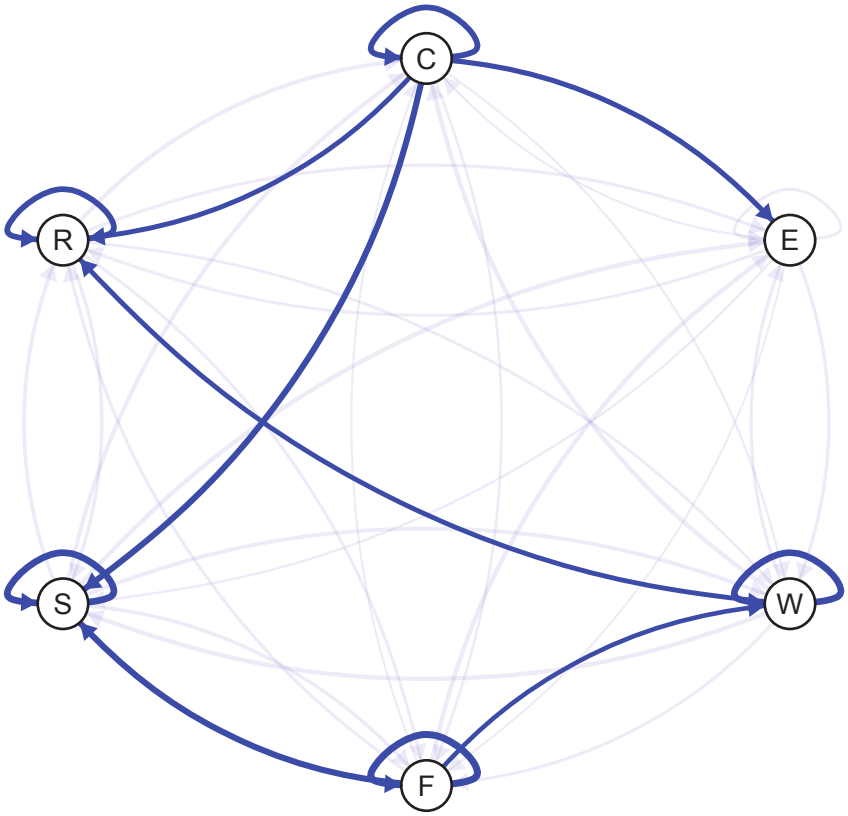


Figure 2.2: *Inter-individual differences of the arrows of the network from Figure 2.1.* The thickness of the arrows is based on the size of the standard deviation of the random effects. To construct the figure, we have put a cutoff of 0.1 on the standard deviation and only the standard deviations above the cutoff are shown with a non-transparent arrow. As the threshold for the standard deviation of the random effects 0.1 was chosen because it represents large inter-individual differences. The average coefficient of the self-loops (i.e., autoregression coefficients) is about 0.2 with a random effects standard deviation of 0.1. Therefore, assuming a normal distribution, the range from 0 to 0.4 represents 95% of the individual self-loop coefficients. With a larger cutoff, such as 0.2, also individuals having negative self-loops would be taken into account. However, more than 95% of the population has a positive self-loop strength.

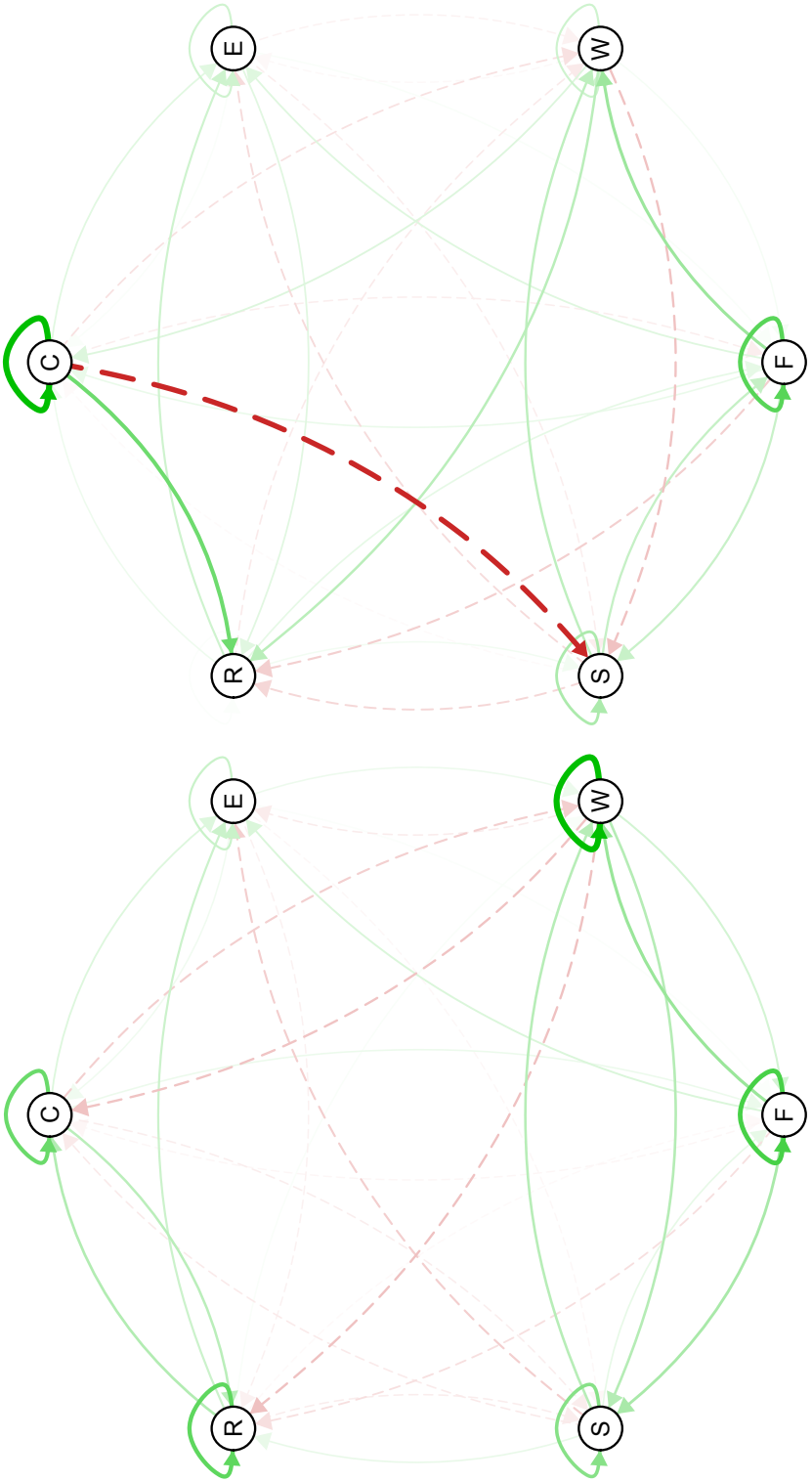


Figure 2.3: *Individual networks (at baseline) of two different persons.*

Therapy: Local network analysis

To evaluate the effect of therapy on the local network structure we compared the arrows in the networks of the therapy and the control group. After correcting for multiple testing (using FDR controlled at 5%), none of the arrows indicated a significant effect of therapy on the network structure, meaning that there was no significant three-way interaction between the arrow, the post-baseline indicator and the therapy-indicator. However, this does not imply that there is no effect of therapy at all. First of all, as shown in previous research (Geschwind et al., 2011), therapy has an effect on the average levels of some variables, and also in this study we can detect effects of therapy on the mean level of, for instance, cheerfulness. Secondly, the fact that we did not find an effect of therapy on the network structure here could also be due to a lack of power. Correcting for multiple testing always leads to a decrease in power, which can lead to missing an effect on the network structure that is small but still relevant.

Neuroticism: Global network analysis

To assess the effect of neuroticism on the global network structure, we tested whether the structure of the network regarding betweenness centrality changes as a function of neuroticism. Figure 2.4 presents the results of the betweenness analysis for low, middle and high neuroticism at baseline. For every item, the model-based estimate of betweenness is calculated, together with a bootstrap simulated 50% and 95% confidence interval. Plotting both 50% and 95% confidence interval gives an indication of the asymptotic distribution of the estimate.

Although the distributions of the betweenness coefficients are quite wide (as are the associated confidence intervals), the data do suggest some interesting trends. In order to get a good interpretation of the effects of neuroticism on betweenness, it is insightful to look at the effects on the entire betweenness distribution. Whereas the centrality of the nodes fearful and event are low and stable across groups, the positive nodes cheerful and relaxed become less central as neuroticism increases. This is indicated by the distribution, which clearly shifts downwards. Notably, worry has a higher centrality distribution in the high neuroticism group than in the low and mid neuroticism group. That is, worry becomes one of the most central nodes in the high neuroticism group. This result is in line with studies suggesting that worry is an important manifestation of neuroticism (Muris, Roelofs, Meesters, & Boomsma, 2004), and with the idea that worry is a cognitive concomitant of neuroticism (Segerstrom et al., 2000).

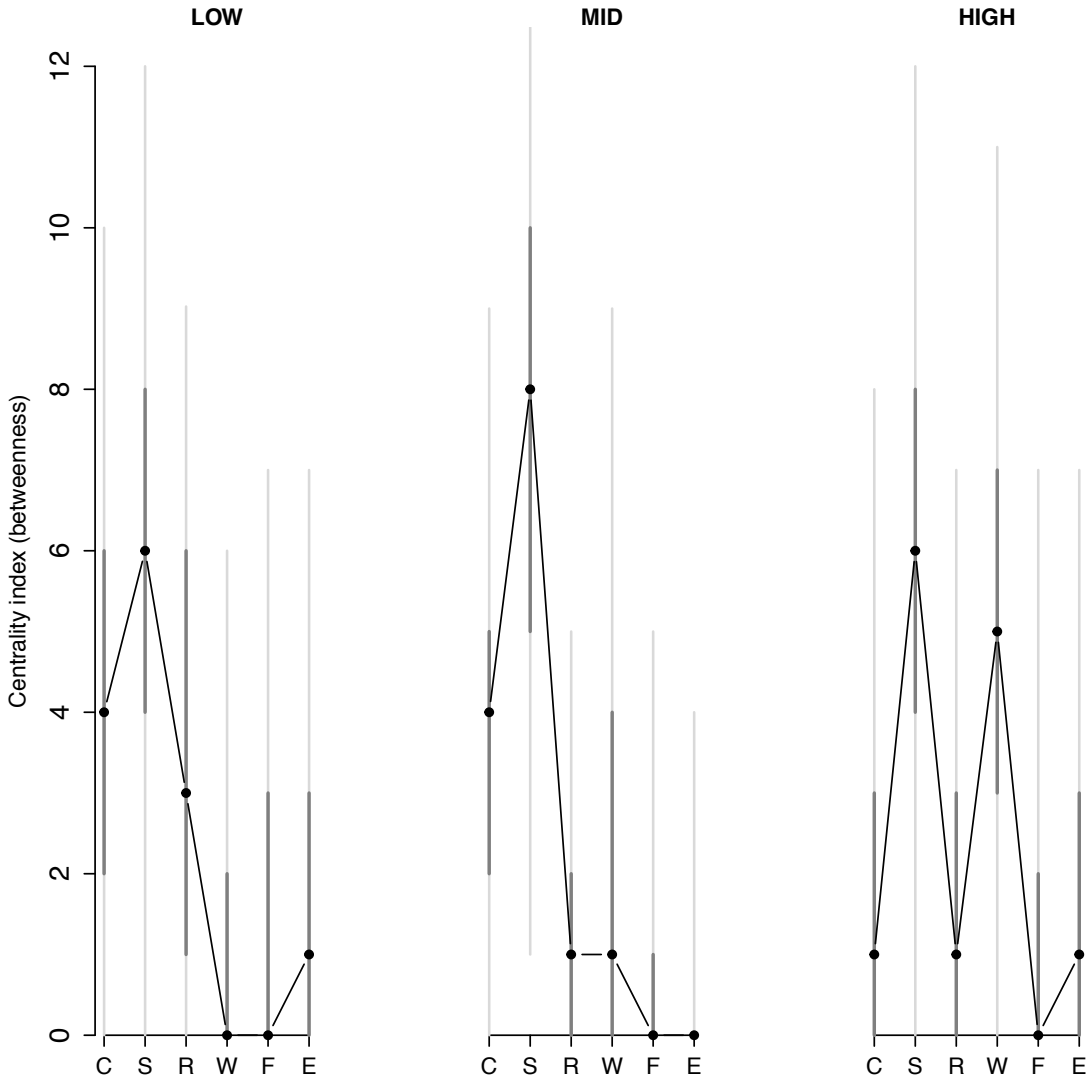


Figure 2.4: *Centrality (betweenness) of each item in the network as a function of level of neuroticism at baseline.* Low, mid, and high neuroticism are shown from left to right. The labels of the items are abbreviated by their first letter (C=cheerful, S=sad, R=relaxed, W= worry, F=fearful and E=event). The black dots are the model-based estimate of betweenness, the darkgrey vertical lines represent 50% confidence intervals and the light grey vertical lines represent 95% confidence intervals (as estimated from the bootstrap method). Together, the median, 50% and 95 % confidence intervals give information on how the node centrality for every item in all three networks is distributed.

2.3 Replication of the results: A validation dataset

Assumptions

The assumptions for applying a multilevel-VAR model were also met in the validation dataset. Excluding the first measurement of each day and lagging the data led to a reduction in the number of reports included in the analysis: The average number of useable data points went down from an av-

erage of 60 to an average of 53 reports. In this dataset, the assumption of equally spaced time points was also only slightly violated. The variation in between-measurement points was relatively small with an average of 1.2 hours and a standard deviation of 0.49. Regarding stationarity, the KPSS test indicated that a vast majority of the data was stationary (about 70%). In addition, the BIC indicated that the models without trend (BIC= 202100) were a better fit to the data than the models with linear trend (BIC= 202203), indicating that overall the data is stationary. Because the higher order analyses did not reveal any substantially different conclusions and the aim was to compare the results from the two datasets, we pursued an order-1 analysis.

Population network

In the left panel of Figure 2.5, the correlation between the connection strengths of the links of the main population network and the links of the corresponding validation network is shown. The product moment correlation between the connection strengths of the two networks is 0.95 ($p < 0.0001$; the rank order correlation is r=0.96, $p < 0.0001$). This indicates that the population networks between both datasets agree almost perfectly. The networks inferred for the validation data are not shown here, but can be found in Appendix 2.B.

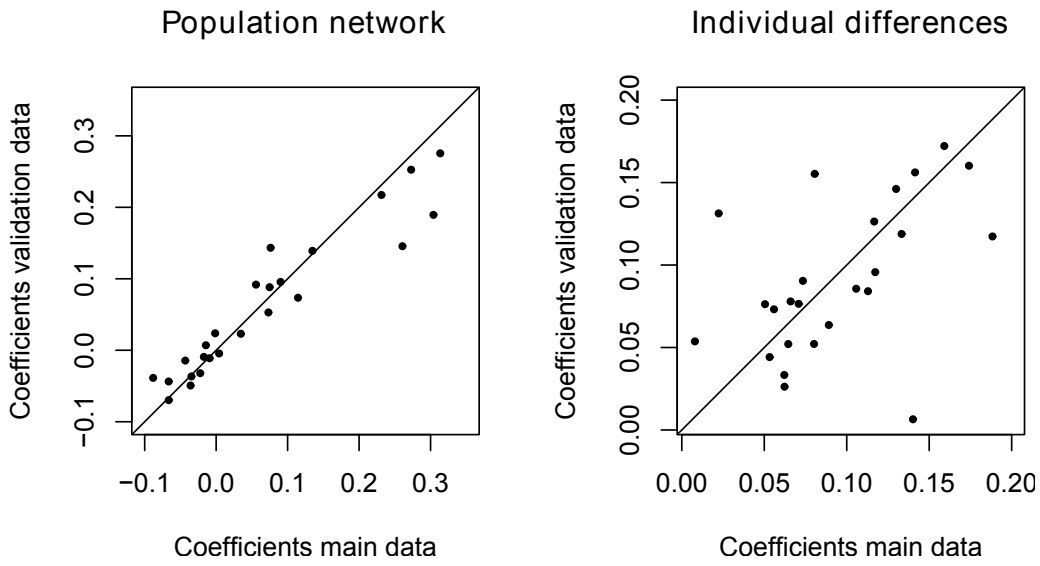


Figure 2.5: *Correspondence between the basis dataset and the validation dataset.* Left panel: Representation of the correspondence between the population network coefficients (fixed effects) of the basis dataset (x-axis) and the validation dataset (y-axis). Right panel: Representation of the correspondence between the inter-individual differences (random effects) of the basis data (x-axis) and the validation data (y-axis).

Individual Differences

In the right panel of Figure 2.5 the correlation between the connection strengths of the links of the main inter-individual differences network and the links of the corresponding validation network is shown. The product moment correlation between the connection strengths of the two networks is a sizeable correlation of 0.50 (p =0.01; rank order correlation is r=0.56, p=0.004). This indicates that although some links in the inter-individual differences networks differ between the two datasets, the majority of them reflect a similar degree of individual variability. We refer again to Appendix 2.B for a visual illustration of the individual differences networks in both datasets.

Neuroticism: Global network analysis

In Figure 2.6 the results of the betweenness centrality analysis for low, middle and high neuroticism of the validation dataset are shown. These results can be compared with Figure 2.4, since the results of the main dataset with five variables are very similar to those with six variables (see Appendix 2.B for the betweenness centrality figure of the main dataset with only five variables). Although worry is again one of the most central nodes in the high neuroticism group, there is no clear shift in centrality between the groups, which we found in the main dataset (see Figure 2.4). In fact, worry seems to be also one of the most central nodes in the low and mid neuroticism group in this dataset. The difference in centrality between the datasets could be related to the overall level of neuroticism. After applying a linear transformation to approximately equate the neuroticism measures in the two groups, we found that in the main dataset the average neuroticism score (M = 40.7; SD = 7.4) was markedly higher than in the validation set (M = 31.1; SD = 12.1; $t(148.68) = -6.9$, $p < 0.0001$). Furthermore, as noted in the Method section, worry was assessed slightly differently in the two datasets, which could also account for the difference in the centrality of worry.

2.4 Discussion

In this paper, we have presented a combination of vector autoregressive (VAR) modeling and multilevel modeling which, to the best of our knowledge, is the first method suited for inferring networks from ESM data. The modeling technique combines time series with individual differences. This strategy allows us to cope with the peculiarities of ESM data (e.g., short time series, significant individual differences) but also opens up unique possibilities for studying individual differences in dynamic structure. Thus, the methodology is an important addition to network methodologies that are currently being developed in personality and clinical psychology and psychiatry (Borsboom et al., 2011; Cramer et al., 2010). For simplicity, we limited the analysis to six variables in this paper, but in principle

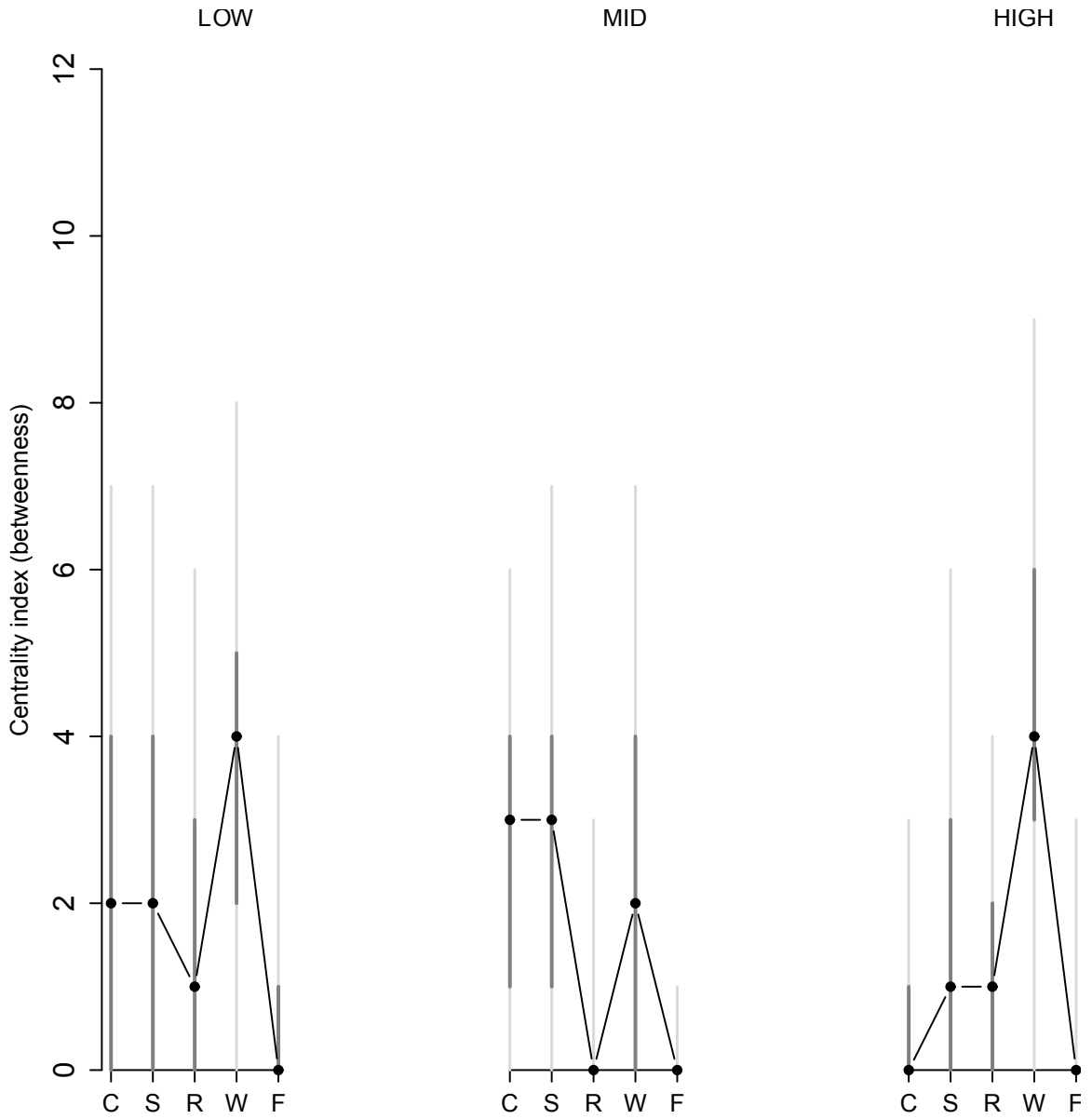


Figure 2.6: *Centrality (betweenness) of each item in the network as a function of level of neuroticism in the validation dataset.* Low, mid, and high neuroticism are shown from left to right. The labels of the items are abbreviated by their first letter (C=cheerful, S=sad, R=relaxed, W= worry and F=fearful). The black dots are the model-based estimate of betweenness, the darkgrey vertical lines represent 50% confidence intervals and the light grey vertical lines represent 95% confidence intervals (as estimated from the bootstrap method). Together, the median, 50% and 95 % confidence intervals give information on how the node centrality for every item in all three networks is distributed.

the analysis is generalizable to larger datasets and to different time series models (e.g., models with different lags). Thus, the methodology is sufficiently flexible to give rise to a relatively comprehensive approach. Furthermore, it is a great advantage that such complex dynamics between several variables can be easily visualized as a network with the *R* package *qgraph* (Epskamp et al., 2012). Illustrating the dynamical interaction between several variables helps to give an immediate intuitive understanding of the complex structure of the model, and is more insightful than a mere verbal explanation

(Gather, Imhoff, & Fried, 2002; Wild et al., 2010).

The multilevel-VAR method combines a nomothetic approach, which makes it possible to generalize findings to a population level, with an idiographic approach, which models dynamical processes at the level of the individual person. In our study, for instance, the fixed effects of the model form a plausible network at group level, which shows the average dynamics between six mood related variables at baseline. Importantly, this population network was replicated in the validation dataset. In both datasets the same dynamics between the variables were found, supporting the validity of the multilevel-VAR method.

In addition, individual heterogeneity can be easily assessed using the random effects estimated in the model. Again, a similar network of individual heterogeneity was found in the validation data. Although some links in the networks of individual heterogeneity differed between the two datasets, the majority of them showed a similar degree of individual variability. Because the two datasets contain different populations, it is to be expected that not all links show the same amount of individual heterogeneity. Intra-individual time series can also be studied by combining fixed and random effects for each subject, which results in individual networks. Thus, our method successfully combines nomothetic and idiographic approaches to data analysis.

In time, the latter approach may lead to improved understanding of intra-individual functioning; this may in turn lead to better therapeutic interventions. A network analysis of a subject receiving therapy may show, for example, that the link between rumination and sadness is the strongest link in the network and that a therapy should intervene on that link to improve the overall mood.

In addition to the visualization of the multilevel-VAR analysis, the inferred networks open a range of new questions and possibilities that arise from network theory, and thus open a whole new research field. On the one hand, the local structure or specific connections can be studied with a local network analysis; on the other hand, the overall structure of the network can be studied with a global network analysis.

An example of a network analysis is node centrality as assessed through a betweenness measure. With this global network analysis we identified the most central node in a network for three groups with a different neuroticism level (low, mid and high). Our results revealed that in general, the node worrying was more central in the high neuroticism group than in the low or mid neuroticism group. This could be interpreted as indicating that worrying in general has a greater influence on the network in the high neuroticism group than in the low and mid neuroticism group. In the validation dataset there was no clear shift in worry in the high neuroticism group compared to the low and mid neuroticism group, but in this study worry was assessed slightly differently than in the main study, and furthermore, the population was different (college students instead of older subjects with residual depressive symptoms); in general, the subjects had lower neuroticism scores. More research is needed

to study the relation between neuroticism and node centrality in different kinds of populations.

Future research may focus on developing similar local and global network analyses, specifically suited for networks inferred from ESM data, and on evaluating the implications of these results. Thus, the presented methodology enables the use of network approaches in clinical research and open new possibilities to analyze and understand the structure of disorders, not only by inferring and visualizing the interaction between the variables, but also by further analyzing the new inferred networks.

Appendix 2.A Pseudo-likelihood method and simulation study

Pseudo-likelihood method

Fitting a combination of VAR (for modeling the intra-individual change) and multilevel modeling (for modeling inter-individual differences) comes with certain difficulties. Therefore, we apply a model fitting procedure that is similar to estimating a traditional (i.e., non-multilevel) VAR by fitting a series of multiple regression models (J. D. Hamilton, 1994). In our case, we fit a series of multilevel models, one for each item. Such an approach can be considered a specific case of the pseudo-likelihood method (Arnold & Strauss, 1991; Fieuws & Verbeke, 2006) in which not the likelihood itself is optimized to estimate the model's parameters, but rather an easier-to-calculate proxy to the likelihood (i.e., the pseudo-likelihood), which is constructed by considering a set of conditional and/or marginal densities. Our approach can be illustrated with a simple example (see also Arnold & Strauss, 1991): If one wants to estimate the parameters of a bivariate normal distribution (two means, two variances and a correlation) then one can estimate four out of five parameters (means and variances) by relying on the univariate marginals.

In this study, estimating the model's parameters is deferred to estimating the parameters from the marginal distributions of the six variables. As a result, the covariance matrix for random effects will not be estimated in a single step and not all of the covariance parameters will be estimated directly. Only eight-by-eight block matrices on the main diagonal from this general matrix pertaining to the same univariate multilevel analysis (i.e., $cov(b_{kpj}, b_{kpj'})$) are estimated (there are six such block matrices). The remaining covariances in the 48-by-48 matrix (related to covariances between random effects of different univariate models, i.e., $cov(b_{kpj}, b_{k'pj'})$) can be estimated in a subsequent step from the covariances between the predicted random effects. The error correlations, signifying the common disturbances to different variables, and the correlations between random effects of the different regression equations are not estimated in our approach. However, these parameters can be estimated in a second step by calculating the correlations between the level 1-residuals and level 2-residuals of the different univariate models. Relying on such an approach will probably lead to a small loss of efficiency compared to direct estimation. We show by means of the simulation study, described in the next section, that using our approach, the point estimates of most directly and indirectly estimated parameters are on average close to the true values.

Simulation study

Goal

In order to investigate the performance of the multilevel-VAR model in recovering the network structure for the type of data used in this paper, we performed a simulation study. To optimize validity of the simulation study, we simulated data based on the parameter estimates obtained from the empirical study and fitted the data with the procedure outlined above and in the main text of this article. Specifically, we took the estimates based on the results of the items cheerful and worry.

As indicated above, we did not fit the multilevel-VAR model by fitting the multivariate model at once, but instead by fitting a series of univariate multilevel models. In these models, several of the parameters can be estimated directly (i.e., all fixed effects and random regression coefficients, variances and covariance of random effects parameters within one model), but some of the parameters could only indirectly be estimated (i.e., the covariances between errors and the covariances between random effects that are in different univariate models). Through this simulation study, we aimed to show that a pseudo-likelihood fitting of the multilevel-VAR model yields a reasonable approximation of all parameters.

Data simulation model

A multilevel-VAR model with random intercept and slopes was used for the simulation. For reasons of computational tractability, we have reduced, without loss of generalizability, the original six-variable multilevel-VAR model to a bivariate model. The model equations are (for $j = 1, 2$; cheerful and worry respectively):

$$Y_{pdtj} = \gamma_{0pdj} + \gamma_{1pdj} \cdot Y_{p,d,t-1,1} + \gamma_{2pdj} \cdot Y_{p,d,t-1,2} + \varepsilon_{pdtj}, \tag{2.A.1}$$

where Y_{pdtj} represents the measurement for person p ($p = 1, 2, \ldots, 129$) at day d ($d = 1, 2, \ldots, 6$) at time t of the j-th variable. In addition, it is assumed that the regression coefficients can be decomposed as follows (for $j = 1, 2$), were β_{kj} represents the common effect of lagged variable k (for $k = 0$, this is the intercept) on the dependent variable j, and b_{kpj} is the person-specific deviation of this general effect:

$$\gamma_{0pdj} = \beta_{0j} + b_{0pj}, \tag{2.A.2}$$

$$\gamma_{kpj} = \beta_{kj} + b_{kpj}. \tag{2.A.3}$$

The intercepts (i.e., β_{01} and β_{02}) are set to 2.87 and 2.04, respectively. The other fixed effects parameters were fixed to $\beta_{11} = 0.28$, $\beta_{21} = -0.035$, $\beta_{12} = -0.048$ and $\beta_{22} = 0.26$. Where, for example, β_{21}, stand for the effect of worry on cheerful. The two error terms $(\varepsilon_{pt1}, \varepsilon_{pt2})$ follow a bivariate normal distribution with mean vector zero, variances of $\sigma^2_{\varepsilon_1} = 1.3$ and $\sigma^2_{\varepsilon_2} = 1.56$ respectively and a correlation of 0.4. The two random intercept components (b_{0p1}, b_{0p2}) come from a bivariate normal population distribution with zero mean vector, variances 1.2 and 1.1 respectively and a correlation of 0.4. The four component vectors of random regression weights $(b_{1p1}, b_{2p1}, b_{1p2}, b_{2p2})$ are multivariate normally distributed with zero mean vector, variances (0.0169, 0.00810, 0.000784, 0.0256) and correlation of 0.4 among all pairs of components. Note that the random intercepts and random regression weights are independent.

Design of the simulation study

In our simulation study, we manipulated the number of time points as follows: $T = 20$, 60, or 500. The number of participants was $N = 20$, 129, or 500. We did not cross the factors, but instead, we started from the settings of the empirical example (i.e., $N = 129$ and $T = 60$) and then manipulated either the number of time points or the number of participants separately. In every condition, the number of simulated data sets (i.e., replications) was 500.

Results of the simulation study

In all figures, the left plot indicates the three different settings for the sample size and the right plot the three different setting of the number of time-points. All the fixed effects regression coefficients (i.e., the β's referring to intercepts and regression weights) were estimated very accurately within all different settings (Figure 2.A.1 and 2.A.2). The variance of the errors and the variance of all person specific regression weights (i.e., the b's) including the intercept, are shown in Figures 2.A.3-2.A.5. From these plots, it can be seen that true point estimations of these parameters are accurate, often with less subjects or time points than used in the empirical study.

Figures 2.A.6 and 2.A.7 show how accurately the correlations between parameters of the models were estimated in an indirect way. This was done for the error correlation between the models (i.e., the correlation between: ε_{pdt1} and ε_{pdt2}), the random effects within one model (i.e., the correlation between: b_{1p1} and b_{2p1}), and the random effects between the two models (i.e., the correlation between: b_{0p1} and b_{0p2}; b_{1p1} and b_{1p2}; b_{1p1} and b_{2p2}; b_{2p1} and b_{1p2}; b_{2p1} and b_{2p2}). Figure 2.A.6 shows that although in the first model the correlation between the two random effects of the betas could be estimated quite accurately with 60 time points and 129 subjects, in the second model the correlation

between the random effects were estimated less accurately with 60 time points and 129 subjects than with more time points or subjects. There was also an estimation bias when the correlations of random effects (b's) between the models was estimated (Figure 2.A.7). However, the correlation of the error variances and random intercepts were estimated highly accurately, also between the two models (Figure 2.A.6). Thus, the random effects, except the random intercepts, were more difficult to estimate accurately and more subjects or time points are needed in that case. However, the model accurately estimated all the parameters that are of immediate relevance for this study.

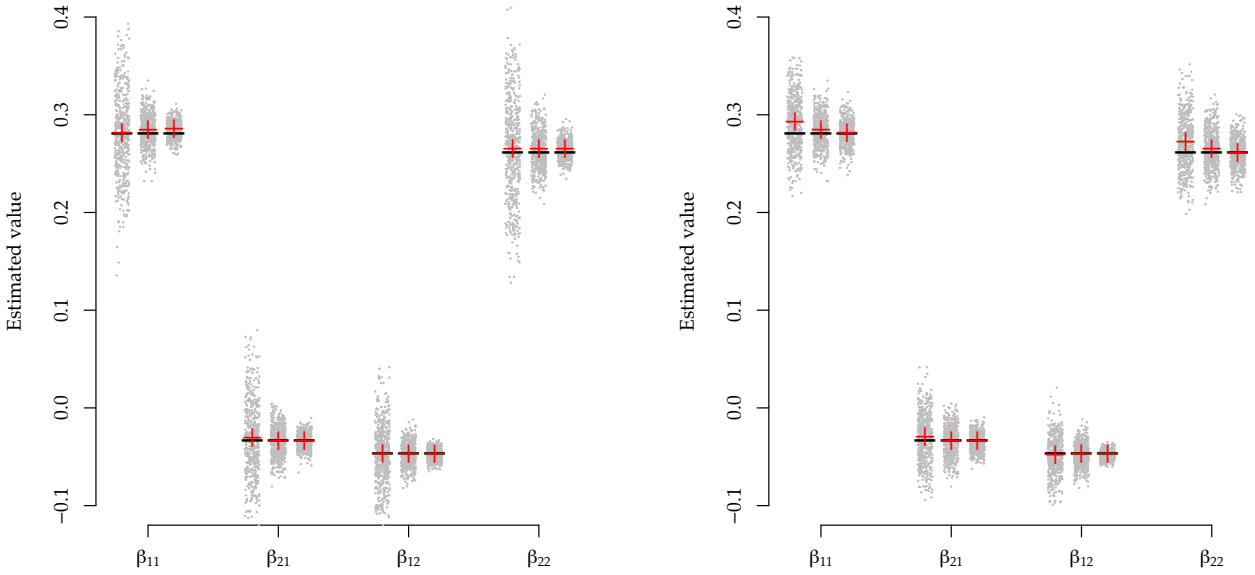


Figure 2.A.1: The recovery of the four average beta weights (β_{11},β_{21},β_{12},β_{22}) for a varying number of participants (right panel, with $T = 60$) and a varying number of time points (left panel, with $N = 129$). The black line indicates the true value, and the red cross indicates the average estimate (from 500 replications). The grey dots are the 500 individual estimates (jittered along the x-axis for visual understanding). The middle condition is always the setting corresponding to the empirical example, with 60 time points and 129 participants.

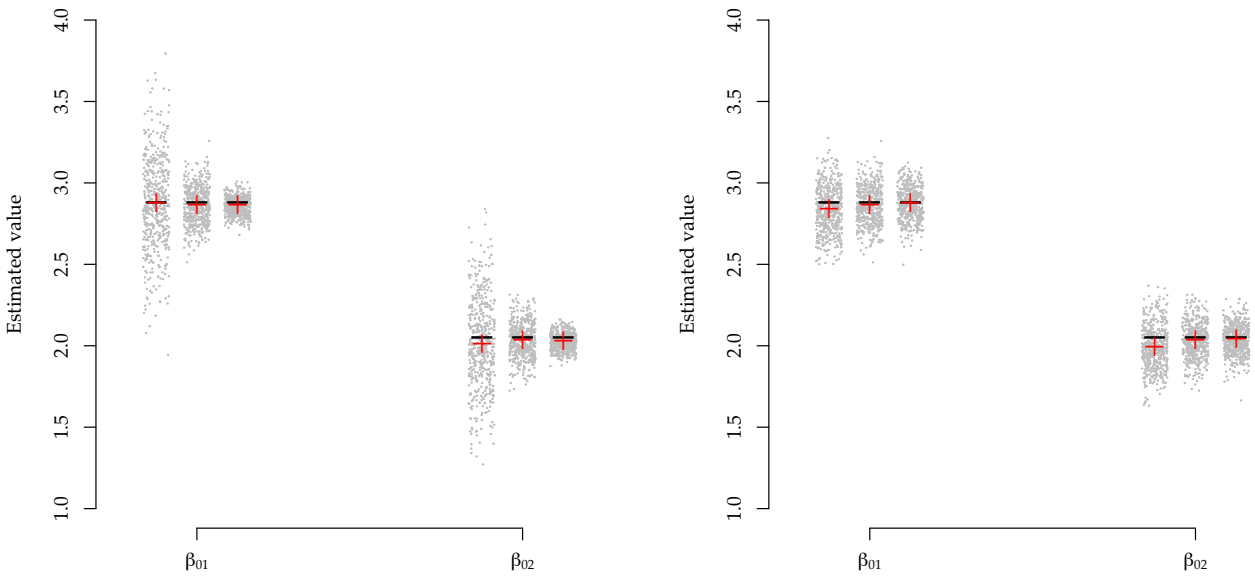


Figure 2.A.2: The recovery of the two average intercept coefficients (β_{01},β_{02}) for a varying number of participants (right panel, with $T = 60$) and a varying number of time points (left panel, with $N = 129$). The black line indicates the true value, and the red cross indicates the average estimate (from 500 replications). The grey dots are the 500 individual estimates (jittered along the x-axis for visual understanding). The middle condition is always the setting corresponding to the empirical example, with 60 time points and 129 participants.

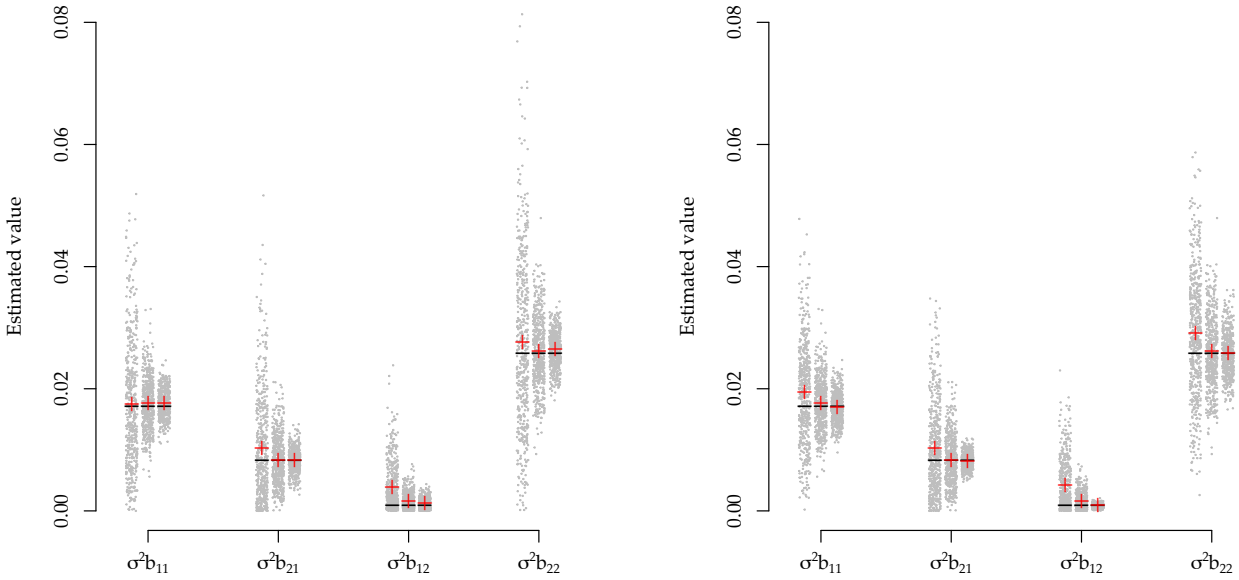


Figure 2.A.3: The recovery of the four variances of the person specific regression weights (b_{11},b_{21},b_{12},b_{22} see Equation 2.8) for a varying number of participants (right panel, with $T = 60$) and a varying number of time points (left panel, with $N = 129$). The black line indicates the true value, and the red cross indicates the average estimate (from 500 replications). The grey dots are the 500 individual estimates (jittered along the x-axis for visual understanding). The middle condition is always the setting corresponding to the empirical example, with 60 time points and 129 participants.

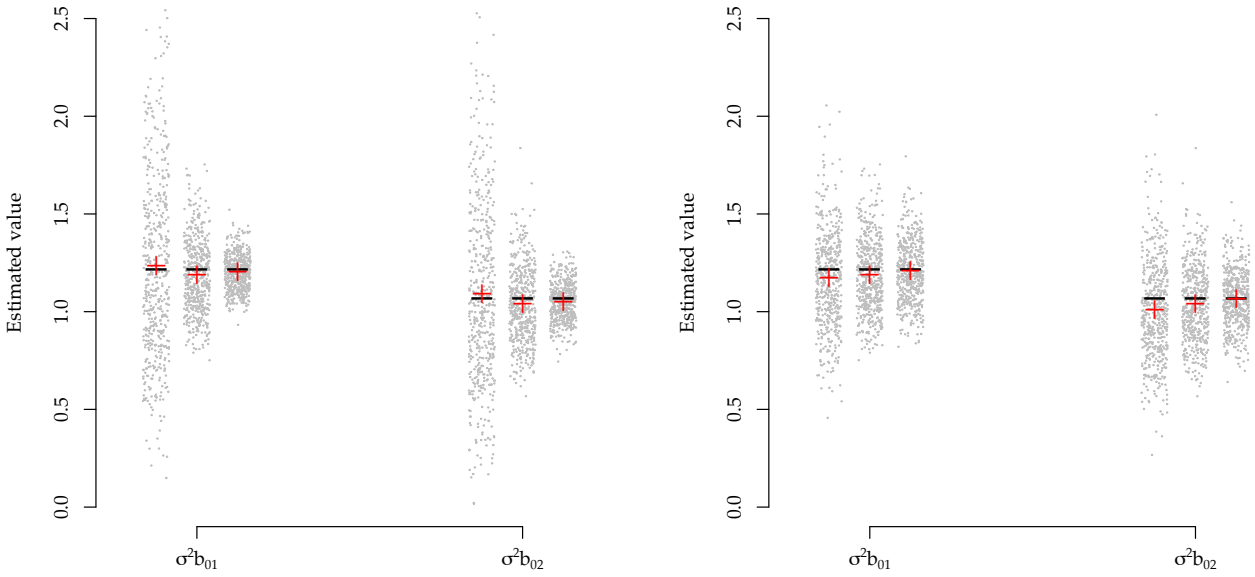


Figure 2.A.4: The recovery of the two variances of the person specific intercepts (b_{01},b_{02}; see Equation 2.7) for a varying number of participants (right panel, with $T = 60$) and a varying number of time points (left panel, with $N = 129$). The black line indicates the true value, and the red cross indicates the average estimate (from 500 replications). The grey dots are the 500 individual estimates (jittered along the x-axis for visual understanding). The middle condition is always the setting corresponding to the empirical example, with 60 time points and 129 participants.

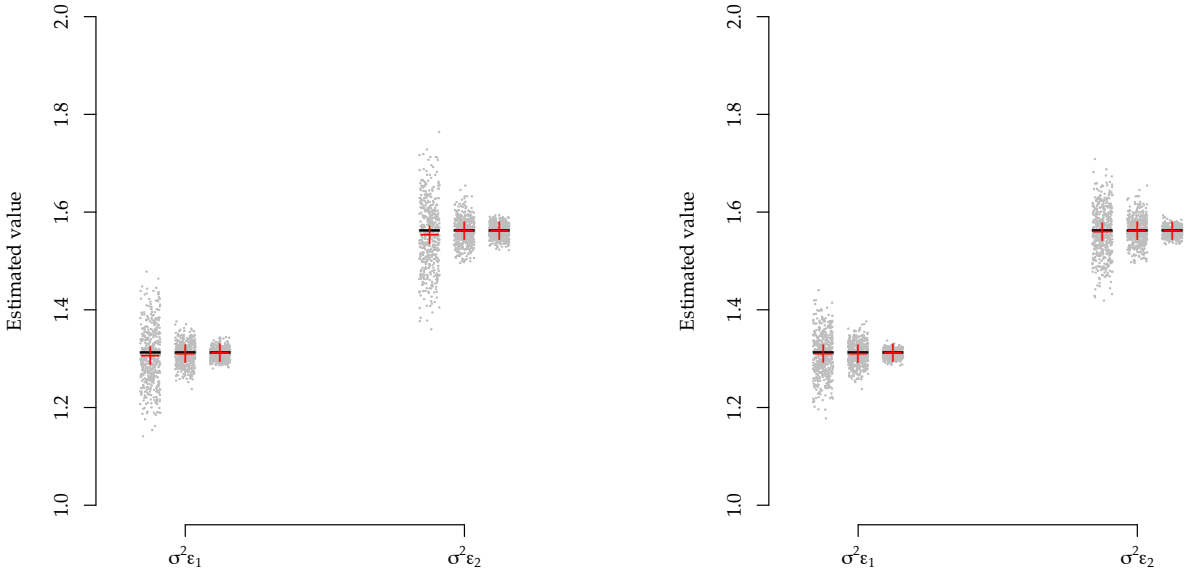


Figure 2.A.5: The recovery of the variances of the two error terms (ε_1 and ε_2) for a varying number of participants (right panel, with $T = 60$) and a varying number of time points (left panel, with $N = 129$). The black line indicates the true value, and the red cross indicates the average estimate (from 500 replications). The grey dots are the 500 individual estimates (jittered along the x-axis for visual understanding). The middle condition is always the setting corresponding to the empirical example, with 60 time points and 129 participants.

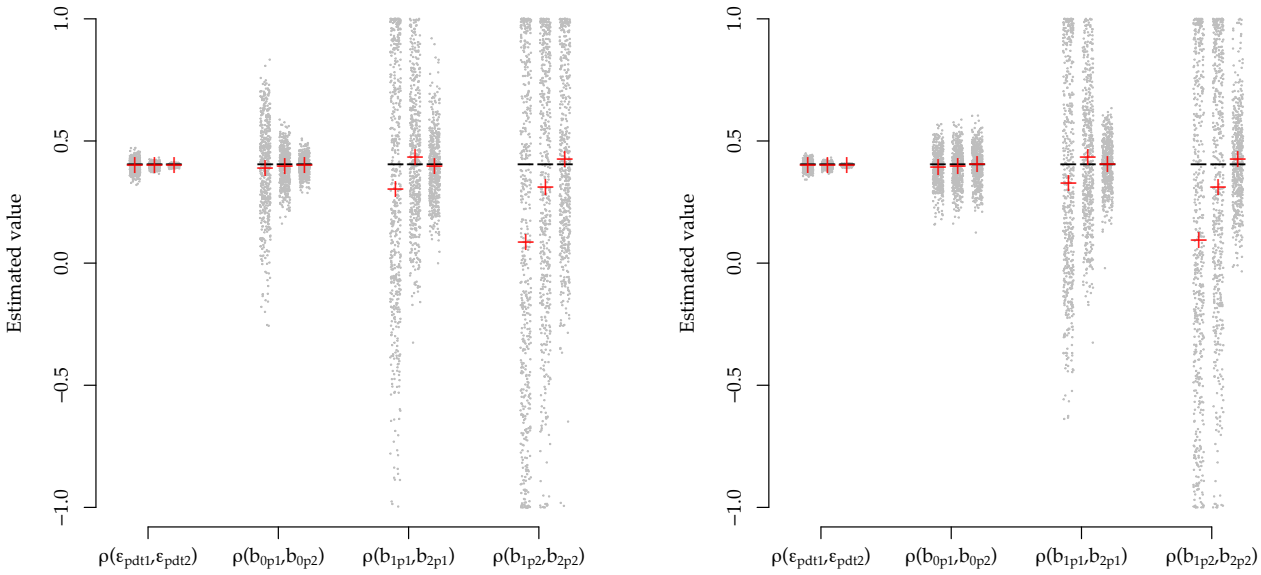


Figure 2.A.6: The recovery of the error correlations, the random intercept correlations and the correlations of the random effects within model 1 and 2, respectively. The simulation was done for a varying number of participants (right panel, with $T = 60$) and a varying number of time points (left panel, with $N = 129$). The black line indicates the true value, and the red cross indicates the average estimate (from 500 replications). The grey dots are the 500 individual estimates (jittered along the x-axis for visual understanding). The middle condition is always the setting corresponding to the empirical example, with 60 time points and 129 participants.

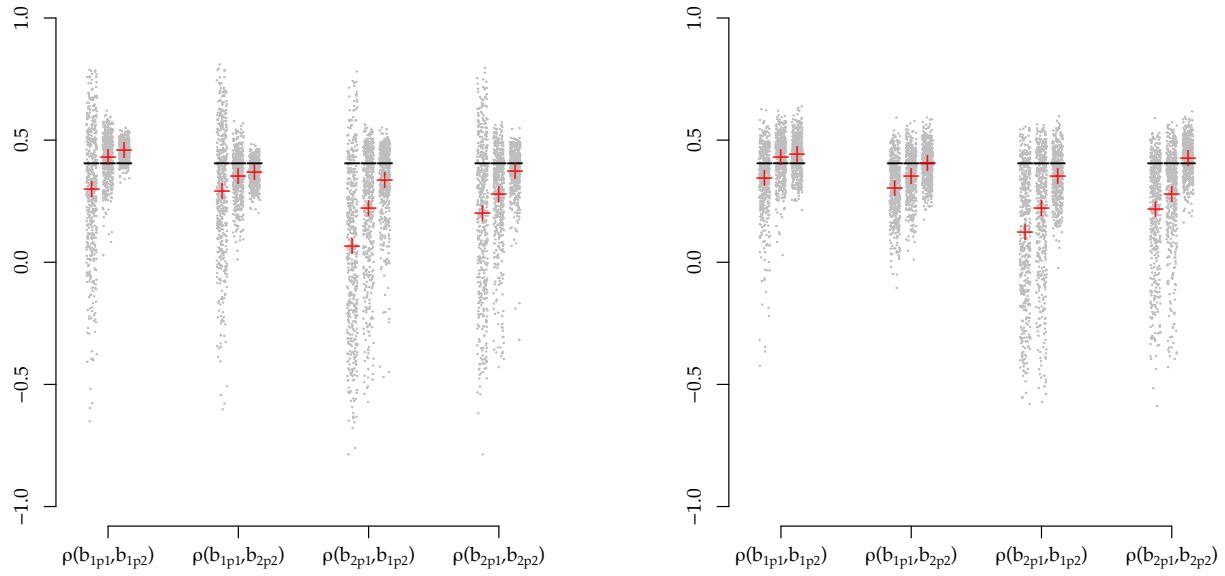


Figure 2.A.7: The recovery of correlations of random effects between the models of cheerful and worry. The simulation was done for a varying number of participants (right panel, with $T = 60$) and a varying number of time points (left panel, with $N = 129$). The black line indicates the true value, and the red cross indicates the average estimate (from 500 replications). The grey dots are the 500 individual estimates (jittered along the x-axis for visual understanding). The middle condition is always the setting corresponding to the empirical example, with 60 time points and 129 participants.

Appendix 2.B Figures of the replication study

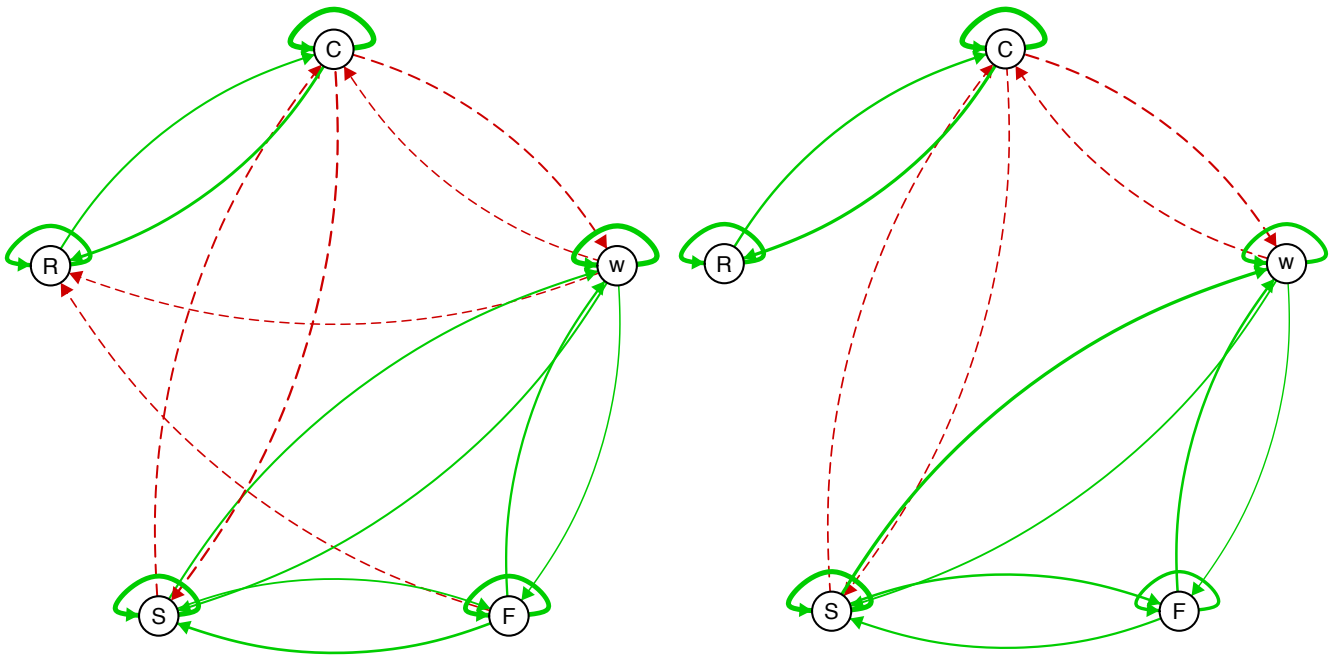


Figure 2.B.1: This figure shows the population network of the main data set (left panel) and the validation dataset (right panel) for five items (C=cheerful, W=worry, F=fearful, S=sad and R=relaxed). Solid green arrows correspond to positive arrows and red dashed arrows to negative connections. Only arrows that surpass the significance threshold are shown (i.e., for which the p-value of the t-statistic is smaller than 0.05). Arrows can be either red, indicating a negative relationship (i.e., $\beta < 0$), or green, indicating a positive relationship (i.e., $\beta > 0$). The two networks are almost identical (Pearson correlation of 0.95, see main text for more information). However, more links are significant in the main dataset and are shown in the figure than in the validation dataset (using FDR controlled at 5%). This is likely to be due to the fact that there are more subjects (129 vs. 97) and more time points (70 vs. 53) in the main dataset than in the validation dataset.

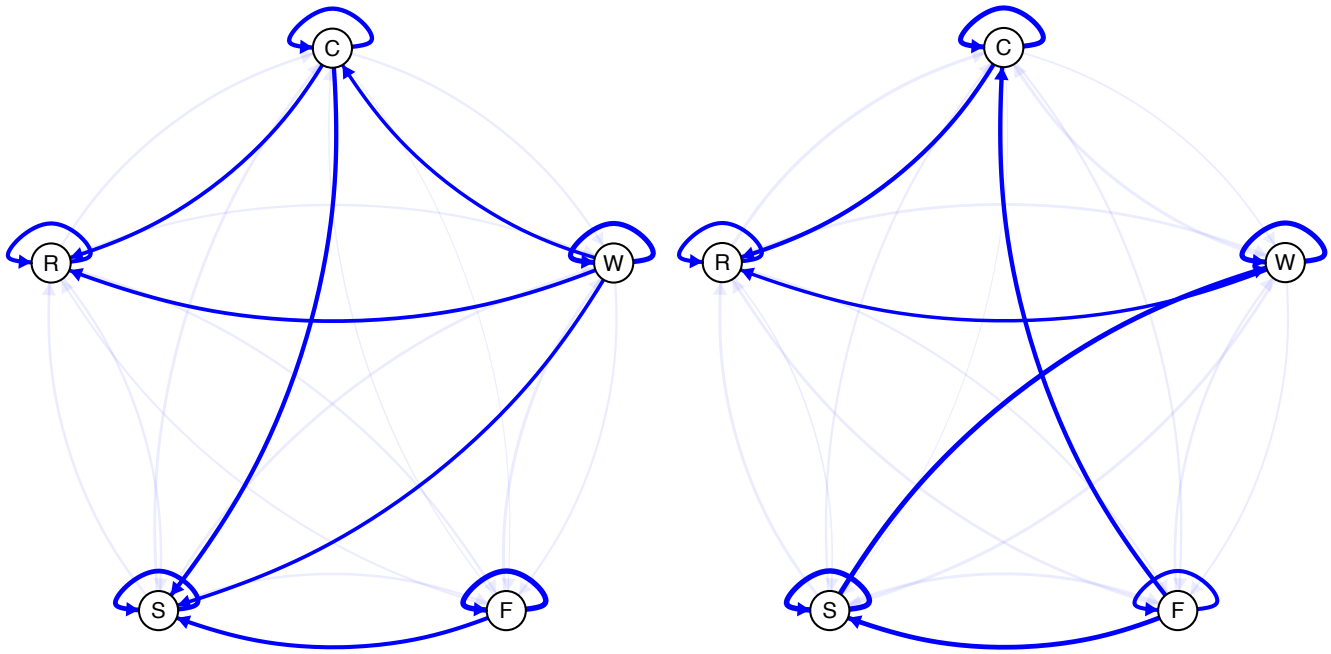


Figure 2.B.2: This figure shows the inter-individual network of the main dataset (left panel) and the validation dataset (right panel) for five items (C=cheerful, W=worry, F=fearful, S=sad and R=relaxed). The thickness of the arrows is based on the size of the standard deviation of the random effects. To construct the figure, we have put a cutoff of 0.1 on the standard deviation and only the standard deviations above the cutoff are shown with a non-transparent arrow.

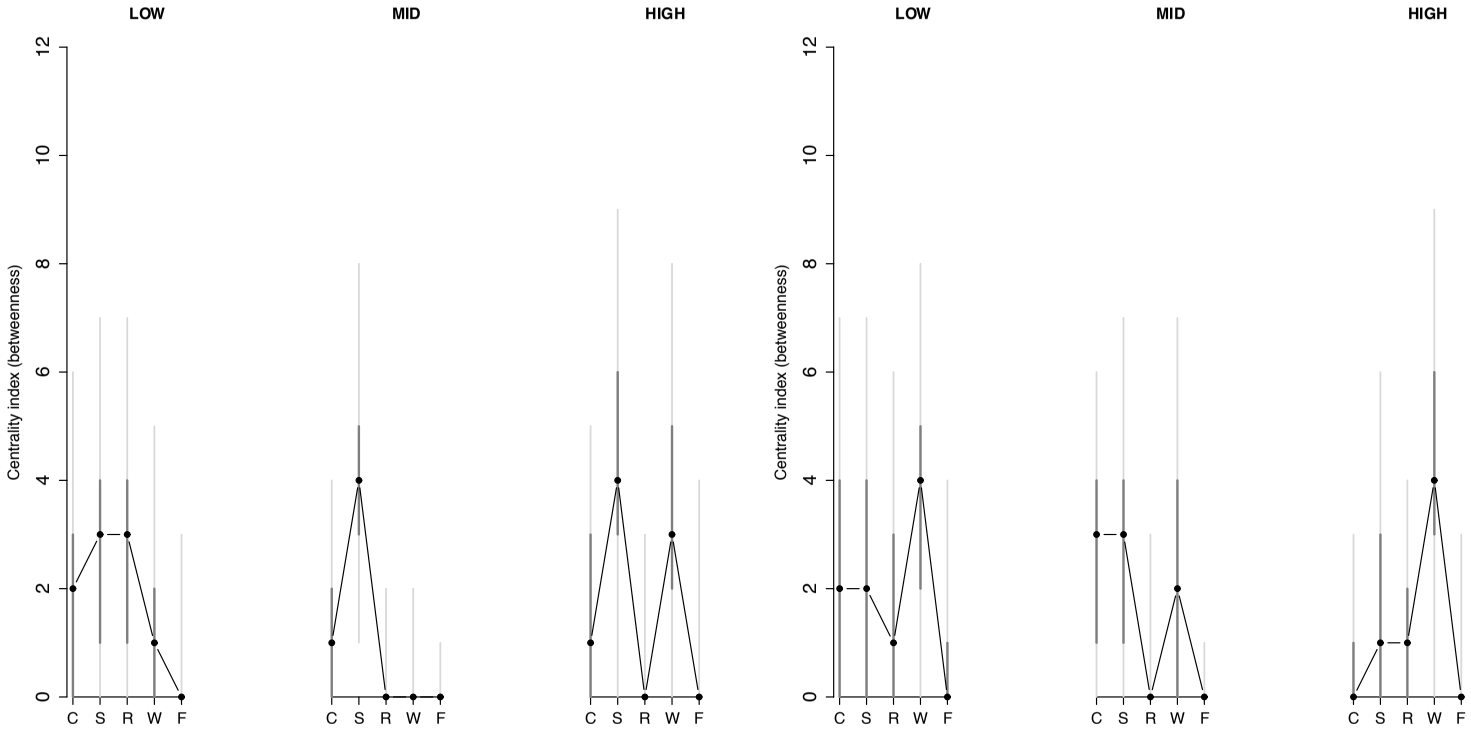


Figure 2.B.3: For both the main dataset (left panel) and the validation dataset (right panel) the centrality index (betweenness) of each item in the network as a function of level of neuroticism (low, mid, and high neuroticism are shown from left to right) at baseline are shown. The labels of the items are abbreviated by their first letter (C=cheerful, S=sad, R=relaxed, W=worry, F=fearful). The black dots are the model-based estimates of betweenness, the darkgrey vertical lines represent 50% confidence intervals and the light grey vertical lines represent 95% confidence intervals (as estimated from the bootstrap method). Together, the median, 50% and 95% confidence intervals give information on how the node centrality for every item in all three networks is distributed.

3 Assessing temporal emotion dynamics using networks

Experience sampling methods (ESM; Csikszentmihalyi & Larson, 2014; Trull & Ebner-Priemer, 2013) and ecological momentary assessment (EMA; Shiffman & Stone, 1998; Stone & Shiffman, 1994) are being increasingly used to study dynamic psychological processes such as mood (aan het Rot et al., 2012; Hamaker, Ceulemans, Grasman, & Tuerlinckx, 2015; Jahng, Wood, & Trull, 2008; Wichers, Wigman, & Myin-Germeys, in press). A particularly relevant aspect thereof is their temporal dynamics (Nesselroade, 2004). When studying temporal dynamics, the focus is not on detecting a gross underlying trend, as is often the case in developmental research, but rather on the intricate temporal dependence of and between variables, or how variables within an individual influence each other or themselves over time (Brandt & Williams, 2007; Molenaar, 1985; Walls & Schafer, 2006). Often the models used to study temporal dynamics are multivariate in nature, and both the influence that a variable has on itself (e.g., how self-predictive is sad mood) as well as its effects on other variables (e.g., how does sad mood augment or blunt subsequent anger emotions) are analyzed (Koval, Pe, Meers, & Kuppens, 2013; Kuppens, Stouten, & Mesquita, 2009; Kuppens, Allen, & Sheeber, 2010; Pe & Kuppens, 2012; Suls, Green, & Hillis, 1998).

One increasingly popular approach to study, visualize, and analyze multivariate dynamics is network analysis (Borsboom & Cramer, 2013; Bringmann, Vissers, et al., 2013; Bringmann, Lemmens, Huibers, Borsboom, & Tuerlinckx, 2015; Fried et al., 2014; McNally et al., 2015; Ruzzano, Borsboom, & Geurts, 2015; Wichers, 2014). This network perspective leads to a new way of thinking about the nature of psychological constructs, phenomena or processes by offering new tools for studying dynamical processes in psychology. In the network approach, psychological constructs, processes or phenomena are represented as complex systems of interacting components (Barabási, 2011; Costantini et al., 2015; Cramer, Borsboom, et al., 2012). For instance, emotional well-being can be considered to consist of a number of dynamically interacting components, such as behavioral, physiological, and experiential emotion components. Likewise, mental disorders can be viewed as a result of the mutual interplay of symptoms of the disorder. These components interact with each other across time, making up the internal dynamics and by that, the very nature of the phenomenon under study. It is

these dynamics that are studied in a network approach (Borsboom et al., 2011; Cramer et al., 2010; Schmittmann et al., 2013). In this paper, we will illustrate the network approach using an empirical example focusing on the relation between the daily fluctuations of emotions and neuroticism.

The network approach

A network consists of nodes (i.e., the components of the phenomenon, construct or process) and edges (or links) connecting the nodes (Barrat, Barthelemy, Pastor-Satorras, & Vespignani, 2004). In our approach, the links have a certain strength that indicates the strength of the (positive or negative) relationship between the nodes (Opsahl et al., 2010). The nodes and edges can be easily visualized graphically (see for example Figure 3.1). Networks can be constructed based on different kinds of data such as cross-sectional or longitudinal data and using different kinds of models for inferring the edges. Depending on the data and model used to infer the network, the edges connecting the nodes have a specific meaning. In this article, we focus on longitudinal data and on the vector autoregressive (VAR) model (Brandt & Williams, 2007). A VAR-based network allows studying the dynamics among the components that constitute a certain construct, phenomenon or process across time. For example, in the network of Figure 3.1, the edges on the nodes are the self-loops, or the effect the emotion has on itself from one time point to the next, and the edges between the emotions are the cross-regressive effects, or the effect a variable has on other variables from one time point to the next, controlling for the other variables.

In addition, several features of the network can be derived that can shed light on central properties of the dynamical interplay between the components or nodes. Such features can involve the overall network or specific parts of the network. One interesting characteristic of the overall network is its density, which indicates how strongly the network is interconnected. The denser a network is, the more strongly the variables interact (Newman, 2010). Another, more specific, feature of the network is node centrality. Centrality refers to the importance or how focal one specific variable or node is in the network (Freeman, 1978).

Empirical example

We will illustrate how networks can be inferred using a multilevel extension of the VAR model (Bringmann, Vissers, et al., 2013), and how they can be used to gather new insights on temporal emotion dynamics. In particular, we will focus on the relation between emotion dynamics and neuroticism in healthy subjects, using two previously collected ESM datasets. Neuroticism is one of the main dimensions reflecting individual differences in personality, and is particularly relevant for emotional experience. Specifically, it reflects a tendency to experience negative emotions, and is considered

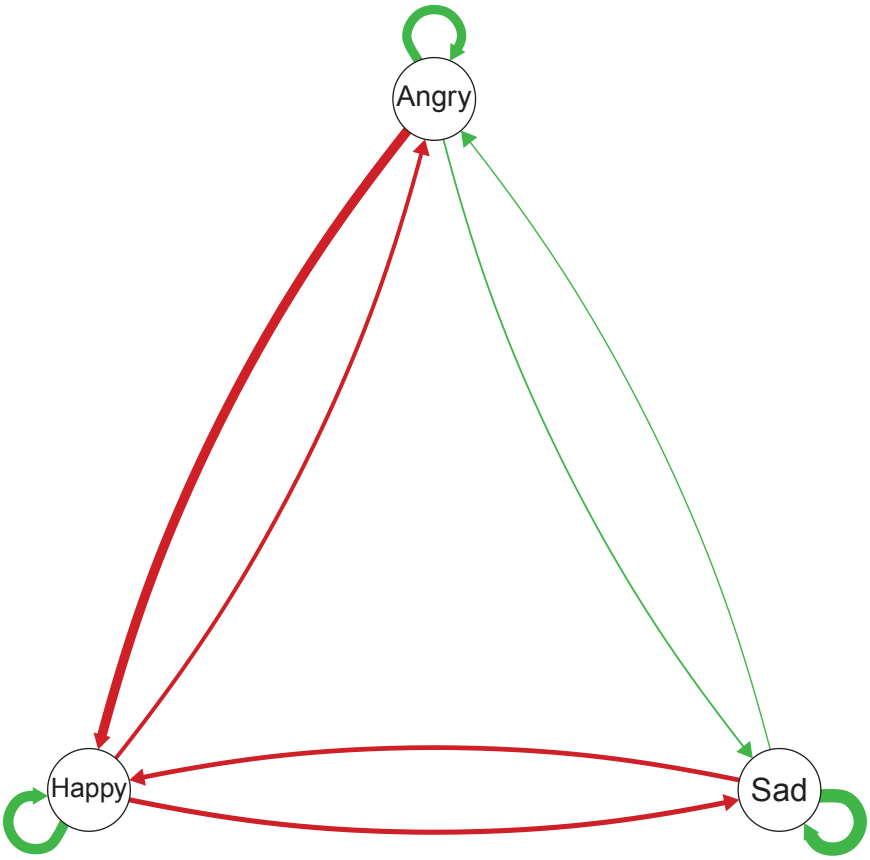


Figure 3.1: *A hypothetical example of an emotion network.* The three nodes are the three emotions: *Happy*, *Angry* and *Sad*. The red arrows are the negative (i.e., inhibitory) edges and the green arrows the positive (i.e., excitatory) edges. The thickness of the arrows represents the strength of the edges. For example, the edges on the nodes (the self-loops) are the strongest links in the network.

to constitute a broad risk factor for mood disorder and psychopathology (Barlow, Sauer-Zavala, Carl, Bullis, & Ellard, 2014).

In this application, we will first look at the general patterns of edges connecting the emotion variables, which are referred to as the population networks. Second, we will assess features of the network structure by studying the density of the individual emotion networks and their relation to neuroticism. In a third step, we will study whether several centrality measures of the individual networks (strength, closeness and betweenness) and the self-loops are related to neuroticism. To our knowledge, this is the first time that both the full temporal emotion network and its parts are studied and related to neuroticism, giving a more complete picture of moment-to-moment dynamics in emotion as a function of the trait of neuroticism. The method used here will be described in detail. Moreover, *Matlab* and *R* code to replicate the main results of the first dataset will be given, so that

other researchers can apply the network method to their own data (see online material).[1]

3.1 Method

Dataset 1

Parts of dataset 1 have been published elsewhere (Bringmann, Vissers, et al., 2013; Koval et al., 2012; Pe, Koval, & Kuppens, 2013; Pe, Raes, et al., 2013). 95 undergraduate students from the University of Leuven in Belgium (age: $M = 19$ years, $SD = 1$; 62% female) participated in an experience sampling method (ESM) study. Over the course of seven days, participants carried a palmtop computer on which they had to fill out questions about mood and social context in their daily lives 10 times a day. Participants were beeped to fill out the ESM questionnaires at random times within 90-minute windows. They had to rate, among other things, their current feelings of negative and positive emotions on a continuous slider scale, ranging from 1 (*not at all, e.g., angry*) to 100 (*very, e.g., angry*). On average, participants responded to 91% of the beeps ($SD = 7$ %). In order to avoid selection bias, we analyzed all six emotion variables measured in this study (positive affect: relaxed and happy; negative affect: dysphoric, anxious, sad and angry), which were selected to capture all quadrants of the affective circumplex defined by the dimensions of valence and arousal (see e.g., Russell, 2003). Furthermore, neuroticism was assessed with the Dutch version of the Ten Item Personality Inventory (Gosling et al., 2003; Hofmans et al., 2008), resulting in a score ranging from 1 to 7 ($M = 3.4$; $SD = 1.5$). Participants were selected from a large pool of participants to ensure a wide range of depression scores. Therefore, the participants in this dataset have a wider range of neuroticism scores than the participants in dataset 2.

Dataset 2

Parts of this dataset have been published elsewhere (Kuppens, Champagne, & Tuerlinckx, 2012; Kuppens, Oravecz, & Tuerlinckx, 2010; Pe & Kuppens, 2012). In this study, the participants consisted of 79 undergraduate students from the University of Leuven in Belgium (age: $M = 24$, $SD = 8$; 63 % female). A similar ESM procedure as in the first dataset was used. Participants were beeped to fill out the ESM questionnaires 10 times a day, again on a scale ranging from 0 to 100, but for a longer time period, namely 14 consecutive days. We extracted all emotion variables, which were 10 in this case (positive affect: relaxed, happy, satisfied, excited; negative affect: dysphoric, anxious, irritated, sad, stressed and angry), again selected to cover all quadrants of the affective space. Participants responded on average to 82% of the programmed beeps ($SD = 10$). Neuroticism was assessed with

[1] To use this code please read first the *R*-file.

the 12-item scale of the Dutch version of the NEO Five-Factor Inventory (Hoekstra et al., 1996), which resulted in a score ranging from 1 to 5 ($M = 3.0$, $SD = 0.7$).

Estimating the networks

To assess temporal emotion dynamics and their relation to neuroticism, an emotion network was created for each individual. The edges or links of the individual networks were obtained using a multilevel VAR model (Bringmann, Vissers, et al., 2013; Bringmann et al., 2015). The standard VAR model (Brandt & Williams, 2007) estimates the extent to which a current emotion (time point t) can be predicted from all other emotions at a previous moment (time point $t-1$), corresponding to the network edges. Each emotion is regressed on its lagged values (autoregressive effect) and the lagged values of each of the other emotions (cross-lagged effects). In the present context, time $t-1$ and time t refer to two consecutive beeps within the same day (overnight lags were removed). It is assumed that the data are stationary, implying that the mean and the moment-to-moment interactions of the emotion processes stay stable over time (Chatfield, 2003; Hamaker & Dolan, 2009). As we study multiple individuals, we implement the VAR model within a multilevel modeling framework, to allow for random, person-specific auto- and crossregressive effects, and so that we can model the temporal emotion dynamics not only within an individual, but also at group level, estimating both average or population (fixed) and individual (random) effects.

Univariate multilevel VAR analyses are conducted for each emotion separately using restricted maximum likelihood estimation. This results in 6 univariate regression equations for the first dataset and 10 univariate regression equations for the second dataset. Taking the first dataset with 6 emotions as an example, we get the following equation for each emotion j (i.e., relaxed, happy, dysphoric, anxious, sad and angry, or $j = 1, \ldots, 6$, respectively):

$$\begin{aligned} Y_{ptj} = \gamma_{0pj} + \gamma_{1pj} \cdot relaxed_{p,t-1} + \gamma_{2pj} \cdot happy_{p,t-1} + \gamma_{3pj} \cdot dysphoric_{p,t-1} \\ + \gamma_{4pj} \cdot anxious_{p,t-1} + \gamma_{5pj} \cdot sad_{p,t-1} + \gamma_{6pj} \cdot angry_{p,t-1} + \varepsilon_{ptj}. \end{aligned} \tag{3.1}$$

Thus, for dataset 1, Y_{ptj} represents the value for the j-th emotion for person p ($p = 1, 2, \ldots, 95$) at beep t ($t = 2, \ldots, 10$). The regression coefficients (i.e., the intercept and the regression weights) of this equation 3.1 are decomposed as follows:

$$\gamma_{kpj} = \beta_{kj} + b_{kpj}, \tag{3.2}$$

where the slopes β_{kj} ($k > 0$, since $k = 0$ codes for the intercept) represent the fixed effects (the edges in the network), or the extent to which the emotions at time $t-1$ can predict the emotion j

at time t over all individuals. The person-specific deviation (random effect) from the average effect is captured in the component b_{kpj}. The random effects are assumed to come from a multivariate normal distribution, estimating an unstructured covariance matrix of the random effects. Using the empirical Bayes estimates of the random effects, emotion networks for each individual are constructed. Specifically, for each edge in the network, the individual random effect is added to the fixed effect for each emotion variable. For instance, the edge from emotion k to emotion j has a value of $\gamma_{kpj} = \beta_{kj} + b_{kpj}$ in the individual network of person p. To reduce the likelihood of errors in the analyses, all multilevel analyses were run in Matlab (Mathworks, Inc.) as well as in Mplus (Muthén & Muthén, 2012) and by different researchers. Visualization and computation of the measures of centrality relied on the *qgraph R* package (Epskamp et al., 2012).

Regarding the analysis, there are three important additional aspects to mention here. First, as we estimate multivariate networks with both autoregressive and cross-lagged effect, all predictors were person-mean centered (centered around each individual's mean score) before the analysis (Hamaker & Grasman, 2014). Note that this might lead to a slight underestimation of the autoregressive effects. Second, to control for differences in variability between individuals, i.e. to make sure that associations between neuroticism and network characteristics were not driven by differences in emotion variance, we conducted analyses involving both non-standardized and standardized coefficients.[2] Within-person standardization of the coefficients was done as described in Schuurman, Ferrer, de Boer-Sonnenschein, and Hamaker (2016).[3] Third, note that the edges only represent the unique direct effects of the variables and not the shared effects (just as in standard multiple regression; Bulteel, Tuerlinckx, Brose, & Ceulemans, in press). This means that a part of the explained variance cannot be taken into account and thus an edge might be less strong or stronger if this shared variance was taken into account.

3.2 Network analyses

The population networks

Before we focus on individual networks and their relationship to neuroticism, we will first look at the average networks. These population networks show the general patterns of connections between the emotion variables. The edges in the population networks represent the slopes β_{kj} ($k > 0$; i.e., the fixed effects). The population networks are presented in Figure 3.2, made with the R-package qgraph (Epskamp et al., 2012).

[2] One exception is the analyses using self-loops. In order to standardize the edges of the network, the standard deviations of the predictor and outcome variables are used. Since a self-loop has the same predictor as outcome variable the standardized and unstandardized edges are equal.

[3] Note that there are different ways to standardize that lead to slightly different results.

Density

For each individual network, the density was computed of 1) the overall network (all emotions), 2) the negative emotions only and 3) positive emotions only. This was done by averaging over the absolute values of the slopes or edges in the network of the emotions of interest. We used the absolute values so that negative and positive edge values do not cancel each other out.

Further, to illustrate the relation between density and neuroticism, we created three neuroticism groups (i.e., low, medium and high neuroticism) by ranking the neuroticism scores. In a next step, we constructed networks for the low and high neuroticism group separately (eventually resulting in two networks for overall, negative and positive emotion density for both datasets). If we focus on the overall network for simplicity of explanation, then the arrows indicate the edge strengths of the temporal connections between emotions. The average absolute value of the edge strength and the corresponding standard deviation (SD) is calculated across all participants and pairs of variables. Next, edges get classified: $1SD$ below the mean (weak connection strength, dotted arrows), between $1SD$ below and above the mean (moderate connection strength, dashed arrows) and $1SD$ above the mean (strong connection strength, solid arrows).

Centrality

We calculated the most common centrality measures degree (or in case of a weighted network the term strength is used), closeness and betweenness. Each centrality measure defines centrality of a node (variable) in the network in a different way (Freeman, 1978; Newman, 2010). To explain these concepts, it is instructive to think metaphorically that the nodes transmit information across time to each other. As the network used here is a directed network, we can study both the out-strength centrality and the in-strength centrality. Out-strength indicates the (summed) strength of the outgoing edges or how much information a node sends away to the other nodes, and thus a node with a high out-strength centrality tends to excite or inhibit many other nodes in the network. In-strength indicates the strength of the incoming edges, or how much information a node receives from the other nodes, and thus its susceptibility to being excited or inhibited by other nodes in the network.[4] Both out- and in strength take only into account the edges to which a node is directly connected.

A node high in closeness centrality is at a relatively short distance from the other nodes in the network, and is thus likely to be influenced quickly by them. Closeness thus represents how fast an emotion can be reached from the other nodes in the network. Distances between nodes are calculated based on edge strength, taking into account direct and indirect edges connecting the node to other nodes (See for more information: Borgatti, 2005; Costantini et al., 2015; Opsahl et al., 2010).

[4]We thank an anonymous reviewer for suggesting this interpretation.

Betweenness centrality is a measure of how many times a node appears on the shortest paths between other nodes in the network. Thus, a node with a high betweenness centrality is a node through which the information in the network has to pass often and can be seen as an important node in funneling the information flow in the network. This measure also takes into account direct and indirect edges connecting the node to other nodes. Note that all the centrality measures are based on the absolute values of the edges.

The relation between the network characteristics and neuroticism

Neuroticism scores of all individuals were correlated with density of the individual networks (calculated on the overall, negative and positive networks) and centrality measures (out-strength, in-strength, closeness and betweenness) using Pearson's product moment correlations. Since the centrality measures are concerned with the influences between variables or nodes (cross-regressive effects) in the network, self-loops or autoregressive effects (in the emotion literature also known as emotional inertia; Suls et al., 1998) are ignored in these focal network measures. Therefore, the correlation between the self-loops and neuroticism was calculated separately for each emotion.

3.3 Results

The networks in Figure 3.2 represent the average patterns between the emotions. Only edges that were significant (i.e., a p-value of less than 0.05) are shown, which is purely for visualization purposes. The figures show that emotions can either augment or blunt each other (Pe & Kuppens, 2012). Augmenting refers to the increase of the experience of other emotions. For example, there exist clusters of negative and positive emotions. Within these clusters, emotions of the same valence tend to in general augment each other. In contrast, emotions of different valence (for example, sad and happy) seem to blunt or decrease each other. Furthermore, the self-loops in the networks are among the strongest edges. For example, in general when a person feels sad, (s)he is not only less likely to feel happy at the next moment, but also likely to still experience sadness at the next moment.[5] These results correspond with the theoretical expectations and empirical findings based on the nomothetic relations in an emotion circumplex, namely that emotions of the same valence are more likely to be correlated with each other than with emotions of different valence (Vansteelandt, van Mechelen, & Nezlek, 2005).

The results in Table 3.1 show a consistent and strong positive relation between neuroticism and overall emotion density as well as negative emotion density. This pattern is not only consistent across datasets, but also when controlling for variability (i.e., after standardization), indicating that

[5]Note that the number of possible edges is proportional to the number of nodes and thus the network for dataset 2 is not necessarily more strongly connected than the network for dataset 1.

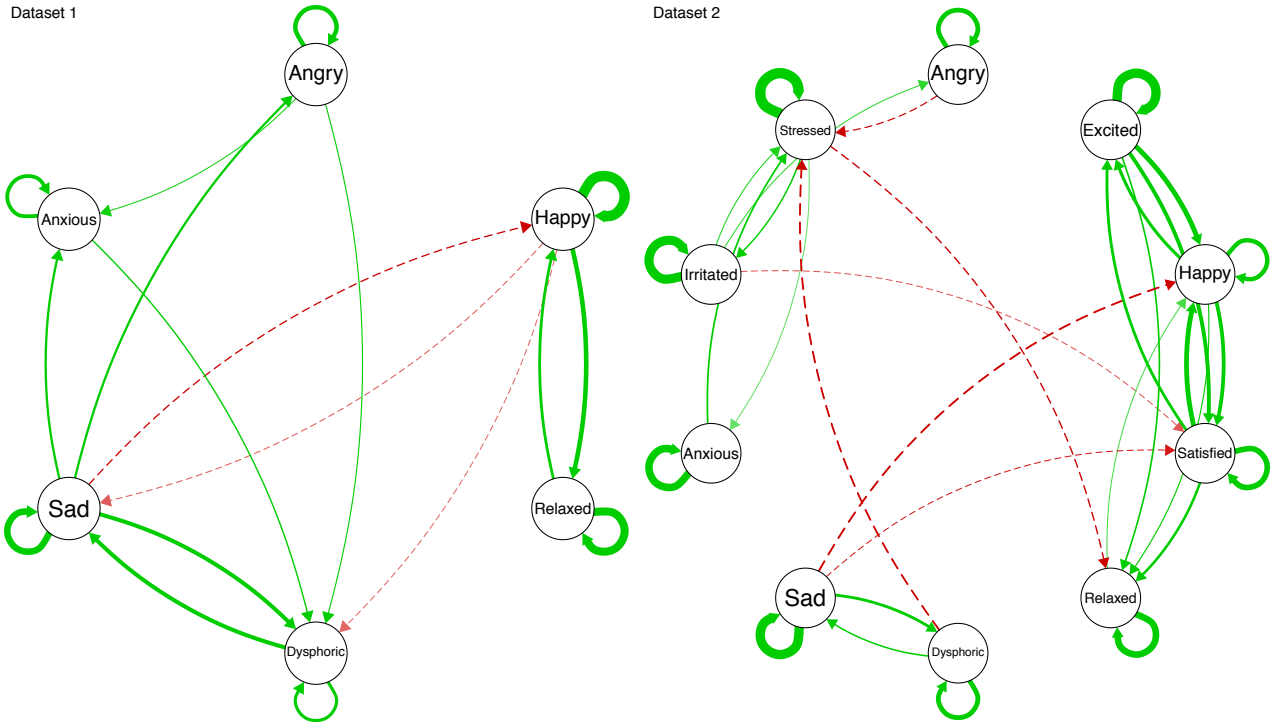


Figure 3.2: *This figure shows the population network of the dataset 1 (left panel) and the dataset 2 (right panel).* Solid green edges correspond to positive and dashed red edges to negative connections. Only edges that surpass the significance threshold are shown (i.e., for which the p-value of the t-statistic is smaller than 0.05). The emotions in the networks are organized so that they align with the emotion circumplex from which they were selected.

individuals high in neuroticism also have a significantly denser overall network and negative emotion network than individuals low in neuroticism. The results for the positive emotion network were less consistent. The relation between the positive emotion network and neuroticism was only significant in the second dataset and was less strong than the relationship between neuroticism and the overall and negative emotion networks. Figures 3.3 and 3.4, focusing on the high and low ends of neuroticism, also features this pattern: The difference between emotion density in individuals with a high and low score in neuroticism is more pronounced for the overall emotion density and negative emotion density than for positive emotion density.

Table 3.1: Density and its relation to neuroticism

	Non-standardized				*Standardized*			
	Data 1		*Data 2*		*Data 1*		*Data 2*	
Emotion Network	r	p	r	p	r	p	r	p
Overall	.49	<.001	.42	<.001	.49	<.001	.41	<.001
Negative	.51	<.001	.44	<.001	.51	<.001	.43	<.001
Positive	.12	.26	.30	.008	.11	.27	.30	.007

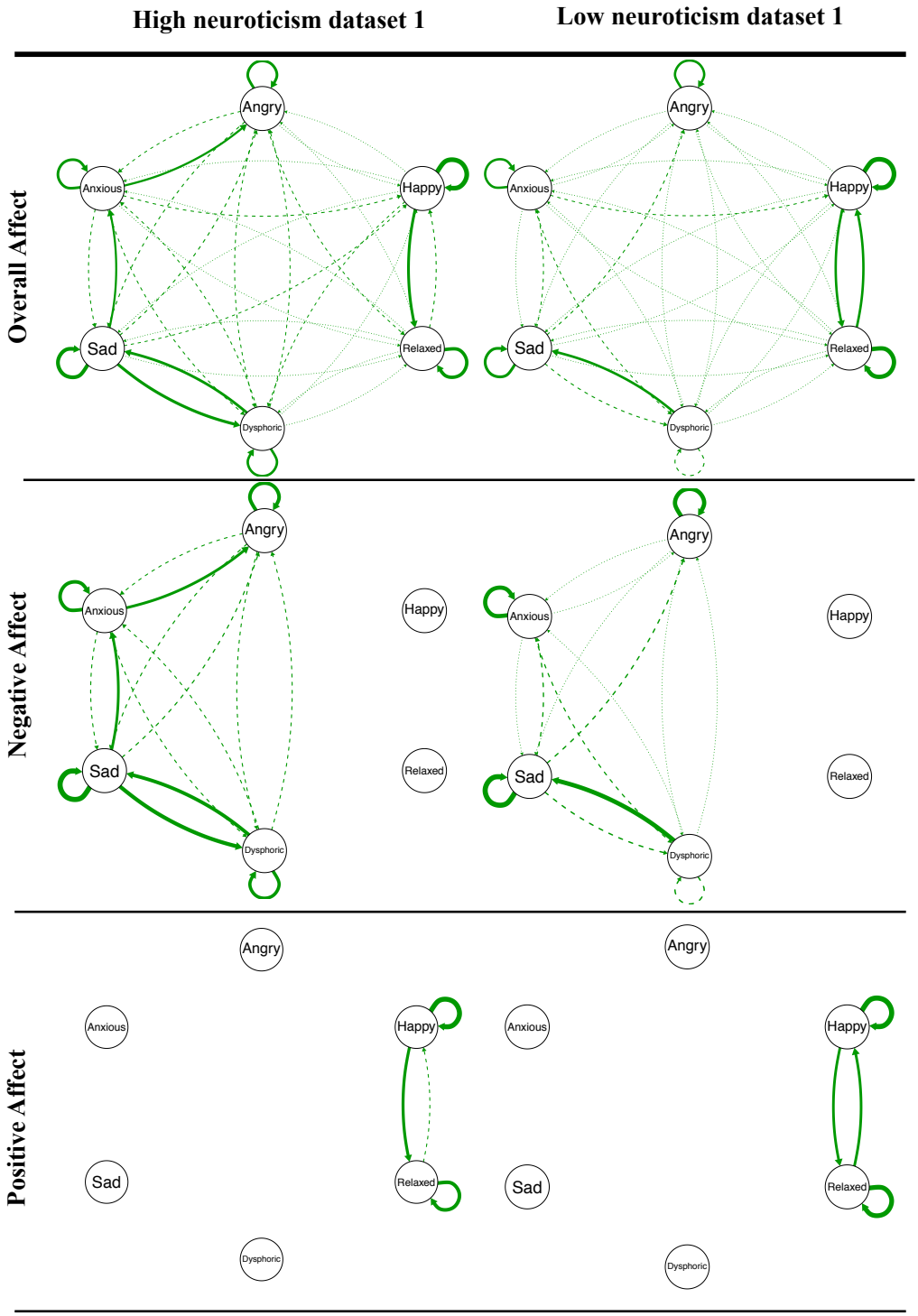


Figure 3.3: *The emotion networks for dataset 1 for individuals with a high and low neuroticism score.* In the network, the arrows indicate the absolute strengths of the temporal connections between emotions. Arrows that are dotted indicate weak connection strength, arrows that are dashed indicate moderate connection strength and bold arrows indicate strong connection strength.

Tables 3.2 and 3.3 show that there is a difference across the datasets in the out- and in-strength centrality. In dataset 1 individuals with high neuroticism scores have significantly high out-strength

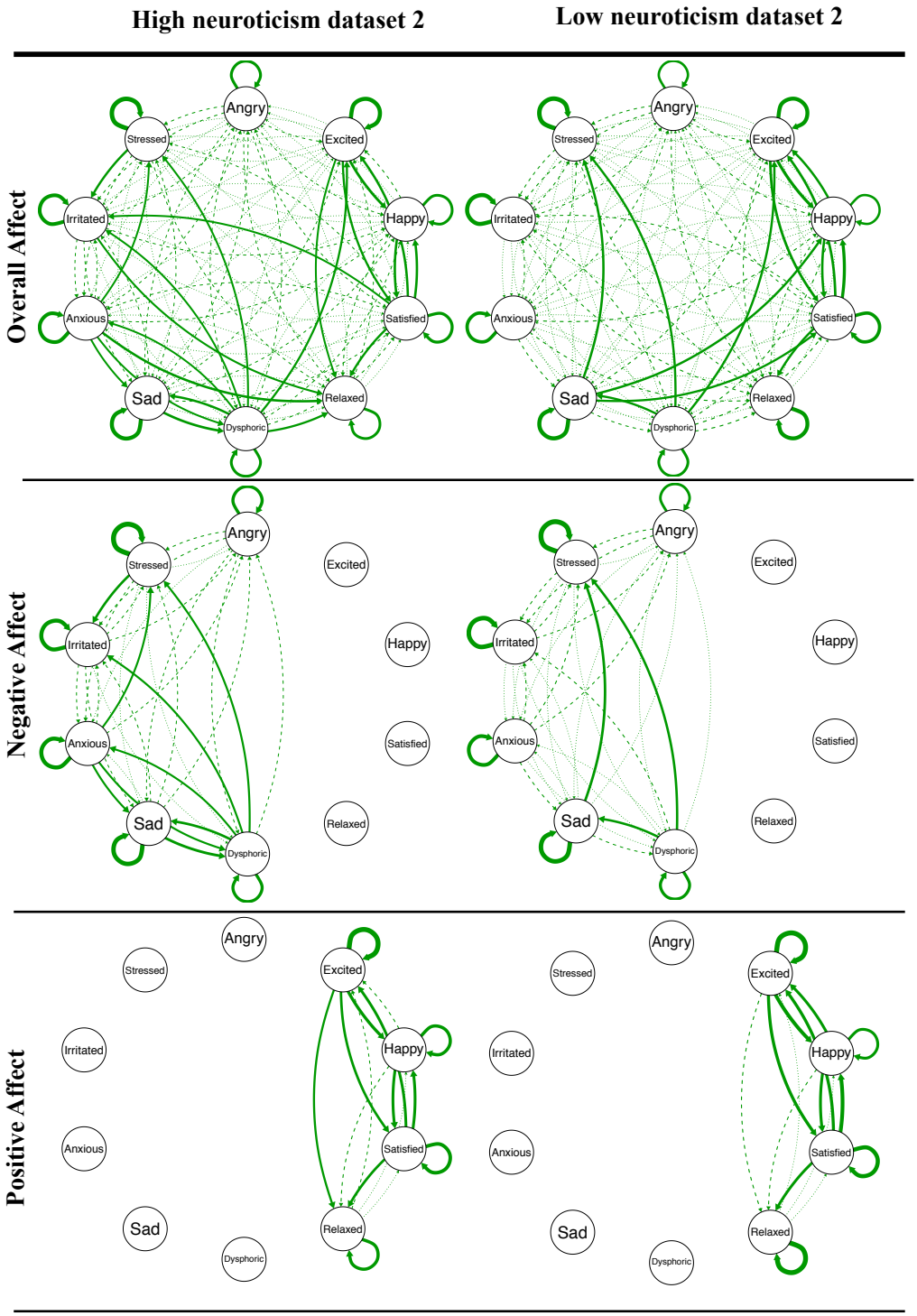


Figure 3.4: *The emotion networks for dataset 2 for individuals with a high and low neuroticism score.* In the network, the arrows indicate the absolute strengths of the temporal connections between emotions. Arrows that are dotted indicate weak connection strength, arrows that are dashed indicate moderate connection strength and bold arrows indicate strong connection strength.

centrality for all negative emotions and even for the positive emotion 'happy'. However, none of these results replicated for dataset 2, although the correlations are consistently positive. In contrast,

the positive significant relation between neuroticism and in-strength centrality of all five emotions (happy was non-significant in both datasets) of dataset 1 was also found in dataset 2. Thus, there is more evidence for a positive relation between in-strength centrality of emotions and neuroticism than out-strength centrality of emotions and neuroticism.

Table 3.2: Out-strength centrality and its relation to neuroticism

	Non-standardized				*Standardized*			
	Data 1		*Data* 2		*Data* 1		*Data* 2	
Out-strength	r	p	r	p	r	p	r	p
Angry	.441	<.001	.219	.052	.495	<.001	.238	.035
Dysphoric	.266	.009	.126	.267	.249	.015	.088	.44
Sad	.407	<.001	.085	.458	.402	<.001	.029	.799
Anxious	.289	.004	.188	.098	.313	.002	.252	.025
Relaxed	.157	.128	.15	.188	.18	.082	.195	.085
Happy	.42	<.001	.204	.071	.431	<.001	.252	.025
Satisfied			.298	.008			.331	.003
Excited			.361	.001			.349	.002
Irritated			.315	.005			.35	.002
Stressed			.283	.012			.267	.017

Table 3.3: In-strength centrality and its relation to neuroticism

	Non-standardized				*Standardized*			
	Data 1		*Data* 2		*Data* 1		*Data* 2	
In-strength	r	p	r	p	r	p	r	p
Angry	.357	<.001	.232	.040	.346	.001	.212	.060
Dysphoric	.407	<.001	.406	<.001	.394	<.001	.423	<.001
Sad	.348	<.001	.338	.002	.376	<.001	.352	.002
Anxious	.374	<.001	.470	<.001	.371	<.001	.395	<.001
Relaxed	.391	<.001	.347	.002	.341	.001	.280	.013
Happy	−.003	.98	−.093	.416	.024	.815	−.196	.084
Satisfied			−.068	.553			−.053	.641
Excited			−.057	.616			−.048	.672
Irritated			.192	.09			.163	.152
Stressed			.151	.184			.130	.254

As is apparent in Table 3.4, closeness centrality is positively related to neuroticism for almost all emotions (except 'stressed') in both datasets, even after standardization. This is in contrast to the relationship between betweenness centrality (influencing the overall information flow) and neuroticism (see Table 3.5). Although in some cases the relation was significant, it was not very strong, and none of the findings replicated in both datasets.

Finally, regarding the self-loops and their relation to neuroticism, it is evident that only the self-loops of emotions 'sad' and 'anxious' were significantly related to neuroticism in both datasets (see Table 3.6).

Table 3.4: Closeness centrality and its relation to neuroticism

	Non-standardized				*Standardized*			
	Data 1		*Data* 2		*Data* 1		*Data* 2	
Closeness	r	p	r	p	r	p	r	p
Angry	.493	<.001	.386	<.001	.503	<.001	.438	<.001
Dysphoric	.373	<.001	.22	.052	.358	<.001	.261	.020
Sad	.501	<.001	.363	.001	.475	<.001	.381	.001
Anxious	.305	.003	.312	.005	.310	.002	.381	<.001
Relaxed	.353	<.001	.351	.002	.394	<.001	.368	.001
Happy	.436	<.001	.386	<.001	.481	<.001	.435	<.001
Satisfied			.408	<.001			.437	<.001
Excited			.447	<.001			.450	<.001
Irritated			.456	<.001			.464	<.001
Stressed			.376	.001			.367	0.001

Table 3.5: Betweenness centrality and its relation to neuroticism

	Non-standardized				*Standardized*			
	Data 1		*Data* 2		*Data* 1		*Data* 2	
Betweenness	r	p	r	p	r	p	r	p
Angry	.156	.131	−.033	.771	.133	.198	−.093	.414
Dysphoric	.151	.143	.287	.010	.179	.083	.282	.012
Sad	.120	.248	.291	.009	.124	.230	.160	.158
Anxious	.096	.354	.324	.004	.027	.797	.299	.007
Relaxed	−.101	.329	−.089	.434	−.116	.263	−.046	.687
Happy	−.058	.579	−.100	.382	−.094	.364	−.043	.704
Satisfied			−.271	.016			−.163	.152
Excited			−.047	.679			−.056	.621
Irritated			−.016	.888			−.069	.548
Stressed			−.225	.046			−.150	.187

Table 3.6: Self-loops and their relation to neuroticism

	Data 1		*Data* 2	
Self-loops	r	p	r	p
Angry	.226	.028	.219	.052
Dysphoric	.295	.004	.167	.141
Sad	.417	<.001	.265	.018
Anxious	.257	.012	.362	.001
Relaxed	−.122	.240	−.133	.243
Happy	.169	.102	.303	.007
Satisfied			.108	.341
Excited			.327	.003
Irritated			.128	.262
Stressed			.240	.033

Note: In order to standardize the edges of the network the standard deviations of the predictor and outcome variables are used. Since a self-loop has the same predictor as outcome variable the standardized and unstandardized edges are equal.

3.4 Discussion

In this study, we found that for individuals with high levels of neuroticism, the associations found in the population level network were qualitatively the same but more pronounced (i.e., denser) than for their less neurotic peers. This effect was especially clear in the negative emotion network and was found in both datasets irrespective of standardization. Although in some cases the positive emotion network was significantly denser in individuals with a high neuroticism score, this effect was rather weak and not consistent across both datasets. These findings are noteworthy because they further reinforce the idea that neuroticism is characterized specifically by negative emotions that tend to co-occur (even across time). This is also in line with the results of (Pe et al., 2015), who found that individuals with the clinical diagnosis of depression have especially a denser negative emotion network than non-depressed individuals (see also Wigman et al., 2015 for a similar result).

These results also support previous research on early warning signs reflecting vulnerability for emotional disorder. Individuals who experience a higher autocorrelation have slower dynamics, which can be seen as predictive of a transition into depression (van de Leemput et al., 2014). In the same way, people who are highly neurotic and have strong self-loops (autoregressive effects) and strong connections between their emotions (cross effects) can be seen as being prone to experience a critical slowing down and thus an episode of depression.

Regarding the relation between centrality measures of the specific emotions and neuroticism, the results were more mixed. Although in the first dataset there were strong associations between the out-strength of individual emotions and neuroticism, this was not replicated in the second dataset. This could be due to the larger differences in neuroticism between individuals in the first versus the second dataset; alternatively these differences may reflect sampling error, as centrality indices are composites of many distinct parameters each of which is subject to random fluctuations due to the sampling of individuals from the population and the sampling of time points within individuals. The association between in-strength centrality and neuroticism, however, did replicate: Individuals experiencing a high degree of neuroticism were more likely to have a network in which angry, dysphoric, sad, anxious or relaxed had a high in-strength centrality, i.e., these emotions were more likely to be directly affected by the other emotions at the next time point.

Moreover, closeness centrality (how fast an emotion variable can be reached) was positively related with neuroticism for all emotions except for stressed. Betweenness centrality (the importance of a variable in funneling the emotion flow), on the other hand, did not reveal a clear association with neuroticism. Finally, the self-loops indicated that individuals with higher emotional inertia or overspill of especially the emotions sad and anxious were more neurotic. This is in line with previous research, which found that high negative emotional inertia or the spillover of negative emotions was linked to

neuroticism (Suls et al., 1998; Suls & Martin, 2005).

Thus, the more strongly connected emotion networks in highly neurotic individuals seem to be driven by in-strength centrality or the fact that emotions are affected by the other emotions of the network in a negative way (negative emotions get augmented whereas relaxed gets mostly blunted by the other emotions). Additionally, an important feature of the networks of highly neurotic individuals seems to be that most emotions can be reached fast (closeness centrality). These results show that, to better understand the relationship between neuroticism and emotions, not only the full network density should be taken into account, but also the local structure of the network.

A limitation of this study is its generalizability. Even though the results often replicated in the two datasets, in both datasets the participants were undergraduate students living in Belgium and the studies were conducted in the same lab. To be able to generalize the results, it would be interesting to use studies from other labs with different participants (e.g., older individuals and from different countries) to replicate the results. In addition, only a limited number of emotions were assessed, especially regarding positive emotions. Additionally, only the unique effects of the edges in the network are taken into account and thus a (possibly large) part of the explained variance is not included in the network. Solutions to take both unique and shared variance into consideration, such as the relative importance matrices, are currently only suitable for VAR models, and are not straightforward to generalize to a multilevel framework (Bulteel et al., in press). A further problem concerns spurious relationships in networks. As emotion processes are complicated dynamic systems it is unlikely that we have captured the full emotional process with the limited number of variables used in this paper, and thus spurious relationships might have been revealed. A promising solution to see if an edge is truly direct or spurious is through the use of ancestral graphs, which have been used in fMRI research for studying connectivity. Ancestral graphs are able to explicitly model whether there are relevant variables missing from a network model (Bringmann, Scholte, & Waldorp, 2013; Waldorp, Christoffels, & van de Ven, 2011). Future research should focus on developing these kinds of techniques further so that they can also be used in multilevel analyses.

As this study was based on mere correlations between neuroticism and emotions networks, it would be fruitful to have a more experimental setup in which one studies temporal emotion dynamics within individuals having different levels of neuroticism at different points in time. It is likely that individuals do not experience the same level of neuroticism continuously (Fleeson, 2001, 2004). Therefore, it would be interesting to see if in periods when neuroticism is, for example, less severe, one indeed would find less dense emotion networks than in periods when neuroticism is more severe. Note that in order to study such changing dynamics, extensions of the multilevel VAR technique will be needed, such as the multilevel threshold autoregressive model (de Haan-Rietdijk, Gottman, Bergeman, & Hamaker, 2014) or a time-varying autoregressive model (Bringmann et al., in press).

In this paper we have illustrated some of the possibilities of the network approach for studying temporal dynamics of psychological phenomena. More specifically, we have applied the network approach to an empirical example: the daily fluctuations of emotions and neuroticism. Whereas most studies have focused on aggregated or summed negative emotions and found that individuals with neuroticism tend to have a longer recovery of their negative emotions (i.e., higher emotional inertia; Suls & Martin, 2005), network analyses give a deeper understanding of this process. We have shown that there are emotion-specific effects, and moreover, it seems that the inflow and the speed of flow from other emotions was especially driving the stronger connectivity in more neurotic individuals. These new ways of analyzing emotions and other psychological phenomena can provide important information for better understanding how emotions are related to psychopathology, and for example how individuals get stuck in their emotions.

4 Revealing the dynamic network structure of the Beck Depression Inventory-II

Major Depressive Disorder (MDD) is a complex and burdensome mental health disorder made up of a wide variety of symptoms (APA, 2000; Hardeveld, Spijker, De Graaf, Nolen, & Beekman, 2010; Kessler et al., 2003; WHO, 2001). Structured interviews and questionnaires, such as the Hamilton Depression Rating Scale (HDRS; M. Hamilton, 1960) and Beck Depression Inventory (BDI & BDI-II; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961; Beck, Steer, & Brown, 1996) are important and commonly used tools to screen for, study, and follow the course of MDD (Beck, Steer, Ball, & Ranieri, 1996). In longitudinal studies, a total score, which results from simply adding all symptom scores, is often used as a measure of changes in depression severity. Relatively few studies use a more fine-grained analysis, in which the reduction of depression severity is studied by examining specific (clusters of) symptoms of depression instead of using the total score (e.g., Bhar et al., 2008; Fournier et al., 2013; Stewart & Harkness, 2012).

What all the above studies have in common is that they are based on the latent variable model. According to this model, symptoms of a given disorder are assumed to share an essential property; namely, their causal dependence on a latent variable, from which all symptoms arise (Borsboom, 2008; Kendler et al., 2011). In this perspective, symptoms experienced by patients are merely effects of the relevant latent variable (in this case depression). Standard models assume the symptoms to be statistically independent given the latent variable, and as a result, symptom associations are viewed to be spurious (Borsboom, 2008). Specifically, in the standard model, symptoms are not considered to have autonomous influence on one another. The latent variable approach, as utilized in standard models, is therefore not suitable to examine the dynamic relations between symptoms. The recently developed network approach (Cramer et al., 2010) steps away from this latent variable model by proposing that research should no longer focus exclusively on the mean level of symptoms or change therein (e.g., an overall score or a reduction of symptoms). Instead, this approach emphasizes that clinical research should also focus on the relation between individual symptoms from one time point to another, which we denote here as short-term dynamics.[1]

[1]Note that short-term dynamics refers to the dynamics between time points that are close to one another (e.g. time

The importance of examining these short-term dynamics is supported by various sources. First, theories of treatments for depression tend to focus on the short-term symptom dynamics when describing their proposed mechanisms of change. For example, according to Beck's cognitive theory (Beck, 1979), change in cognitive processes (e.g. negative thinking) leads to changes in symptoms such as affect (Beck, 1964; Rush, Kovacs, Beck, Weissenburger, & Hollon, 1981). Second, in clinical practice it is commonly observed that if patients experience relief in one symptom (for example sleeping problems), other symptoms start to wane as well, indicating the start of recovery; this is especially notable when symptoms are systematically assessed at the start of each therapy session, as is the case in cognitive therapy (Beck, 1979). Third, recent studies indicate that depression risk factors and stressful life events have differential effects on depressive symptoms (Cramer, Borsboom, et al., 2012; Fried et al., 2014). As Cramer, Borsboom, et al. (2012) showed, correlations between symptoms were directly influenced by the stressful life events and could not be explained by changes in an underlying common cause, in this case the risk to develop depression. This further supports the idea that symptoms have an autonomous influence on one another. Being able to objectively describe such symptom-by-symptom interactions can give important clues for clinical research and practice.

Apart from their substantive plausibility, network approaches open up a new range of research questions. For example, estimating a network of symptoms from depression questionnaires allows for an objective assessment of the centrality of symptoms (Boccaletti et al., 2006; Opsahl et al., 2010). Symptoms with a central position in the network are probably the most important or influential ones and are therefore likely to cause the symptom spread to continue. Studying these central symptoms can give clues for further clinical research. One could investigate, for instance, the commonly held assumption that anhedonia (loss of pleasure and interest) and depressed mood are central symptoms of depression as stated by the most prevalent diagnostic systems DSM-IV-TR (APA, 2000) and ICD-10 (WHO, 2008). In addition, once the network has been estimated, the community structure of the network can be examined (Girvan & Newman, 2002). A community is present if some clusters of symptoms are more strongly interconnected with each other than with symptoms that are not part of the cluster. In this way, the dynamic architecture of depression can be investigated.

This paper will be the first to investigate the short-term dynamics of one of the most widely used psychological questionnaire for depression: the Beck Depression Inventory (BDI-II). Inspired by the possibilities of the network approach, we will apply a novel method developed by Bringmann, Vissers, et al. (2013) that is able to explore these symptom dynamics, and infer a network structure of the BDI-II symptoms. Until recently, it was not possible to infer these kinds of directed and weighted networks from clinical questionnaires since two important requirements for studying short-

point $t-1$ and t). This in contrast to looking at changes in average values (mean levels), which can also be seen as a long-term dynamics. However, we do not use the latter term, since the term "change in mean level" is more intuitive.

term dynamics, intensive longitudinal data in which a set of symptoms is measured frequently across time, and a suitable statistical method, were lacking. Intensive longitudinal data is still sparse, but in a recent study of Lemmens et al. (2011, 2015) such data for the BDI-II were collected. Second, the newly developed vector autoregressive (VAR) multilevel method, which is a combination of multilevel (hierarchical) and time series models, is suited for analyzing these kinds of clinical longitudinal data. These data have rather short time series (ca. 20 time points) for a large sample of patients (Bringmann, Vissers, et al., 2013). Note that since only few studies have investigated single specific (clusters of) symptoms of the BDI-II or even interactions between symptoms in general, all analyses are exploratory.

The structure of the paper is as follows: First, we will infer the network(s) representing the short-term dynamics of the BDI-II symptoms. Second, we will study the centrality of symptoms. Based on the DSM-IV and ICD-10, one would expect the BDI-II items that are intuitively most closely related to the main symptoms, anhedonia and depressed mood (namely items: loss of interest, loss of pleasure and sadness), to be the most central ones in the network(s). In the third and last part, we will analyze whether communities are present in the BDI-II network(s). Since the network(s) consists of a fair number of symptoms (i.e., 21), we expect the emergence of new clusters of symptoms or community structures.

4.1 Method

Data

The data in the current study come from a large randomized clinical trial (RCT), which examined the effectiveness, relapse prevention and mechanisms of change of Cognitive therapy (CT) vs. Interpersonal psychotherapy (IPT) for depression (Lemmens et al., 2011, 2015). In this study, 182 patients (age between 18 and 65) with a DSM-IV diagnosis of Major Depressive Disorder (MDD) were randomly allocated to one of three conditions: (a) CT ($n = 76$), (b) IPT ($n = 75$), or (c) an 8-week waiting-list control (WLC) condition followed by treatment of choice (CT or IPT; $n = 31$). In the current study, we did not differentiate between patients who started therapy immediately and who started after 8 weeks. This resulted in a sample size of 99 for the CT condition (mean age of 40 years and $SD = 12$; 80% female) and a sample size of 83 for the IPT condition (mean age of 41 years and SD=12; 64% female).

There were no significant differences in demographic and clinical characteristics between the groups. Each patient participated in 3 to 20 weekly individual sessions, depending on the progress of the patient or due to drop out. On average, patients completed 14 sessions ($SD = 5$).[2] The BDI-II

[2] Analyses indicated that there were no differences in demographic and clinical characteristics between subjects that dropped out and those who finished therapy (at least 12 sessions).

was administered before each session to assess depression severity. Of the 2661 sessions, 2.5% of the BDI-II data were missing. Further details concerning the design of the trial and effectiveness of the interventions have been fully reported elsewhere (Lemmens et al., 2011, 2015).

Beck Depression Inventory-II (BDI-II)

The Beck Depression Inventory-II (BDI-II Beck, Steer, & Brown, 1996; van der Does, 2002) is one of the most widely used and empirically validated questionnaires for screening depression. The BDI-II is a self-report questionnaire measuring the severity of depression with 21 items. Each item is rated on a 4 point Likert-scale ranging from 0 to 3. The total score, ranging from 0 to 63, is constructed by adding the item scores, with higher scores reflecting more severe depressive symptomatology.

Interventions

CT and IPT are two of the most empirically validated psychotherapies used for treating depression (Cuijpers et al. 2008, 2011; Hollon et al. 2002). CT is based on Beck's cognitive theory (Beck, 1979), which states that depression results from maladaptive information-processing strategies that are maintained by dysfunctional behavioral responses. CT focuses on identifying and changing dysfunctional cognitions, schemas and attitudes in order to treat depression. In IPT, the interpersonal model of depression is central (Klerman et al. 1984). According to this model, major disturbances in the interpersonal domain may cause and maintain depression. It is assumed that depressive symptoms can be reduced through the improvement of interpersonal functioning.

4.2 Statistical analysis

The BDI-II network

First, we inferred the BDI-II network by analyzing the short-term dynamics between the 21 symptoms across the 20 weeks of therapy with a modified version of the multilevel-Vector Autoregressive (VAR) method (Bringmann, Vissers, et al., 2013)[3]. In the multilevel-VAR method, the time dynamics between the 21 symptoms of the BDI-II from one moment to the other are represented by a VAR model (see also Tschacher et al. (2012) for a similar approach). In the VAR model, the dependent variable

[3]Note that we deviate in this paper slightly from the procedure as proposed by Bringmann, Vissers, et al. (2013). With 21 items, it is not computationally possible to include all 21 random effects in the multilevel-VAR model simultaneously. Instead, we included only 5 random effects (including the autoregressive coefficient and the intercept) at the same time in a stepwise matter. Simulations indicated that the fixed effects could still be estimated precisely with this number of subjects and time points, which means that this is a feasible approach for estimating the current average network.

(e.g., symptom sadness; item 1) at time point t (e.g., session 2) is regressed on the lagged $t-1$ (e.g., session 1) versions of the independent variables (Box et al., 1994; Walls & Schafer, 2006)[4].

The independent variables in this study are all the symptoms of the BDI-II, measured at the previous time point (in this case the previous session). To account for differences between patients, all regression coefficients were assumed to be normally distributed at the population level. As a consequence, we obtained a multilevel model consisting of fixed (average) and random (individual) effects.[5] Each BDI-II symptom was used as a criterion variable once, which means that 21 multilevel-VAR models were estimated.

In order to estimate a multilevel-VAR model, data need to be stationary. An implication of this assumption is that that the variables will fluctuate around the same mean over time (Lütkepohl, 2007). Since the BDI-II symptoms decreased over the course of treatment (Lemmens et al., 2015), the means changed significantly, which indicates a non-stationary process. For this reason, a linear trend in the multilevel-VAR model was included, making the data trend stationary (Hamaker & Dolan, 2009). This implies that the short-term dynamics or the session-to-session fluctuations of the symptoms (as represented by the network) and the decrease of symptoms across the sessions (as represented by the linear trend) are modeled separately. Therefore, change in the short-term dynamics is in principle unrelated to change in the mean level of the BDI-II symptoms. Note further that stationarity also implies the assumption that the effects of symptoms on other symptoms are stable across time.

In order to obtain the BDI-II network, the estimated fixed effects of the multilevel-VAR analyses were used (Snijders & Bosker, 2012). Fixed effects represent the average connection strengths of the arrows in the network among the 21 symptoms and indicate whether the symptoms are positively or negatively related to each other. The fixed effects represent either autoregressive effects (self-loops) or cross-regressive effects (connections between different variables) in the network. Note that the network only represents the dynamic relations between the symptoms (the slopes of the multilevel-VAR model) and not the mean scores (the intercepts of the multilevel-VAR model) of the symptoms.

The estimated fixed effects or connections of the network resulted in a directed weighted network structure of the BDI-II, which was visualized using qgraph (Epskamp et al., 2012), a package for the statistical programming language *R*. Arrows or connections in the network represent more than mere associations between symptoms: because symptoms are measured over time, the connections can be viewed as an approximation of causality, resembling Granger-causality (Granger, 1969; Tschacher et

[4]Theoretically further lags are also possible. For example, a lag 2 model would indicate how symptoms are related to all symptoms experienced two sessions and one session ago. However, model comparison indicated that lag 1 was a more likely model than a lag 2 model (BIC lag 1: 71162, BIC lag 2: 71539).

[5]Simulations (not reported here) have indicated that because it is computationally not possible to include all 21 random effects at once in the multilevel-VAR model, the variance components (random effects variances) cannot be estimated accurately enough. For this reason, they will not be discussed further in the paper. The random effects should not be left out of the model though, because their inclusion leads to a more precise estimate of the fixed effects.

al., 2012). The network analyses were based on all the connections of the network. However, for reasons of clarity, we only visually present the strongest connections in the inferred network; that is, those connections which surpass the significance threshold (5%) using the False Discovery Rate (FDR) method (Benjamini & Hochberg, 1995). In the visually presented network, symptoms that are more strongly related to each other tend to be closer together in the figure (this is a result of the node placement algorithm, see: Fruchterman & Reingold, 1991).

Since the current study included two different therapy groups (CT and IPT), it is possible that two different network structures give rise to the data. We tested this in two ways: First, we fitted a model with the two networks separately and cross-correlated their estimated network links. Secondly, we compared a model in which we included two networks with a model which had one common network; for this purpose, we used the Bayesian Information Criterion (BIC; Schwarz et al., 1978). The model with the lowest BIC is the preferred model.[6]

Centrality analysis

In the second analysis, the inferred network was further analyzed by estimating the centrality of the BDI-II symptoms. In a centrality analysis, one can determine the relative importance or influence of a symptom in the network. We performed three types of centrality analyses: outdegree, indegree and betweenness centrality (see: Opsahl et al., 2010).[7] Outdegree centrality indicates how many outgoing arrows or how much information a symptom sends to other symptoms it is directly connected to. In the same way, indegree centrality indicates how many incoming arrows a symptom receives from the directly connected symptoms. Betweenness centrality takes into account both the direct and indirect connections of a symptom. A symptom with a high betweenness centrality is a symptom located on many paths between other symptoms and thus is a symptom through which the information in the network has to pass often. Therefore, a symptom with a high betweenness centrality is important in funneling the information flow or the symptom spread in the network.

Community structure analysis

As a third analysis, we performed a community structure analysis. In complex networks, new structures of clusters can often be found. An example of such a cluster is a *community*, in which groups of symptoms are densely interconnected among each other, but sparsely connected to the overall network. We used the *Walktrap* algorithm, which is suited for weighted networks (Pons & Latapy, 2005). This algorithm does not take directions of the arrows into account, so we summed the connection strengths

[6] The BIC was calculated by taking the average of the BICs of the separate univariate models.

[7] Since we want to estimate the centrality between the symptoms, self-loops are not taken into account in the centrality analyses. However, in all other analyses self-loops are taken into account.

(arrows) between two symptoms to have an appropriate undirected network suitable for analysis. The *Walktrap* algorithm uses random walks on the network to find communities or densely interconnected symptoms. The algorithm reveals how many groups can be found and also to which group a symptom of the network belongs to. All the analyses were done in the statistical software *R*.

4.3 Results

The BDI-II network

Figure 4.1 shows the inferred network of the dynamics between the 21 BDI-II symptoms. The analysis of cross-correlations and the model fitting approach using the BIC indicated that the network structure did not differ across the two therapy groups ($r = 0.86$, $p < 0.0001$; one network: $BIC = 77367.65$ versus two networks: $BIC = 80574.09$). Therefore, only one network was needed, representing both treatment groups. From the figure, it is evident that the strongest connections between symptoms are all positive in sign. Thus, when a symptom score increases, it is likely that other symptom scores also increase the next session, leading to an increase in the severity of symptoms in general. For example, if a participant reports feelings of guilt ('guilty feelings', item 5) in one session, that participant is more likely to report feelings of failure about the past ('past failure', item 3) the next session. The strength of the relation between symptoms translates into the thickness of the arrows in the figure: the stronger the symptoms are related the thicker the arrow between two symptoms, and the closer the symptoms tend to be together in the figure. This is expressed in, for example, the placement of the symptoms 'past failure' (item 3) and 'worthlessness' (item 14).

Apart from the connections between the symptoms, self-loops can contain important information. For example, the self-loop of the symptom 'loss of interest in sex' (item 21) is clearly the strongest connection of the network, meaning that when a participant reports loss of interest in sex one session, he or she is highly likely to report this in the next session as well. Furthermore, self-sustaining loops are apparent in the network. For example, 'worthlessness' (item 14) and 'guilty feelings' (item 5) seem to mutually influence each other. It should be mentioned that there are negative connections in the complete network as well. However, since these are rather weak, they did not pass the threshold for visualization in Figure 2.1. Note, however, that all connections are taken into account in the further analyses.[8] [9]

[8] We also confirmed that the connections in the network and thus the relationships between symptoms are not driven by differential variability. Standardizing the data per patient and per symptom led to a network that was highly similar to the original network; the correlation between parameters in the original and standardized network was 0.99. As a result, the conclusions of this paper are robust with respect to standardization of the data and are unlikely to reflect differential symptom variability.

[9] Proportional odds logistic regression (POLR), which is a regression model for ordinal response variables, also showed highly similar results; the correlation between parameters in the original and POLR network was 0.96 and led to similar centrality and community cluster results.

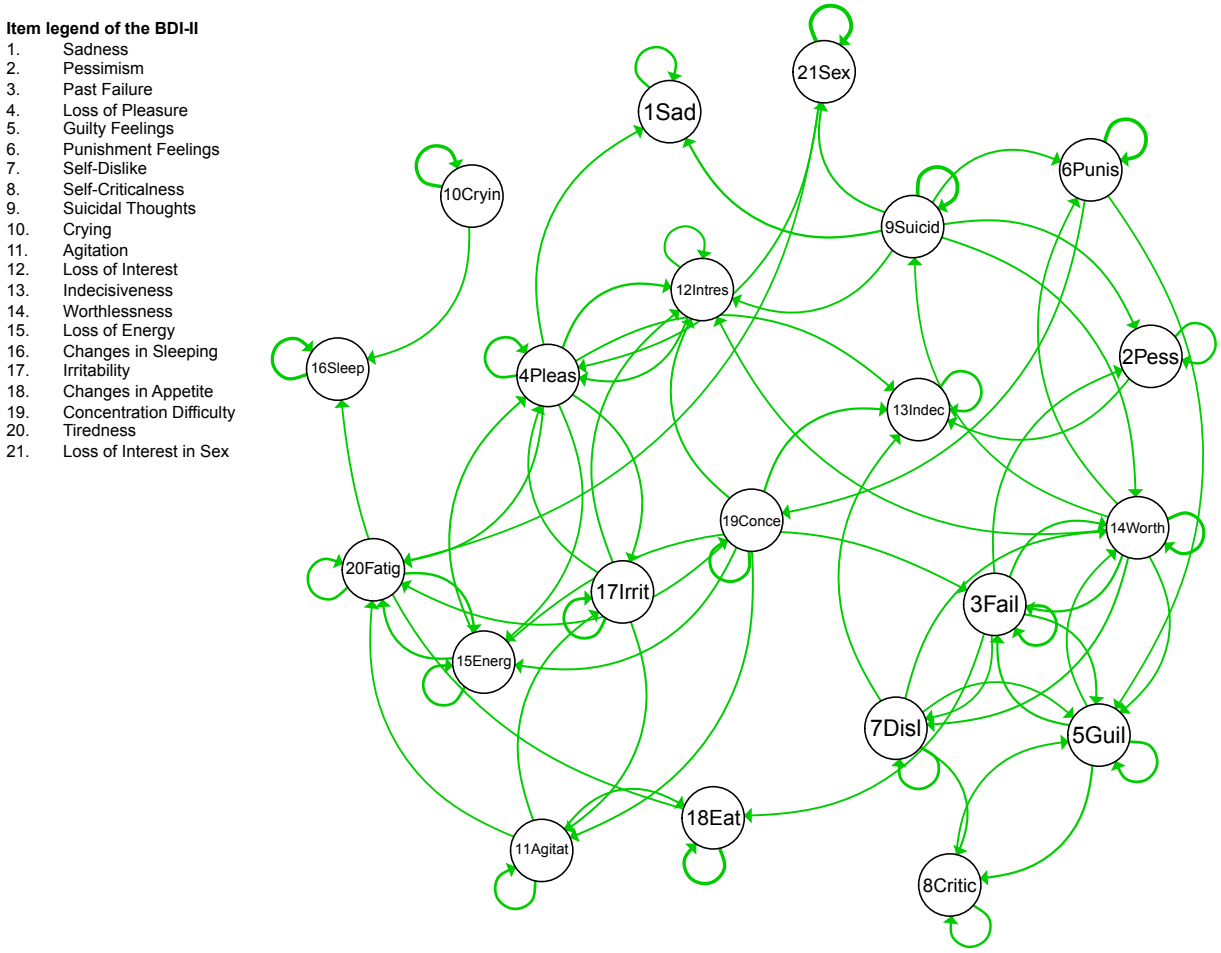


Figure 4.1: *The BDI-II network.* In this network, the connections between the 21 symptoms that surpass the significance threshold are visualized. Because of multiple hypothesis testing, we do not use the traditional 0.05 cutoff for p-values as the standard (which would inflate the number of unimportant links to be visualized); instead, we control the False Discovery Rate or FDR at 5% (Benjamini & Hochberg, 1995). Here, the 75 connections that pass the FDR threshold are visualized.

Centrality analysis

Figure 4.2 presents the results of the centrality analysis. The left panel of the figure indicates that the symptom 'loss of pleasure' (item 4) has one of the highest outdegrees, meaning that when one reports loss of pleasure in one session, it is likely that one will also report an increase in other symptoms in the next session. This is in contrast to, for instance, the symptom 'changes in sleeping patterns' (item 16), which is less likely to directly affect other symptoms the next session.

The middle panel indicates that the symptoms 'indecisiveness' (item 13), 'loss of interest' (item 12), 'past failure' (item 3) and 'sadness' (item 1) feature higher indegrees and thus receive a lot of information from other symptoms. This is in contrast to 'suicidal thoughts' (item 9): this symptom

is unlikely to be influenced by other symptoms, and is more likely to influence other symptoms (see also the first panel again).

The right panel indicates that the symptoms 'loss of pleasure' (item 4) and 'past failure' (item 3) feature the highest betweenness centralities, but they also have one of the highest outdegree ('loss of pleasure') and indegree ('past failure') centrality scores, respectively. Thus, the symptoms 'loss of pleasure' and 'past failure' are important in funneling the activation flow or symptom spread in the network.

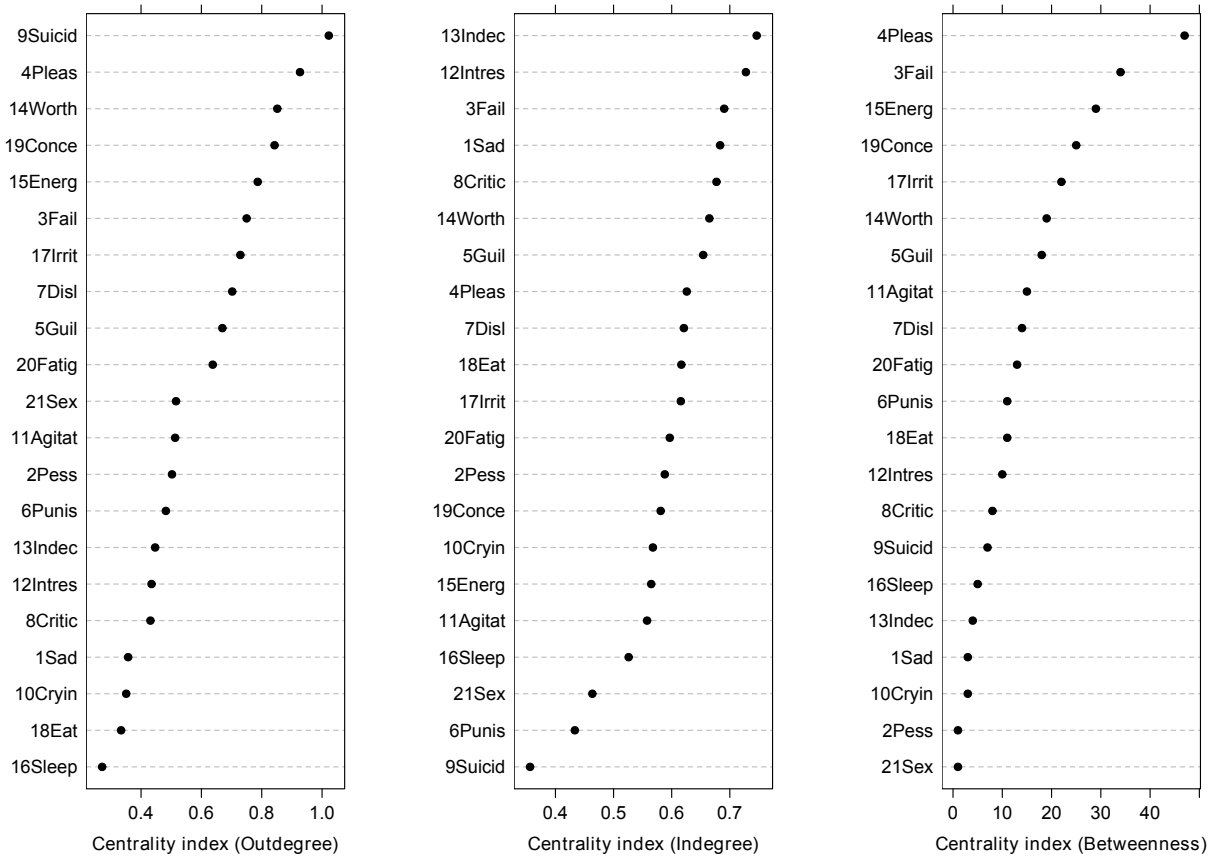


Figure 4.2: *Outdegree, Indegree and Betweenness centrality for all BDI-II symptoms.* The black dots are the model-based estimates of outdegree, indegree, and betweenness centrality. The higher the centrality index score the more central the symptom is in the network.

Community structure of the BDI-II network

The community structure analysis using the *Walktrap* algorithm indicated a two-cluster solution (see Figure 4.3).[10] This community structure means that symptoms in one cluster are more densely interconnected among themselves and more sparsely connected to symptoms in another cluster. The

[10] A hierarchical cluster analysis on the sum of the weighted links gave highly similar results.

green cluster in Figure 3 consists of the symptoms 'guilty feelings' (item 5), 'past failure' (item 3), 'self-dislike' (item 7), 'self-criticalness' (item 8), 'worthlessness' (item 14), 'punishment feelings' (item 6) and 'pessimism' (item 2), which are often described as cognitive symptoms. Items in the yellow cluster mainly consist of physical and affective symptoms of depression that appear related to loss of energy and pleasure.

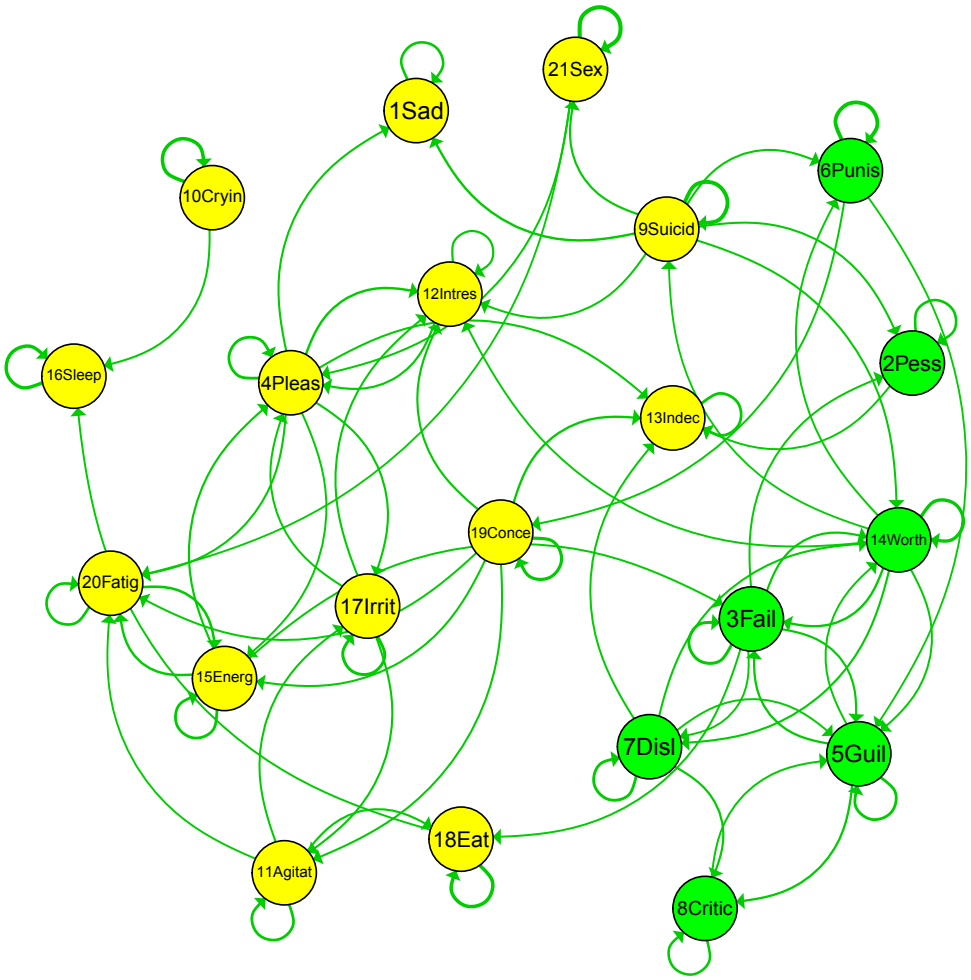


Figure 4.3: *Community structure of the BDI-II network with the two clusters indicated in two different colours.*

4.4 Discussion

In this paper, we derived for the first time a network that represents the session-to-session dynamics of one of the most widely used and empirically validated self-report measures for assessing the severity of depression: the BDI-II (Beck, Steer, & Brown, 1996). Results indicate that, in this network,

all BDI-II symptoms are directly or indirectly connected. In addition, the strongest connections between the symptoms are uniformly positive, indicating that, in general, when a symptom changes in severity, other symptoms tend to change in the same direction. This pattern of symptom dynamics is independent of the overall decrease in symptom scores as this trend was modeled separately; hence, the pattern does not reflect the common influence of treatment or recovery. In addition, for each symptom-symptom relation, we controlled for the effect of all other symptoms in the network. Thus, the evidence from this study points to the conclusion that direct effects among symptoms of the BDI-II are prevalent, and in fact connect all symptoms assessed in the questionnaire. In other words, symptoms of depression form a network of direct interactions.

Centrality analyses of the symptoms suggested that some symptoms are likely to have a larger influence on the symptom spread than other symptoms. As one may expect based on, e.g., the DSM-IV, the symptom 'loss of pleasure' (item 4) is one of the most central items in the symptom network and thus has a relatively large effect on the enduring of depressive symptoms in general. Somewhat more surprisingly, the symptoms 'sadness' (item 1) and 'loss of interest' (item 12) have a high indegree centrality, but quite a low outdegree and betweenness centrality, which means that they tend to increase in severity as other symptoms become more severe, but do not play a large role in funneling the symptom spread themselves. Thus, these symptoms may serve a mainly reactive role in the short-term dynamics of depression. Additional studies are needed to confirm these results, preferably engaging different depression questionnaires, such as the Hamilton Depression Rating Scale (HDRS).

Based on theory, one may expect a difference in symptom dynamics for subjects receiving CT and IPT because both treatments are assumed to work through different mechanisms. We did not observe such differences. A potential explanation for our findings could be that the dynamics between symptoms are similar when the treatments that are being compared are equally effective in reducing pathology, a fact that has been well established for CT and IPT for depression (e.g., Cuijpers, van Straten, Andersson, & van Oppen, 2008; Hollon & Ponniah, 2010). Alternatively, it might be the case that differences between CT and IPT actually do exist, but that we did not capture them in the current study because the BDI-II is, due to its design, insensitive for the differences between the two treatments. For example, even though the BDI-II includes items on several cognitive components (key elements of CT), items referring to problems in the interpersonal domain (core of IPT) are lacking. Further research involving other questionnaires is necessary to indicate if there are differences in symptom dynamics between therapies. A final possibility is that the difference between CT and IPT does not lie in symptom-symptom interaction, as studied in this paper, but in differences that arise in, e.g., stepwise changes in symptomatology. In this case, therapy effects might be detected in the way symptoms decrease or increase from one time point to another. Models that may be used to analyze

such differences, while accounting for the network of symptom-symptom interaction, are currently unavailable; however, non-linear statistical network inference techniques that may be used to model such processes are within reach, and could be used to investigate this issue in the future.

In the present study, community analyses revealed two groups of symptoms. The result appears to accommodate emerging evidence from the biomedical literature, which points to two types of depression: melancholia and atypical depression (Lamers et al., 2010, 2012). The current community clusters resemble these different depression types, as the green cluster in Figure 3 has similarities to melancholic type, whereas the yellow cluster resembles atypical depression. It is also interesting to note that the community structure result, based on multiple time points, is similar to the two-factor solution of the BDI-II, based on pooling across subjects at one time point (as found in e.g., Beck, Steer, & Brown, 1996; see also Arnau, Meagher, Norris, & Bramson, 2001; Steer, Ball, Ranieri, & Beck, 1999). Except for 'suicidal thoughts', all other symptoms in the green cluster of Figure 4.3 are the same as in the cognitive dimension of the two-factor solution of the BDI-II, whereas the yellow cluster could be interpreted as the somatic-affective or non-cognitive dimension. Although it is a good sign that the results we find are consistent with what one typically finds using factor analysis, our approach leads to a different way of thinking, different strategies for intervention, and to very different conclusions. In the latent variable approach, there are just two clusters of symptoms, which is a static result. In the network view, the result concerns the communication between symptoms that is denser within the cluster than with symptoms that are not in the cluster, leading to new hypotheses on how interventions should be operationalized, namely focusing on the interaction between symptoms. Thus, the existence of such patterns of influence is not a replication of the results of factor analysis on individual differences; rather, it may be seen as a potential explanation for these results (Wichers, 2014).

Several findings of this paper suggest further research. One important issue is how our results, which only involved participants with a diagnosis of depression, compare to results from unaffected individuals. For example, it is important to investigate whether a similar network characterizes healthy individuals. One hypothesis would be that there are no distinct symptom clusters in healthy subjects, but that instead all symptoms are similarly (and weakly) connected. Such a network would be more resilient, since activation would not spread as easily, and it would be less likely to get stuck in a cluster of symptoms. Another important topic for future research involves the difficult question of how to relate different time scales (Boker, Molenaar, & Nesselroade, 2009). This is because the symptoms that characterize depression are likely to influence each other in different time windows. For example, sleep problems are likely to exert effects in a pattern of a day-to-day variation, whereas mood states are much quicker and may affect each other within minutes. The question of how the dynamics of these different time scales interact with each other is, in our view, one of the main puzzles to be solved

in the study of symptom dynamics.

Regarding clinical practice, the relevance of the methodology and results of our approach may lie in opportunities to determine symptom centrality. For example, network analyses may be used to indicate which symptoms should be targeted first, and in this sense may help in setting up treatment strategies. Ideally, such analyses should be based on person-specific analyses (cf. Molenaar & Campbell, 2009). Unfortunately, at the moment such analyses are not computationally feasible for large networks of 21 symptoms. However, future development of the multilevel-VAR method, combined with a higher frequency of within subject assessment, should make it possible to take this procedure a step further, which may eventually lead to person-specific therapeutic interventions. Information about person-specific network centrality would not necessarily require pre-treatment assessment, and the high frequency assessment could be informative at any point, even if started during therapy. For example, if a centrality analysis of an individual network reveals that for that specific person 'loss of pleasure' is the most central symptom, therapy that intervenes on this symptom would be more effective than treatment that intervenes on non-central symptoms; for other persons, different interventions may be preferable. In a similar vein, one could hypothesize that if 'suicidal thoughts' is the most central symptom for a given person, this may signal acute need for care. Furthermore, since suicidal thoughts has a high outdegree, and is thus likely to trigger other symptoms, but a low indegree, and is thus not likely to be influenced by the other symptoms, interventions should be directly targeted at this symptom. Given the increased opportunities for assessing highly intensive time series within individuals, person-specific treatment protocols based on networks of symptom dynamics are rapidly becoming a realistic possibility. Thus, the network perspective is a promising new research field, which can give guidance to research on depression and to psychological research in general.

5 Changing dynamics: Time-varying autoregressive models using generalized additive modeling

Humans are complex dynamic systems, whose emotions, cognitions, and behaviors fluctuate constantly over time (Nesselroade & Ram, 2004; L. P. Wang, Hamaker, & Bergeman, 2012). In order to study these within-person processes, and to determine how, why, and when individuals change over time, individuals need to be measured on a relatively large number of occasions (Bolger & Laurenceau, 2013; Ferrer & Nesselroade, 2003; Molenaar & Campbell, 2009; Nesselroade & Ram, 2004; Nesselroade & Molenaar, 2010), resulting in intensive longitudinal data that, if $N = 1$, are typically designated as time series (Walls & Schafer, 2006). Currently, a spectacular growth of studies gathering intensive longitudinal data is taking place (aan het Rot et al., 2012; Bolger et al., 2003; Mehl & Conner, 2012; Scollon, Prieto, & Diener, 2003). With this development, it has become possible to study dynamical processes of psychological phenomena in much greater detail than has hitherto been possible (Trull & Ebner-Priemer, 2013).

There are various aspects of within-person processes that one can choose to study in order to gather insights into psychological dynamics, of which *temporal dependence* is one particularly informative aspect (Boker et al., 2009; McArdle, 2009). Temporal dependence concerns the degree to which current observations can be predicted by previous observations, for example, the degree to which an individual's emotional state at a given time point is predictive of her emotional state at subsequent time points (Jahng et al., 2008; Kuppens, Allen, & Sheeber, 2010).

A popular approach to handling such temporal dependency is autoregressive (AR) modeling, a family of statistical models in which the structure of the time-dependency in the data is explicitly modeled through regression equations. Some autoregressive models are suited to study time dependence within a single individual (e.g., Hertzog & Nesselroade, 2003; Molenaar, 1985; Rosmalen, Wenting, Roest, de Jonge, & Bos, 2012; Stroe-Kunold et al., 2012), whereas multilevel techniques can model time dependence within multiple individuals simultaneously (e.g., Bringmann, Vissers, et al., 2013; de Haan-Rietdijk et al., 2014; Song & Ferrer, 2012; Oravecz et al., 2011). In addition, AR techniques can be applied in various frameworks, such as the Bayesian (e.g., Pole et al., 1994) and the structural equation modeling framework (SEM; e.g., Hamaker et al., 2003; McArdle, 2009; Voelkle et

al., 2012).

A drawback of most AR models is that they are based on the assumption that the average value around which the process is fluctuating as well as the variance and the temporal dependency of the process are time-invariant. This is also known as the *stationarity assumption* (Chatfield, 2003). However, in the context of psychology this may not always be a realistic assumption. In fact, it could be argued that in many psychological time series studies a form of non-stationarity can be expected to be present (e.g., Bringmann et al., 2015; Molenaar, De Gooijer, & Schmitz, 1992; Rosmalen et al., 2012; Tschacher & Ramseyer, 2009). Even more so, often the very reason why it is interesting and important to study dynamics of psychological processes lies in their non-stationary nature (Boker, Rotondo, Xu, & King, 2002; van de Leemput et al., 2014). For example, when an individual receives therapy, the aim is to accomplish change, such as a decrease in symptoms. Thus, instead of considering dynamics, such as temporal dependency, as static characteristics of an individual, it is more realistic to consider them as time-varying, which implies that standard AR models are unsuitable (Molenaar et al., 1992; Boker et al., 2002).

To overcome this limitation, time-varying AR (TV-AR) models have been developed (Dahlhaus, 1997). In these models, the parameters (the intercept and autoregressive parameter) of the AR model (most commonly an AR(1) model) are now allowed to vary over time, so the models can be applied to both stationary and non-stationary processes (Chow, Zu, Shifren, & Zhang, 2011). Most time-varying AR models used in psychology and econometrics are based on the state-space modeling framework (Chow et al., 2011; Koop, 2012; Molenaar, 1987; Molenaar & Newell, 2003; Molenaar, Sinclair, Rovine, Ram, & Corneal, 2009; Mumtaz & Surico, 2009; Prado, 2010; Tarvainen, Hiltunen, Ranta-aho, & Karjalainen, 2004; Tarvainen, Georgiadis, Ranta-aho, & Karjalainen, 2006; West, Prado, & Krystal, 1999). The state-space framework is very general and encompasses a wide variety of models, such as dynamic linear models. Hence, the framework is very powerful due to its generality, but the downside is that it requires learning (state-space) notation with which most psychologists are unfamiliar. In addition, state-space models require the user to specify the way parameters of the time-varying model vary over time (Belsley & Kuh, 1973; Tarvainen et al., 2004; for a notable exception see Molenaar et al., 2009), but in practice the required theories about the nature of the change are often lacking (Tan, Shiyko, Li, Li, & Dierker, 2012), or must be handled via explicit incorporation of spline-based or other nonparametric functions into a (confirmatory) state-space framework (Tarvainen et al., 2006). Doing so may entail high computational demands when the dimension of the unknown change forms to be explored is high. Thus, there is a clear need for a time-varying AR method that functions without pre-specification and moreover is easy to apply for researchers in psychology.

As we will show in this paper, one solution is to implement TV-AR models based on *semi-parametric* statistical modeling using a well-studied elegant and easily applicable generalized addi-

tive modeling (GAM) framework (Hastie & Tibshirani, 1990; McKeown & Sneddon, 2014; Sullivan, Shadish, & Steiner, 2015; Wood, 2006). The crucial advantage of semi-parametric TV-AR models in general is that they are data-driven, and thus the shape of change need not be specified beforehand (Dahlhaus, 1997; Fan & Yao, 2003; Giraitis, Kapetanios, & Yates, 2014; Härdle, Lütkepohl, & Chen, 1997; Kitagawa & Gersch, 1985). Furthermore, no state-space notation is needed, since the TV-AR model is closely related to and can be specified and estimated within the familiar regression framework. Software for applying the GAM framework is freely available in the *mgcv* package for the statistical software *R* (Wood, 2006). The package has well-functioning default settings, making it very user friendly.[1] By showing how the TV-AR model can be applied with existing and easy to use software, we hope to make the TV-AR method accessible for a broad audience of psychological researchers.

The structure of the paper is as follows. In the first section, a detailed explanation of the standard time-invariant AR is given. In the second section, we describe the general structure of the TV-AR model, and in the third section we explain in detail how the time-varying parameters are estimated, and also introduce the *mgcv* package in *R*, with which the TV-AR is estimated (McKeown & Sneddon, 2014; Wood, 2006). In the fourth section, we provide a simulation study and give guidelines on how to use the TV-AR model with the *mgcv* package. In the fifth section, we give an example from emotion dynamics research to illustrate the TV-AR method by applying it to two different subjects whose affect was measured over circa 500 days in the context of an isolation study, the MARS500 project (Basner et al., 2013; Tafforin, 2013; Vigo et al., 2013; Y. Wang et al., 2014). This section is followed by concluding remarks and the Appendix with a description of details of the simulation study. Additional details of the simulation study can be found in the *R*-code online.

5.1 Standard time-invariant AR

In this section, the standard time-invariant autoregressive (AR) model is explained in more detail. Code for the equations and figures in this section can be found in the *R-code*.

Time series data consist of repeated measurements on one or more variable(s) taken from the same system (e.g., an individual, dyad, family, or organization). Typically, such data are statistically dependent, since all measures are taken from the same participant (e.g., answers on a questionnaire are likely to be related over time, Brandt & Williams, 2007; Velicer & Fava, 2003). This statistical dependence or autocorrelation that occurs in repeated measurement data is a central aspect that has

[1] Note that a time-varying effect model that also allows fitting a semi-parametric TV-AR model has recently been developed in SAS (Tan et al., 2012). However, it is less general and has fewer options for fitting a TV-AR model (e.g., at the moment it is only suitable for normally distributed time-varying models).

to be accounted for when studying the underlying process. Furthermore, when this autocorrelation is not taken into account invalid estimates can occur.

In psychology, the standard model used to deal with this statistical dependency is a Gaussian discrete time AR model.[2] An AR model accounts for the statistical dependency by modeling it explicitly, or in other words, the time series is regressed on itself (Hamaker & Dolan, 2009). The most basic form is an AR model of lag order 1 or AR(1):

$$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t. \tag{5.1}$$

This amounts to a linear regression model with an intercept β_0, and the autoregressive coefficient β_1, representing the degree and direction of the relation between a measurement at a previous (lagged) time point $(t-1)$ and current time point (t) of a single variable y (Velicer & Fava, 2003) and can be estimated with ordinary least squares (OLS). The part of observation y_t that cannot be explained by the previous observation y_{t-1} is referred to as the innovation ε_t (Chatfield, 2003). Other terms for the innovation are random shock, perturbation, or dynamic error.[3] The innovations are assumed to be normally distributed with a mean of zero and variance σ^2_ε (J. D. Hamilton, 1994).

The autoregressive coefficient β_1 can also be interpreted as the extent to which a current observation is predictable by the preceding observation (Hamaker & Dolan, 2009). A positive relationship indicates that high values of a variable (e.g., Positive Affect; PA) at one time point are likely to be followed by high values in the next time period (see left panel of Figure 5.1). In contrast, a negative relationship would predict the opposite, namely low values of the variable during the next time period (Chatfield, 2003; Velicer & Fava, 2003), which typically results in a jigsaw pattern (see right panel of Figure 5.1).

An important assumption for an AR(1) model is stationarity. A distinction is made between strictly stationary and covariance-stationary (also known as weakly or second-order stationary) processes. If a process is strictly stationary, the distribution of y_t and all joint distributions of y random variables are the same at all time points, and are thus time-invariant. Covariance-stationarity is a less strong assumption, as in this case only the first two moments of a distribution, the mean and the variance, and thus the parameters β_0 and β_1, have to be time-invariant.[4] Furthermore, stationarity also requires

[2] In discrete time AR models the measurements of the process are assumed to be equally spaced, meaning that the distance between the measurements is the same through the whole study. If time points were not equally spaced, the autoregressive coefficient would have a different meaning across occasions. This is in contrast to continuous time AR models, where the intervals between time points do not have to be equal (see for more information: Bisconti, Bergeman, & Boker, 2004; Deboeck, 2013; Oravecz et al., 2011; Voelkle & Oud, 2013; Voelkle et al., 2012).

[3] The term dynamic error is used to pit this error against the well-known measurement error. The difference between the two error terms is that while measurement error is occasion-specific, affecting the scores only at a single occasion, dynamic error tends to affect subsequent occasions as well due to the underlying temporal dependency in the process (Schuurman, Houtveen, & Hamaker, 2015). In the current study we restrict our focus to processes without measurement error.

[4] As we study normally distributed processes here, it is interesting to note that in this case covariance-stationarity

that the autoregressive coefficient must lie between -1 and 1 (boundaries not included). In this case, the mean μ and variance σ^2 of the process in Equation 5.1 can be expressed as

$$\mu = \frac{\beta_0}{1-\beta_1} \tag{5.2}$$

$$\sigma^2 = \frac{\sigma_\varepsilon^2}{1-\beta_1^2}, \tag{5.3}$$

showing that both are time-invariant (Chatfield, 2003; J. D. Hamilton, 1994).

Figure 5.1 shows two examples of a stationary process. Although the process fluctuates (changes) in both the left and right panel, the intercept, mean, autocorrelation and variance do not change over time. In an AR model, the intercept term β_0 only has a substantial interpretation if a score of 0 is a possible value in the sample.[5] Therefore, we prefer to work with the mean μ, which can be interpreted as the value around which the process fluctuates.

5.2 Time-varying AR

Psychological data are often non-stationary, rendering a standard AR model inapplicable. In this section, we will therefore describe an alternative model, the TV-AR model, which can model non-stationarity. First, we will discuss non-stationarity, illustrated by two simulated examples with 150 time points (representing here the evolution of valence within an individual). Secondly, we will give a general overview of the TV-AR model. Information on statistical inference for the TV-AR model will be given in the next section. The code to make the figure in this section can be found in the *R-code*.

There are several sources that can give rise to a non-stationary process in which the intercept, mean, autocorrelation and (or) variance change over time. In psychological research, the focus has mainly been on detecting a type of non-stationarity that is due to a (gradual) change in the mean of a process, which is visible as a trend in the data. Consider for example the left panel of Figure 5.2, in which a simulated process of hypothetical valence scores for an individual is shown. Here the autoregressive parameter does not change over time ($\beta_1 = 0.2$), but the intercept does, as represented by the dashed line, and therefore the mean also changes. Thus, a trend in the data is present.

To deal with a trend, common approaches in psychology have been *detrending* and *modeling the trend*. In the first method, data are made stationary by subtracting the values of a fitted trend from the individual data-points, thus removing the trend from the data (Hamaker & Dolan, 2009). A

implies strict stationarity, since a normal distribution is completely defined by its first two moments (Chatfield, 2003, p. 36).

[5]The intercept β_0 is the expected score when the observation at the previous occasion was zero (i.e., $y_{t-1} = 0$). When the scale that is used does not include the score zero, the intercept is typically not interesting.

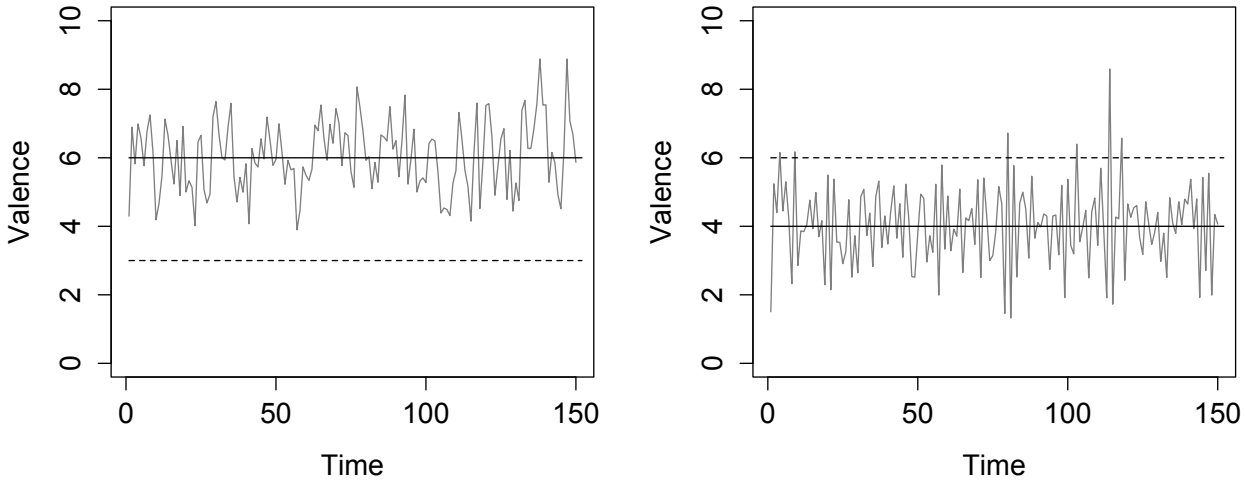


Figure 5.1: *Simulated time series with a positive (left) and a negative (right) autocorrelation for a valence process of a single individual.* The valence process ranged from 0 to 10, with 0 indicating feeling very unhappy and 10 indicating very happy. The process was simulated for 150 time points with an intercept (β_0) of 3 (left) and 6 (right; see dashed line in both graphs) and an autoregressive coefficient (β_1) of 0.5 (left) and -0.5 (right), meaning that there was a positive (left) or negative (right) dependency in the data. Notice that here the intercept as such has no further meaning and is different from the mean. In the left graph, the mean (μ; shown by the solid black line) is $3/(1-0.5) = 6$, indicating that on average this individual felt quite happy. In the right graph, the mean is $6/(1+0.5) = 4$, indicating that on average this individual felt slightly unhappy.

drawback associated with this way of dealing with non-stationarity is that it may remove important information from the data (Molenaar et al., 1992). In the second approach, stationarity is obtained through modeling the trend with, for example, linear growth curve modeling (Tschacher & Ramseyer, 2009). Both modeling the trend as well as detrending require specifying the functional form of the trend, which can be difficult, especially when convenient parametric forms are not applicable (Adolph, Robinson, Young, & Gill-Alvarez, 2008; Faraway, 2006; Tan et al., 2012). The TV-AR model that we will present has the advantage that it can detect trends in a data-driven way, and thus no pre-specifications are needed to account for a trend in the data.

Detrending or modeling the trend makes the process trend-stationary. However, when detrending, often only the trend due to a changing intercept is removed, and what is overlooked is that non-stationarity and trends can also occur due to changes in the autocorrelation.[6] For example, Figure 5.2 (right panel) shows a process that is non-stationary due to a change in the autocorrelation. The autoregressive function changes linearly over time, from a high value ($\beta_1 = 0.65$) to a lower one ($\beta_1 = 0.2$). At first, the data are characterized by a high autocorrelation, which disappears towards

[6]Note that a trend can be also caused by a unit root process, such as a random walk. In this case, the process has to be differenced in order to become stationary (see, for example, J. D. Hamilton, 1994).

the end of the time series. This is evident in the figure: First there are large oscillations (a signature of a high autocorrelation), which then become smaller towards the end of the time series (indicating low autocorrelation). Removing or modeling a trend as described above will not deal with this source of non-stationarity, leaving the process covertly non-stationary. This is an important reason why TV-AR models, which can detect and model both changes in the intercept and autocorrelation simultaneously, are important.

Another reason why TV-AR models are useful is that they can test for non-stationarity. There are several tests to check for stationarity, such as the Dickey Fuller test (which can be used to test whether a unit root is present in the time series; Dickey & Fuller, 1979), and the KPSS test (which can be used to test whether the mean is stable over time, or whether it follows a linear trend; Kwiatkowski et al., 1992). However, there is no specific test that checks for non-stationarity due to changing autoregression or a changing mean that follows a different trajectory than a linear trend. With the TV-AR model, we present a method that can test the time invariance of the autoregressive parameter, and simultaneously check whether a trend is due to a time-varying intercept and/or a time-varying autoregressive parameter (see Figure 5.2). Moreover, this method allows for instantly modeling such non-stationarity.

The defining feature of a TV-AR model is that the coefficients of the model are allowed to vary over time, following an unspecified function of time (Dahlhaus, 1997; Giraitis et al., 2014). To this end, we specify

$$y_t = \beta_{0,t} + \beta_{1,t} y_{t-1} + \varepsilon_t \tag{5.4}$$

where the intercept $\beta_{0,t}$ and the autoregressive $\beta_{1,t}$ coefficients are now functions that can change over time.[7] The innovations still form a white noise process so that the values of ε_t are independently and identically distributed, which implies that their variance is constant over time.

An important assumption of the TV-AR model is that, even though the functional form of $\beta_{0,t}$ and $\beta_{1,t}$ can be any function, change in the parameter values is restricted to be gradual, that is, there should be no sudden transitions. This assumption implies that the TV-AR model, as defined here, is not appropriate for time series with abrupt changes or sudden jumps. Thus, researchers should decide whether or not continuous change in parameters is plausible on the basis of the substantive knowledge of the problem at hand. If sudden, qualitative transitions are expected (e.g., as would be the case in some areas of cognitive development or in mental disorders with a sudden onset) then the current methodology would not be advisable. However, if the point at which an abrupt change takes place is known, one can model the change with a TV-AR model. One could specify, for example, a TV-AR model before and after an intervention. Additionally, although a TV-AR model is designed

[7] Note that in Giraitis et al. (2014) $\beta_{1,t}$ is specified as $\beta_{1,t-1}$. Here we use the standard notation used in Dahlhaus (1997).

for handling non-stationary processes, the process is still required to be *locally stationary*, meaning that $-1 < \beta_{1,t} < 1$, for all t (Dahlhaus, 1997).

Assuming that the change is restricted to be gradual and the process is locally stationary, the model implied mean is (Giraitis et al., 2014):[8]

$$\mu_t \approx \frac{\beta_{0,t}}{1 - \beta_{1,t}}. \tag{5.6}$$

Similarly, due to the fact that the autoregressive coefficient is allowed to vary over time, the variance of the time series is now also time-varying, that is,

$$\sigma_t^2 \approx \frac{\sigma_\varepsilon^2}{1 - \beta_{1,t}^2}. \tag{5.7}$$

Note that since μ_t can vary over time, in the literature μ_t is often interpreted as the *attractor* (also known as baseline or equilibrium) rather than the mean of the process (Giraitis et al., 2014; Hamaker, 2012; Oravecz et al., 2011). As is the case in a time-invariant AR model, the intercept and the changing mean (attractor or trend) are distinct features of a process. The intercept typically does not have a direct psychological interpretation, whereas the attractor represents the underlying trend in the time series (see Figure 5.2).

5.3 Inference of the TV-AR model: Splines and generalized additive models

In this section, we discuss how to estimate the time-varying parameters in the TV-AR model using the generalized additive model (GAM) framework. GAM models are expanded general linear models (GLMs), such that one or more terms are replaced with a non-parametric (smooth) function (Keele, 2008; Wood, 2006). This makes GAM models semi-parametric models, since predictor variables (i.e., in our case y_{t-1}) can either be modeled as in standard regression (e.g., β_1) or in a non-parametric way (e.g., $\beta_{1,t}$). We focus in this section on the nonparametric representation. Code for the figures can be

[8] To derive a model-implied mean of the TV-AR, we can write

$$\begin{aligned} \mu_t &= E\big[\beta_{0,t} + \beta_{1,t} y_{t-1} + \varepsilon_t\big] \\ &= E[\beta_{0,t}] + E[\beta_{1,t} y_{t-1}] + E[\varepsilon_t] \\ &= \beta_{0,t} + \beta_{1,t}\mu_{t-1} \\ &\approx \beta_{0,t} + \beta_{1,t}\mu_t \end{aligned} \tag{5.5}$$

where the latter approximation results from the fact that, in contrast to a standard AR model where we have $E[y_t] = E[y_{t-1}] = \mu$, the expectations of y_t and y_{t-1} are not exactly equal for a TV-AR model. However, since the parameters $\beta_{0,t}$ and $\beta_{1,t}$ are only allowed to change gradually, we can assume that μ_{t-1} is reasonably well approximated by μ_t, so that we have Equation 5.6. The derivation of the time-varying variance is similar to the derivation of the time-varying mean.

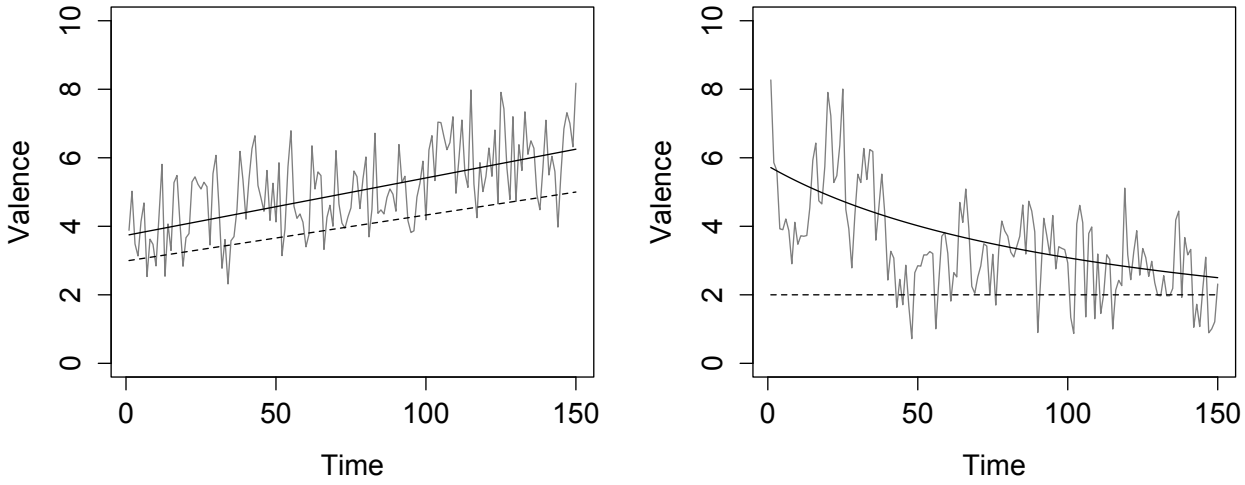


Figure 5.2: *Simulated data of a valence process (with 0 indicating feeling very unhappy and 10 indicating very happy) with time-varying parameters.* In the left panel, the autoregressive coefficient is time-invariant ($\beta_1 = 0.2$), while the intercept is time-varying ($\beta_{0,t}$; ranging from 3 to 5); in the right panel, the autoregressive coefficient is time-varying ($\beta_{1,t}$; gradually changing from 0.65 to 0.2), while the intercept is time-invariant ($\beta_0 = 2$). The attractor in the left panel (μ_t; shown by the solid black line) changes from 4 to 7, indicating that this individual felt a bit unhappy at first, but at the end of the time series felt happy, whereas the attractor in the right panel changes from circa 6 to 2.5, indicating that this individual felt happy at first, but at the end felt unhappy.

found in the *R-code*.

The non-parametric smooth functions used here are based on regression splines. Regression splines are piecewise polynomial functions that are joined (smoothly) at breakpoints called knots (Hastie & Tibshirani, 1990). In order to clarify the concept further, we will give a simulated example (based on Wood, 2006). Specifically, data are simulated for $n = 20$ time points according to a sine wave: $y_t = \sin\left(\frac{2\pi t}{20}\right) + \epsilon_t$, where $\epsilon_t \sim N(0, 0.3^2)$. We denote the time points in the data as t_i with $i = 1, ...20$. The data are represented as the small black dots in the first and last panel of Figure 5.3. To fit these data, we start with a simplified TV-AR model

$$y_t = \beta_{0,t} + \varepsilon_t \tag{5.8}$$

with only a time-varying intercept and no autoregressive parameter.

The goal is to find the function $\beta_{0,t}$ that tracks the general relation between y and t (which for this example is the sine wave underlying the data) as well as possible. In order to find the optimal

smooth function estimating $\beta_{0,t}$, the following penalized least squares loss function is minimized:

$$\sum_{i=1}^{n}[y_i - \beta_{0,t_i}]^2 + \lambda \int_{-\infty}^{+\infty} [\beta''_{0,t}]^2 dt. \tag{5.9}$$

In the first part of Equation 5.9 one can recognize the ordinary least squares minimization $\sum_{i=1}^{n}[y_i - \beta_{0,t_i}]^2$, which measures the distance between the function and data points. The last part is the roughness penalty $\lambda \int_{-\infty}^{+\infty} [\beta''_{0,t}]^2 dt$. This is an integrated squared second derivative that defines wiggliness, since the second derivative is a measure of curvature of the function whereas the integral sums up this measure along the entire domain of the function (Keele, 2008). Note that the square is needed to treat negative and positive curvature identically. The λ is a tuning parameter that controls the smoothness of the function. Small values of λ practically eliminate the penalty, thereby not penalizing for wiggliness and opening the possibility for wiggly functions. Large values of λ give a lot of weight to the penalty, thereby penalizing for wiggliness and restricting the possibility for wiggly functions. Minimizing the whole function leads to an optimal trade-off between goodness of fit and smoothness.[9]

The solution to the problem in Equation 5.9, denoted $\hat{\beta}_{0,t}$, can be expressed as a finite weighted sum of a set of predefined functions, known as basis functions. This can be written as follows:

$$\hat{\beta}_{0,t} = \hat{\alpha}_1 R_1(t) + \hat{\alpha}_2 R_2(t) + \hat{\alpha}_3 R_3(t) + \cdots + \hat{\alpha}_K R_K(t), \tag{5.10}$$

where we have expressed the solution in terms of K basis functions $R_1(t), \ldots, R_K(t)$ and t represents the predictor variable (time, in our case). The basis functions can be evaluated at every time t_i in the data and therefore the values $R_1(t_i), \ldots, R_K(t_i)$ can be collected in a $n \times K$ design matrix X so that the optimal regression weights can be determined by linear regression methods (see below).

Various options exist for choosing the smoothing basis, that is, the set of basis functions R_1 to R_K. Commonly used smoothing bases are *cubic regression splines* and *thin plate regression splines* (the latter being the default setting in the package *mgcv*), which represent alternative strategies with different properties of how the basis functions are chosen (Wood, 2006). Cubic regression splines are segmented cubic polynomials joined at the knots, and are constrained to be continuous at the knot points as well as to have a continuous first and second derivative (Fitzmaurice, Davidian, Verbeke, & Molenberghs, 2008). With cubic regression splines the locations of knots have to be chosen, the default setting in the *mgcv* package being that the knot points are automatically placed (equally spaced) over the entire range of data.

In contrast, the thin plate regression splines approach automatically starts with one knot per observation and then uses an eigen-decomposition to find the basis coefficients that maximally account

[9] Note that the least squares criterion can be used here because we assume continuous normally distributed data. In the more general case, the least squares criterion is replaced by minus the likelihood.

for the variances in the data. Thus, thin plate regression splines circumvent the choice of knot locations, reducing subjectivity brought into the model fit (Wood, 2006). Furthermore, unlike cubic regression splines, thin plate regression splines can handle smoothing in high-dimensional problems (e.g., when multiple independent variables occur). However, in one-dimensional problems, such as the one considered here, cubic and thin plate regression splines will lead to very similar solutions.

For our example, we have chosen a thin plate regression spline smoothing basis with $K = 6$ basis functions. The six basis functions are plotted in the panels 2-7 of Figure 5.3. The first two basis function are defined as $R_1(t) = 1$ and $R_2(t) = t$. Here one can recognize the constant and the first predictor variable of a standard linear regression model. The other four basis functions (R_3 - R_6) have a more complicated shape (for examples of such functions, see Gu, 2002; Keele, 2008; Wood, 2006). Additionally, in thin plate regression every basis function that is added is wigglier than the previous basis function. For example, basis function R_6 is wigglier than R_5. Note that in contrast to cubic splines, where the basis functions depend on the knot location, in thin plate splines a basis function cannot be associated with a knot location. Furthermore, the basis functions are evaluated at every value of t (also with the cubic spline smoothing basis). This is important to point out, as regression splines are defined as segmented polynomials that are joined at the knot points, so evaluations of the basis functions may prima facie seem to be restricted to particular segments.

After choosing the smoothing basis and the number of basis functions, estimating the time-varying function $\beta_{0,t}$ simply boils down to the estimation of the weights (denoted as α_i above) of the linear combination in a penalized regression sense (see below). In Figure 5.3, the final panel shows the weighted basis functions as well as the sine wave that is the final smooth function (i.e., $\hat{\beta}_{0,t}$, the thick dashed line).

Using a regression spline based method to estimate a smooth function raises the question of how many basis functions are needed to get a good fit. The usual approach is to place more basis functions than can reasonably be expected to be necessary, after which the function's smoothness is controlled by the roughness or wiggliness penalty as described earlier ($\lambda \int_{-\infty}^{+\infty} [\beta''_{0,t}]^2 dt$; see Wood, 2006). An attractive feature of spline regression methods is that the penalized loss function eventually boils down to a relatively simple penalized regression problem (see Wood, 2006). Thus, one can choose a reasonably large number of basis functions (so that in principle even very wiggly functions can be handled by the model), but then too wiggly components of the basis functions that are unnecessary are downplayed based on the value of the penalization tuning parameter λ. For instance, in our example the wiggliest basis function R_6 (panel 7 in Figure 5.3) is clearly penalized, as it appears as an almost flat horizontal line in the last panel of Figure 5.3.

Of course, the next question is then: What is a good value for the penalty parameter λ? If the value of λ is too small, the estimated function is not smooth enough, but if λ is set too high,

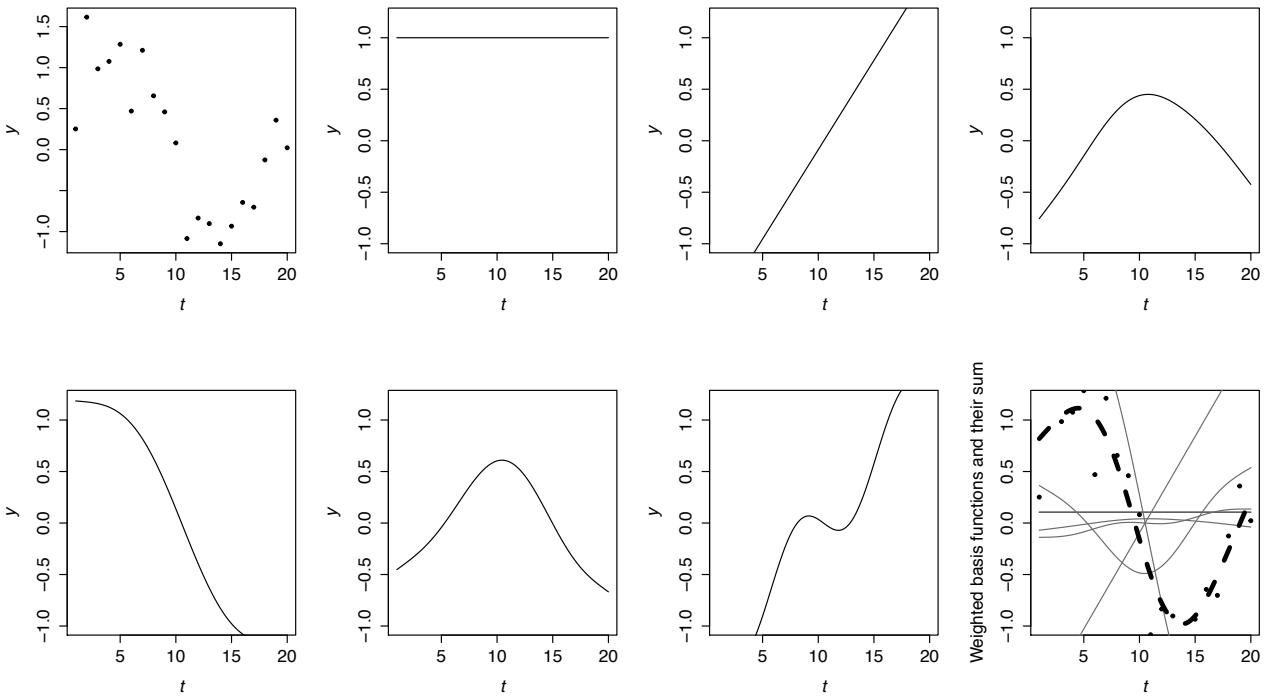


Figure 5.3: *The six basis functions for the curve $\beta_{0,t}$ using a cubic regression spline basis.* Just as in standard regression, all basis functions $R_i(t)$ are weighed by multiplying them with their corresponding α_i coefficients. The contribution of each basis function to the solution is estimated using penalized regression and the $\hat{\beta}_{0,t}$ (the thick black dashed line in the bottom right panel) is a weighted sum.

the function may oversmooth the data. Commonly, the optimal value of λ is determined using the *generalized cross validation* method (GCV; Golub, Heath, & Wahba, 1979). The idea of (ordinary) cross validation is that first a model, in this case a regression spline with a certain value of λ, is fitted on part of the data, for example leaving one datum out. In a second step, it is measured how well the estimated model can predict the other part of the data, for example the datum that was left out. However, with splines this process is computationally intensive and sensitive to transformations of the data (Wood, 2006). Therefore, the generalized cross validation score is used instead, which follows the same principle, but is invariant to transformations (Keele, 2008). The lowest GCV score indicates the optimal λ value and thus optimal smoothness of the estimated smooth function.

All of the above steps are implemented in the *mgcv* package in *R* (Wood, 2006). Using this software, one only has to define sufficiently many basis functions. The default for all splines is 10 basis functions. For the current example, detecting the relation between y and t, the command in *R* would be `gam(y~s(t,bs='tp',k=6))`, where the function `s` indicates the use of a smooth function for its argument (the predictor `t` in this case), `bs` indicates which smoothing basis is used (thin plate in this case), and `k` indicates the number of basis functions (see also the *R-code*). In addition to the GCV

score and the estimated smooth function, the *mgcv* package also provides 1) p-values, 2) a measure of nonlinearity (edf and ref.df), 3) 95% confidence intervals (CIs) and 4) model fit indices, all of which we elaborate on below.

1. The p-values for the smooth function result from a test of the null hypothesis that the smooth time-varying function is actually zero over the whole time range (Wood, 2013).

2. As non-parametric smooth functions (such as $\beta_{0,t}$) are difficult to represent in a formulaic way, a graphical representation is usually needed to get insight into the form of the function (see for instance Figure 5.3; Faraway, 2006). However, besides a plot of the smooth function, the *mgcv* package also provides a measure of nonlinearity in the form of the *effective degrees of freedom* (edf). Basically, the edf refers to the number of parameters needed to represent the smooth functions. At first sight, one may think that this is equal to the number of basis functions, but because of the penalization that is not the case. The reason why the penalization decreases the effective degrees of freedom is that the parameters are not free to vary because of the penalizations (Wood, 2006). The higher the edf, the more wiggly the estimated smooth function is, and an edf of 1 indicates a linear effect (Shadish, Zuur, & Sullivan, 2014). Furthermore, the edf also gives an indication of how much penalization took place and thus may serve as a diagnostic: The closer the edf is to the number of basis functions, the lower the penalization. Usually, an edf close to the number of basis functions means that additional basis functions should be added to capture the shape of the function. The ref.df is the reference degree of freedom used for hypothesis testing (Wood, 2013).

3. The 95% confidence intervals (CIs) around the smooth curve reflect the uncertainty of the smooth function. As the confidence intervals are obtained through a Bayesian approach, they are strictly speaking credible intervals, or Bayesian confidence intervals as referred to by Wood (see Wood, 2006).

4. Finally, model selection criteria can be retrieved with the package (such as BIC and AIC), where the lowest fit indices indicate the best model fit. When using the BIC and AIC for penalized models, note that the degrees of freedom are determined by the edf number and not by the number of parameters (see for more information Hastie & Tibshirani, 1990).

We have assumed a simple model with only a time-varying intercept to explain the fundamentals of splines. For the more realistic general TV-AR model, the time-varying autoregressive function is estimated in a similar way (see for further information Wood, 2006).

5.4 Guidelines regarding the TV-AR model: a simulation study

To evaluate how the TV-AR model performs under different circumstances using the default settings, we carried out a simulation study. In addition, we investigated the robustness of our method against violations of the assumption of gradual change, by considering also functions that change non-gradually. We will give here a general overview of the simulation conditions. In the Appendix the simulation setup is described in detail. In addition, there is *R-code* exemplifying some of the simulation results.

In the simulation study, we varied three factors: the generating function, low or high values for the model parameters, and the sample size. First, we had 5 generating functions for the intercept $\beta_{0,t}$ and the autoregressive parameter $\beta_{1,t}$: 1) both are invariant over time, 2) both increase linearly over time, 3) both follow a cosine function over time, 4) both follow a random walk and 5) both follow a stepwise function (see also Figure 5.4). Note that the random walk and the stepwise function are non-gradually changing functions. Strictly, the TV-AR model is thus not expected to recover these functions. Instead, we consider these functions to investigate the robustness of TV-AR in non-gradual conditions. The second factor we varied was the maximum absolute values of the parameters (low or high maximum value). The third factor was sample size (30, 60, 100, 200, 400, 1000).

Estimation was executed using five models: A) a TV-AR model using the default settings (a thin plate regression spline basis using 10 basis functions); B) a TV-AR model with only a time-varying intercept and a time-invariant autoregressive parameter using the default settings; C) a TV-AR model with only a time-varying autoregressive parameter using the default settings; D) a standard time-invariant AR model; and E) a thin plate regression spline basis using 30 basis functions.

We evaluated the estimates of all models with mean squared errors (MSE) and coverage probabilities (CP). Furthermore, we analyzed how well the BIC, AIC and GCV could distinguish between time-varying and time-invariant processes. Last, we looked at the significance of the parameters and the effective degrees of freedom (edf) if applicable.

Results and guidelines

The results show that the time-varying AR model was able to estimate all gradually changing generating functions (invariant, linear, cosine) very well using the default settings of the *mgcv* package in *R* (i.e., using 10 basis functions and thin plate regression splines; see Figure 5.4 and 5.5). Around 200 time points were needed for detecting a small change, such as a small linear increase over time, but large changes could already be detected with 60 time points.

In general, none of the model selection methods (BIC, AIC and GCV) performed well in selecting the correct model out of models A, B, C and D (e.g., with 100 time points in the high condition of the

linear increase, the BIC selects the correct model (model A) in only 60% of the cases). However, the BIC does relatively well in distinguishing between the time-invariant model D and the time-varying models (the three variants A, B and C combined). For example, with 100 time points in the high condition of the linear increase, the BIC selects the correct class (invariant versus time-varying) in circa 97% of the cases.[10]

As the BIC cannot be used for selecting the exact time-varying model (model A, B or C), additional criteria are needed. One possibility is to fit a TV-AR model and check the significance of the parameters (intercept and autoregressive parameter). If the intercept is significant, one can be confident that the intercept is time-varying, especially with at least circa 100 time points. This is because the TV-AR model automatically includes an (standard time-invariant) intercept, and significance implies that another, time-varying, intercept is needed. In contrast, in the case of the autoregressive parameter, significance entails that the parameter is valuable for the model, and thus should be kept, but it does not give information about whether it is a time-varying parameter or not. Additionally, a high edf is an indication that the parameter is time-varying, but note that the edf cannot be used to discriminate between time-invariant parameters and linearly increasing time-varying parameters, as they will often both have an edf of circa 2.

Even when the assumption of gradual change was violated, the TV-AR model was still able to estimate the general pattern of change (i.e., the trend-like fluctuations in the random walk), but not abrupt changes (such as in the stepwise function) or fast changes (i.e., the small-magnitude fluctuations in the random walk process). An exception was the condition with 1000 time points of the stepwise function, where the large jump could be detected quite well (see Figure 5.5). To get satisfying estimations in these cases, more time points are needed, and the amount of basis functions should be large enough. In general, it is advisable to always check whether you have enough basis functions. A good indication that you do not have enough basis functions and should increase their number is that the effective degrees of freedom (edf) come close to the number of basis functions (Wood, 2006). The simulation study showed that the average coverage probabilities of especially the non-gradually changing functions are clearly improved by increasing the number of basis functions (in this case from 10 to 30 basis functions; see Table 5.1). This lines up well with the advice given in general to have a high enough number of basis functions to allow for enough wiggliness in the estimated function (Wood, 2006).

.

[10]Note that the AIC and GCV were not as accurate as the BIC. For example, with 100 time points in the high condition of the linear increase, the AIC and GCV selected the correct class (invariant versus time-varying) in only 73% and 76% of the cases respectively.

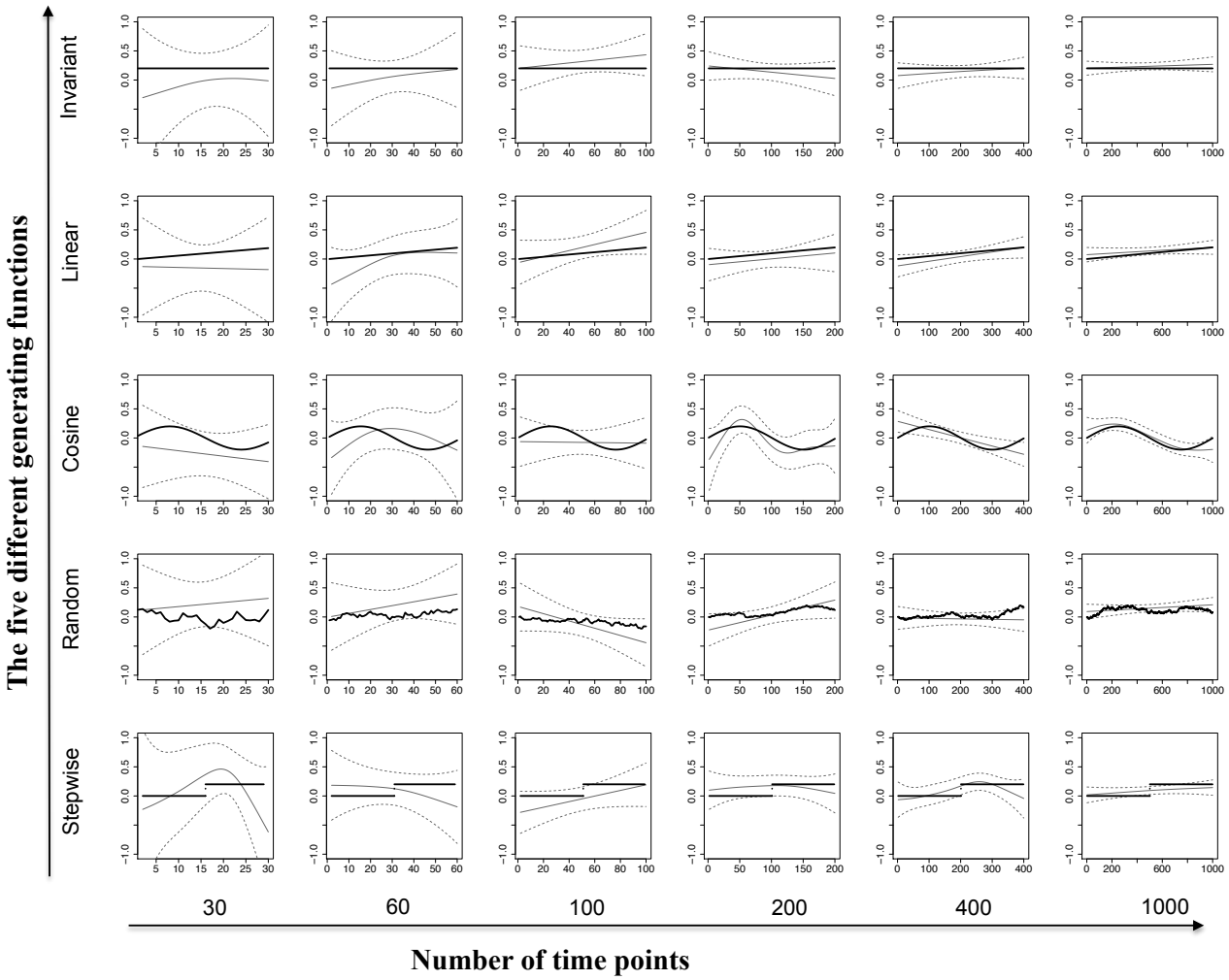


Figure 5.4: *Graphical representations of the generating functions of the autoregressive parameter for the low condition.* The different true underlying functions $\beta_{1,t}$ are represented as thick black solid lines and the estimated $\hat{\beta}_{1,t}$ as grey solid lines, the grey dashed lines being the 95% CIs. The estimations are based on the median of the MSE values of the 1000 replications.

5.5 An empirical example

We applied the TV-AR model to data of two individuals who took part in a long isolation study, the MARS500 project, in which psychological and physiological data have been collected to study the effects of living in an enclosed environment for the duration of a real potential mission to Mars (i.e., 520 days; for more information see `http://www.esa.int/Mars500`). We focus here on emotional inertia, which is studied in the context of affective research. Emotional inertia is defined as the temporal dependency of individual emotions, or the self-predictability of emotions, and is typically modeled with an AR model (Kuppens, Allen, & Sheeber, 2010; Suls et al., 1998). However, a study by Koval and Kuppens (2012) showed that emotional inertia is not a trait-like characteristic, but is itself prone to change, causing the data to be non-stationary (see also de Haan-Rietdijk et al., 2014;

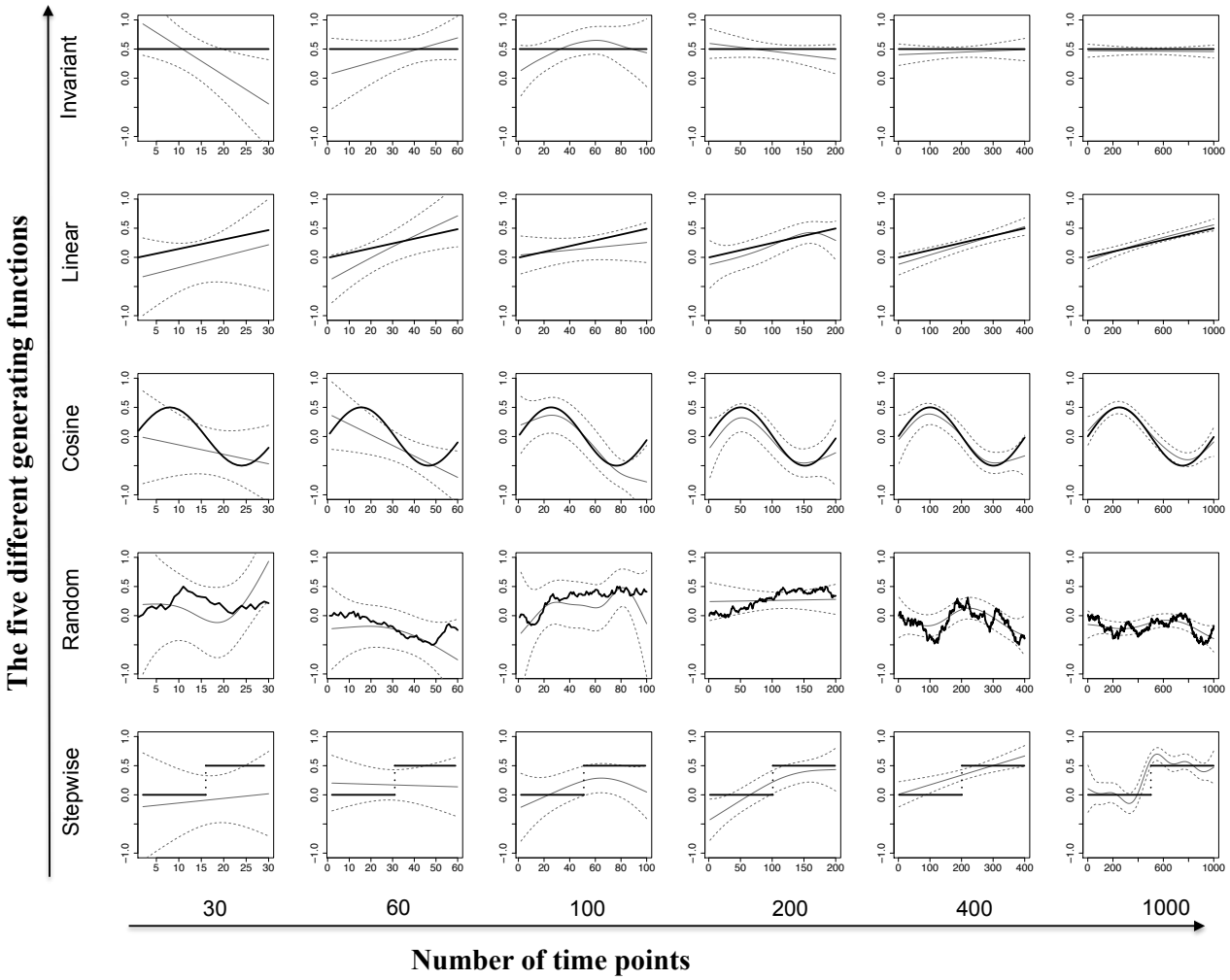


Figure 5.5: *Graphical representations of the generating functions of the autoregressive parameter for the high condition.* The different true underlying functions $\beta_{1,t}$ are represented as thick black solid lines and the estimated $\hat{\beta}_{1,t}$ as grey solid lines, the grey dashed lines being the 95% CIs. The estimations are based on the median of the MSE values of the 1000 replications.

Koval et al., in press). They showed, among other things, that individuals who anticipated a social stressor had a significant decrease in their emotional inertia, which means that to model the process of inertia correctly, the autoregressive parameter should be allowed to vary over time. In the MARS500 example, being isolated can be seen as a social stressor. Furthermore, it is plausible that the longer one is isolated, the more social stress there is. To study if and how inertia changed due to social isolation, we analyzed time series data from two persons involved in the MARS500 study using the TV-AR model.

Table 5.1: Coverage Probabilities (CP) of the autoregressive function in % using thin plate regression splines. Here the average CP of every simulation condition is given. Low and high stand for low and high value conditions for the maximum absolute values of the time-varying parameters. Note that the last line in the table uses the same settings as the previous line, except now 30 instead of 10 basis functions (K) are used.

	True underlying function									
	Invariant		*Linear*		*Cosine*		*Random*		*Step*	
N	*Low*	*High*	*Low*	*High*	*Low*	*High*	*Low*	*High*	*Low*	*High*
30	86	67	89	83	89	83	92	87	89	78
60	92	84	93	91	91	84	94	88	91	83
100	93	90	93	91	92	85	93	86	92	83
200	95	92	95	94	89	92	92	84	90	79
400	95	93	95	94	87	94	91	81	86	80
1000	95	95	95	95	89	96	86	78	82	82
1000 $K = 30$	95	94	94	95	91	96	87	83	84	87

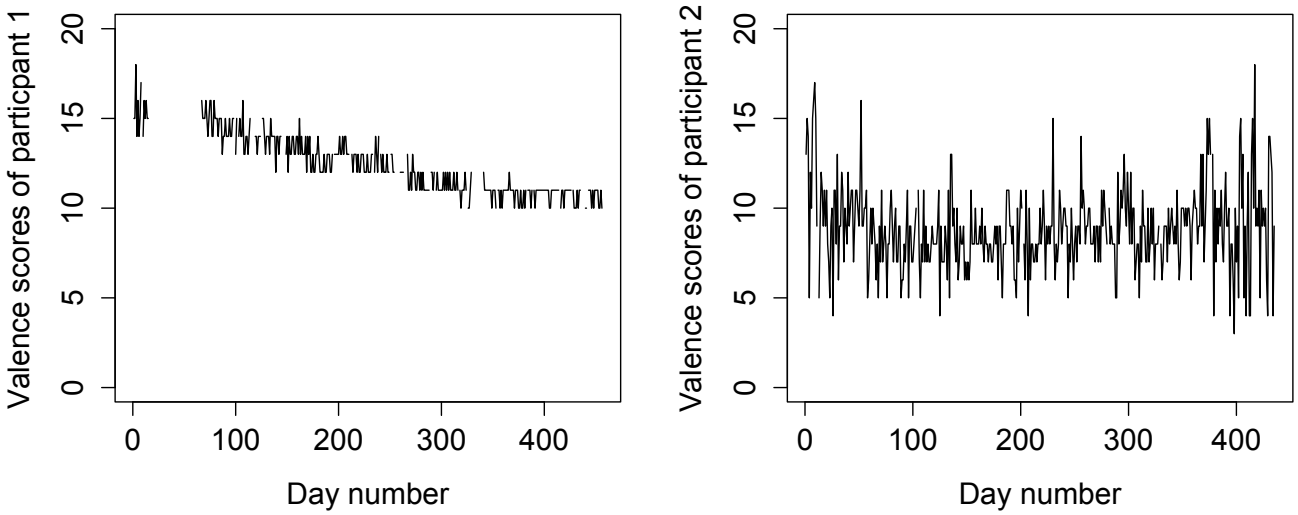


Figure 5.6: *The raw data of the variable valence for participant 1 (left) and participant 2 (right).*

Method

Data description

The MARS500 study consisted of six healthy male participants (average age was 34 years), who all signed a written informed consent before participating in this experiment. In accordance with the Declaration of Helsinki, the protocol was approved by The Ethics Committee of the University Hospital Gasthuisberg of Leuven (Belgium) and the ESA Medical Board before the research was conducted. We focus here on the dynamics of the variable 'valence' of two participants. Each morning, the

participants indicated on a 21×21 grid how they were feeling at that moment. The horizontal axis of the grid referred to valence and the vertical axis to arousal. Only the valence score (on 21-point scale) will be analyzed here. A high score indicates experience of highly positive feelings, and a low score experience of highly negative feelings.[11] There was 29% and 18% missingness in the data of participant 1 and 2 respectively (see Figure 5.6 for the raw data).[12]

Analyses

We consider the following four models:

In *model 1*, both the intercept and the autoregressive parameter are allowed to vary over time. The time-varying autoregressive parameter implies that the temporal dependency or emotional inertia (i.e., how self-predictable the emotion is) changes over time. Since the mean (or the attractor of the process) is a function of the intercept and the autoregressive parameter, it most likely also changes over time in this model:[13]

$$Valence_t = \beta_{0,t} + \beta_{1,t} Valence_{t-1} + \varepsilon_t. \tag{5.11}$$

In *model 2*, the intercept is allowed to fluctuate over time, but the autoregressive parameter is fixed over time, meaning that the temporal dependency (or emotional inertia) is time-invariant. Due to the changing intercept, the person's attractor also changes over time:

$$Valence_t = \beta_{0,t} + \beta_1 Valence_{t-1} + \varepsilon_t. \tag{5.12}$$

In *model 3*, the intercept is fixed over time, while the autoregressive parameter is allowed to vary over time. As indicated in the description of *model 1*, a time-varying autoregressive parameter means that the temporal dependency (or emotional inertia) of the process changes over time. However, fixing the intercept implies that the attractor changes over time, but this is fully accounted for by changes in the temporal dependency (i.e., the autoregressive parameter):

$$Valence_t = \beta_0 + \beta_{1,t} Valence_{t-1} + \varepsilon_t. \tag{5.13}$$

Finally, *model 4* is the standard AR(1) model, in which both the intercept and the autoregressive

[11] Although the measurement was done on a daily basis, on some days there were multiple measures, which was due to extra physiological tests that required additional measurements of valence and arousal. In these cases, we only used the first measure of the day.

[12] Note that the TV-AR model can also be used with missing data, although the more missingness the less power one has to detect the underlying process. Additionally, one has to assume that the missingness is (completely) at random.

[13] Of course it is possible, though unlikely, that the changes in the autoregressive parameter are exactly countered by the changes in the intercept (see Equation 5.6). In this case, the attractor would be time-invariant, while the temporal dependency would fluctuate over time.

parameter are time-invariant; as a result the mean (i.e., a time-invariant attractor) is also fixed over time. This means that the temporal dependency (or emotional inertia) is completely constant over time, that is, both the temporal dependency (or emotional inertia) and the attractor value of the process remain the same over time:

$$Valence_t = \beta_0 + \beta_1 Valence_{t-1} + \varepsilon_t. \tag{5.14}$$

Following the guidelines presented in the previous section, we first checked if the process was time-varying or not. For this purpose, we used the BIC: If the BIC selects model 1, 2 or 3 the process is probably changing over time, and otherwise (i.e., if model 4 is selected) the process is probably time-invariant. In the latter case, a standard AR model should be used; otherwise a TV-AR model is appropriate. Secondly, to check which parameters are time-varying, we considered whether the smooth parameters were significantly different from zero and thus were needed in the model. As noted before, a significant intercept indicates that this parameter is time-varying, whereas a significant autoregressive parameter does not entail that it is time-varying. Therefore, in a third step, when the autoregressive parameter was significant we checked if the edf was higher than 1. Additionally, we checked whether the residuals (estimated innovations $\hat{\varepsilon}_t$) indicated autocorrelation over time, satisfied the equal variance assumption and were normally distributed.

The analyses reported here were based on the default settings, that is, a thin plate regression spline basis with 10 basis functions (i.e., $K = 10$). We also ran all of the analyses with a cubic regression spline basis and thin plate regression splines with 30 basis functions (i.e., $K = 30$), but all results were highly similar and led to the same conclusions.

Results

As can be seen in Figure 5.6 (left panel), in the data of participant 1, a clear trend is apparent, whereas the data for participant 2 do not contain any clear time trend (Figure 5.6 right panel). For both participants the assumptions held for the selected models: The residuals for both participants did not indicate any autocorrelation over time, did not violate the equal variance assumption and were normally distributed.

For participant 1, the BIC indicated that the underlying process was varying over time and thus non-stationary (*model 2* was selected as the best model, although the differences between *model 1* and 2 were fairly small, see Table 5.2). Consequently, fitting the TV-AR model showed that the function of the intercept was significantly different from zero ($F = 3.42$, $p = 0.0046$, $edf = 4.50$, $ref.df = 5.20$), while the function of the autoregressive parameter was not ($F = 0.87$, $p = 0.51$, $edf = 5.01$, $ref.df = 5.62$). Thus, only a time-varying intercept was needed in the TV-AR model.

Based on visually inspecting Figure 5.7, the function of the intercept process (upper panel) is clearly varying over time, whereas the CIs of the function of the autoregressive parameter (middle panel) always include zero (the zero is represented by the dashed gray line) and the function does not clearly go up or down at any point in time. Taking all of these considerations into account, *model 2*, with a time-varying intercept and a time-invariant autoregressive parameter of zero, seems to be the best fitting model.

For participant 2, the BIC indicated that *model 3* had the best model fit and thus a TV-AR model was estimated. In line with this result, *model 1* (Equation 5.11) implied that the function of the autoregressive parameter was significant and should be kept in the model ($F = 8.32$, $p < 0.0001$, $edf = 5.17$, $ref.df = 6.15$), while the function of the intercept was not significant and thus time-invariant ($F = 0.15$, $p = 0.70$, $edf = 1.00$, $ref.df = 1.00$). Although significance does not imply that the autoregressive parameter is time-varying, the edf was clearly higher than 1. In addition, visual inspection of Figure 5.8 also clearly indicates that the autoregressive function (middle panel) of participant 2 changes over time. Thus, *model 3*, with a time-invariant intercept and a time-varying autoregressive parameter, seems to be the best model.

Table 5.2: Model selection for participants 1 and 2 using the BIC indices. Lowest fit indices are in bold.

Model	BIC Participant 1	BIC Particpant 2
Model 1	688	1,896
Model 2	**684**	1,894
Model 3	696	**1,890**
Model 4	868	1,899

In sum, in the data for participant 1, no inertia or autocorrelation of valence in the data is apparent, but rather it is the intercept that changes (see Figure 5.7 panel 3). In this specific case, the attractor is equal to the intercept as the autoregressive parameter equals zero. Participant 1 simply feels less happy as the isolation experiment proceeds, as represented by the changing intercept and attractor. This is not necessarily in contradiction with the results found by Koval and Kuppens (2012) as we do not know how much emotional inertia participant 1 had before the isolation experiment. It is possible, for example, that this participant had some level of emotional inertia before going into isolation, but as soon as the experiment started, his emotional inertia decreased to zero, which would be in line with the previous findings of Koval and Kuppens (2012). In contrast, participant 2 starts the isolation experiment relatively happy and with a high spill-over of valence (high inertia), but already after a few days, his inertia decreases until it gets to zero around 100 days, and also his valence becomes

more negative (see the attractor in the last panel of Figure 5.8). Towards the end of the experiment, there is again a light increase in his feeling of happiness and his inertia. This result is in line with research of Koval and colleagues, which suggests that as stress increases (the longer one is isolated) inertia decreases, and thus affect becomes less predictable (Koval & Kuppens, 2012).

Note that if one had ignored this non-stationarity in the data, a standard autoregressive model (thus, *model 4*) would have led to inaccurate conclusions about these two participants. For participant 1, ignoring non-stationarity would have led to inferring a highly significant autoregressive coefficient ($\beta_1 = 0.85$, $t(325) = 27.43$, $p < 0.0001$), that is, an extremely high inertia or a high predictability of his valence. For participant 2, ignoring non-stationarity would have led to the conclusion that there was a positive inertia ($\beta_1 = 0.20$, $t(420) = 4.29$, $p < 0.0001$), and the fact that his inertia was actually varying over time would have gone unnoticed.

In general, even though inertia is already well known to vary in strength greatly across individuals, it is still often studied as a trait of an individual. With the TV-AR model we can study inertia throughout the whole study period, creating an inertia value for every single time point. In future studies, it would be fruitful to take into account that inertia can change over time, even from day to day or faster, and of course, also in other contexts than social stress.

Furthermore, these two applications show how important it is in general to use a TV-AR model, as different conclusions would have been drawn with a standard AR model. In addition, with the TV-AR model trends as well as (time-varying) autoregressive parameters can be detected in one step: Even though the first example above (participant 1) involves a trend-stationary process and pre-specifying the exact (non-linear as the edf of 4.50 indicates) trend would have led to the same conclusions, this would have been much more difficult than with the TV-AR model. Psychological data can be non-stationary for various reasons, and the TV-AR model offers a simple exploratory tool for detecting such changing dynamic processes.

5.6 Discussion

In this paper, we have introduced a new way to study changing dynamics: the semi-parametric TV-AR model. This model fills a gap in the literature, because most standard autoregressive models do not take into account non-stationarity, even though many psychological processes are likely to be non-stationary. Therefore, there is a need for an easily applicable method for studying such non-stationarity or changing dynamics. The semi-parametric TV-AR model presented in this article is exactly such a tool.

As shown by the simulations and application in this paper, the TV-AR model can estimate non-stationary processes well and has significant potential for studying changing dynamics in psychology.

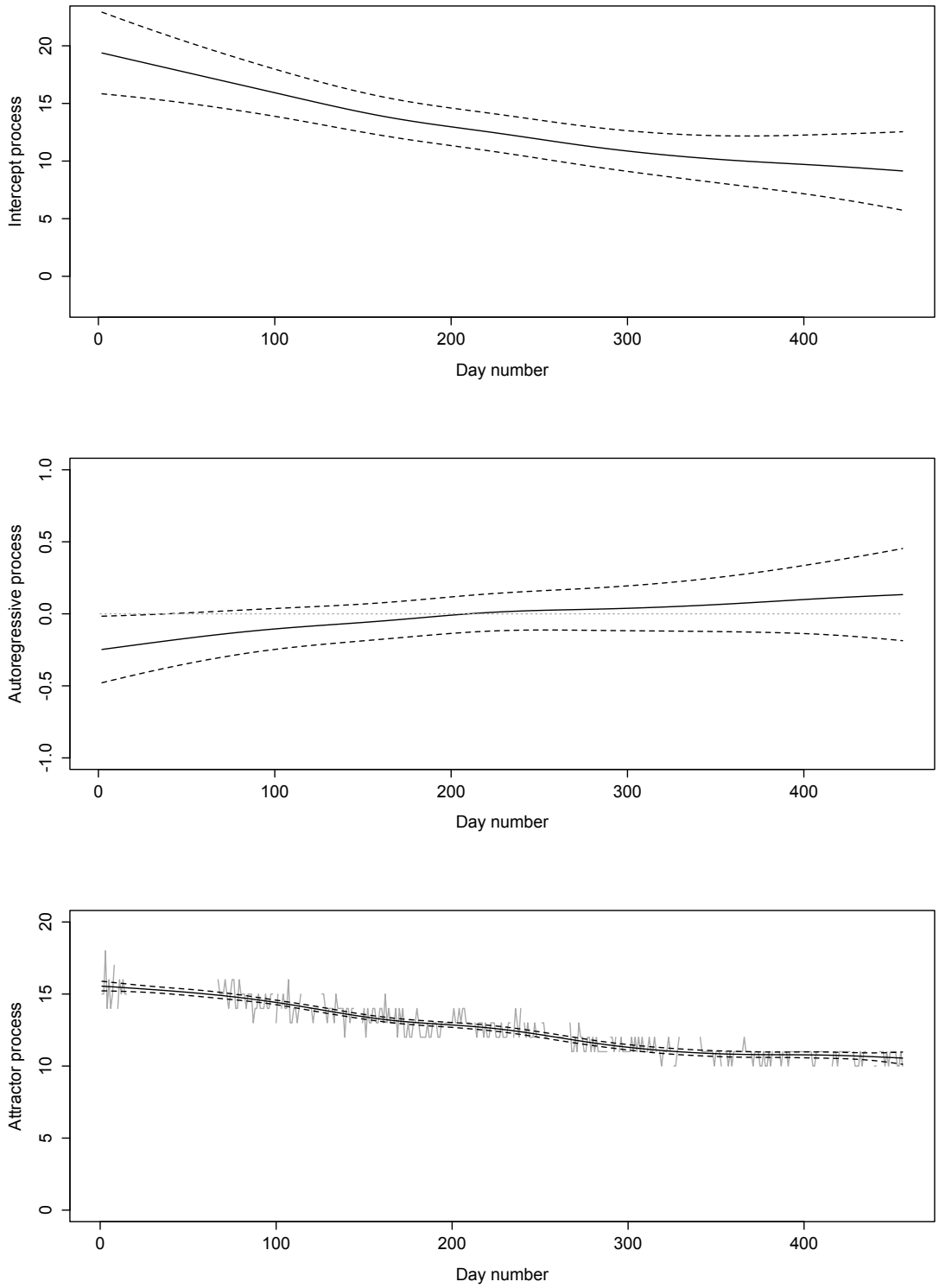


Figure 5.7: *Estimation results for the TV-AR model for participant 1.* Every panel represents a different parameter of the TV-AR model: the upper panel the intercept, the middle the autoregressive and the lowest the attractor. Note that the attractor process is plotted over the actual valence scores (represented in grey).

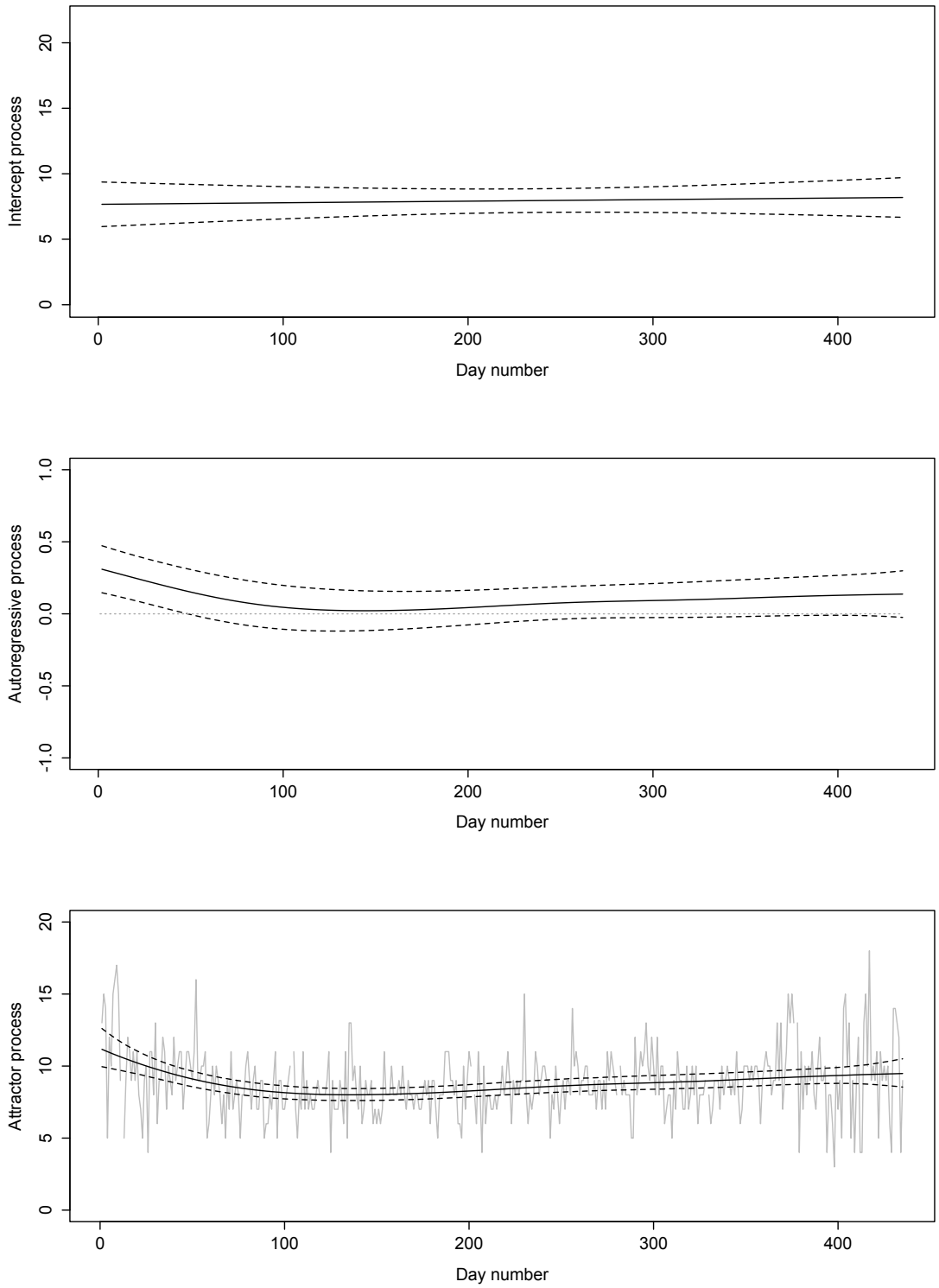


Figure 5.8: *Estimation results for the TV-AR model for participant 2.* Every panel represents a different parameter of the TV-AR model: the upper panel the intercept, the middle the autoregressive and the lowest the attractor. Note that the attractor process is plotted over the actual valence scores (represented in grey).

For example, the TV-AR model can help to detect and specify different kinds of non-stationarity in the data. Currently, it is common practice to focus on the trend that is apparent in the data, and to transform the time series so that it becomes trend stationary. However, even if the trend could be perfectly specified, which is often difficult, non-stationarity may not be fully accounted for, since the autocorrelation structure of the data can also change over time. Furthermore, a changing autocorrelation is not easy to detect visually, nor is there a test to detect such non-stationarity. With the semi-parametric TV-AR model, all such problems can be dealt with in one single step: Trends in the data and changes in the autoregressive process can be detected at once, and even more importantly, no pre-specifications are necessary, as has been shown in the real data application.

It is therefore clear that the semi-parametric TV-AR model is important in the case of non-stationary data. However, its potential range of application is much broader. As little is known about how and when psychological dynamics change, we would recommend to always run a TV-AR model next to a standard AR model as part of regular analysis if enough time points (circa 100) are available. In this way, the model can be used as a diagnostic tool for probing whether there is non-stationarity in the time series, and for detecting and specifying changing dynamics, such as the trend. For example, if the time series turns out to have a trend that is linear instead of non-parametric, a simpler parametric model can be specified based on the TV-AR analyses.

We have considered the simplest form of a TV-AR model, and will now elaborate on some of the extensions that are possible. We studied temporal dependency with a lag order 1 TV-AR model, but one can imagine that the temporal dependency is not only apparent between the two closest occasions, but also between occasions further apart, in which case a TV-AR model with lag order 2 or larger is necessary. Such extra lags can be easily added into a TV-AR model in the same manner as they are added into standard AR models through the inclusion of more lagged predictors.

Another sensible extension involves generalization of the model to multivariate data. The TV-AR model is currently only applicable to the univariate case, while it is often more realistic that a variable is not only predicted by itself, but also by other variables, which evokes the need to analyze psychological dynamics as a multivariate system. Such an extension would lead to a time varying vector AR (TV-VAR) model, and comes with new challenges, as both auto-correlations and cross-correlations would have to be modeled in this case. Yet another natural, but even more challenging, extension would be a TV-AR multilevel extension based on current multilevel (V)AR models (Bringmann, Vissers, et al., 2013; de Haan-Rietdijk et al., 2014; Jongerling, Laurenceau, & Hamaker, 2015). To the best of our knowledge, this is currently not possible, as the *mgcv* software cannot be used to estimate a flexible smooth function for the population (i.e., the population average) and to allow for flexible interindividual variation for that smooth function. An additional extension could be time-varying error variance, so that also the time-varying variance of a process could be fully accounted for. However,

with current software, only the intercept and the autoregressive parameter (and not the error variance) can be modeled as time-varying parameters. Further research should also consider the combination of gradual and abrupt changes, so that when the point of an abrupt change is known, it could be easily adjusted in the TV-AR model.

Even though the TV-AR model is easily applicable, the number of time points needed is a potential limitation. While 100 time points per participant would be preferable, currently most longitudinal studies in psychology gather around 60 time points or less (aan het Rot et al., 2012). Another limitation of the TV-AR is the assumption of gradual change. Although we have shown in the simulation study that with many time points and a large abrupt change the TV-AR model is quite robust and still gives an indication of the sudden jump, other models are probably more suitable for studying sudden change. Such models include the threshold autoregressive model (TAR) (e.g., Hamaker, 2009; Hamaker, Grasman, & Kamphuis, 2010), its multilevel extension, multilevel TAR (de Haan-Rietdijk et al., 2014), or the regime-switching state-space model (cf. Hamaker & Grasman, 2012; Kim & Nelson, 1999).

Furthermore, as the semi-parametric TV-AR model is an exploratory tool, the standard errors of the time-varying parameters are likely to be less satisfactory compared to confirmatory, raw-data maximum likelihood approaches, such as the state-space approach. Additionally, estimating a TV-AR model in a state-space modeling framework has the advantage that measurement error can be taken into account, which is not possible with the semi-parametric TV-AR model (Schuurman et al., 2015). Thus, future research should aim at comparing the exploratory semi-parametric TV-AR model with confirmatory approaches.

In sum, the semi-parametric TV-AR model presented here is an easy to use tool for detecting and modeling non-stationarity. Many extensions are possible, and future research is needed to uncover all the possibilities and limitations of this innovative framework. By introducing the model and explaining its application in standard software, we hope to have made it available to a broad range of psychologists studying human dynamics.

Appendix 5 Details of the simulation setup

The description of the simulation study is divided into three steps: (1) simulation conditions, (2) generating the simulation data and (3) model estimation and evaluation. An overview of these three steps of the simulation is given in Figure 5.A. The programming language *R* was used for all statistical simulations and analyses.

Step 1: Simulation conditions. We varied three factors: 1) the generating function of $\beta_{0,t}$ and $\beta_{1,t}$ (invariant, linear, cosine, random walk and stepwise); 2) the maximum absolute value of the time-varying parameters (low or high); and 3) the sample size (30, 60, 100, 200, 400, 1000). This resulted in 60 (5x2x6) different conditions. A total of $R = 1000$ replications of each condition were simulated. We elaborate on the factors below.

1. *Parameter generating functions.* The intercept ($\beta_{0,t}$) and autoregressive parameter ($\beta_{1,t}$) of the TV-AR model were generated with five different types of functions, three of them gradually changing and two non-gradually changing. The function generating the attractor (μ_t) was indirectly calculated afterwards, see Figure 5.A.

   The first of the gradually changing functions is a *time invariant function*, meaning that the $\beta_{0,t}$ and $\beta_{1,t}$ do not change over time and could therefore also be modeled with a standard AR. The second is a *linear function*. In this case $\beta_{0,t}$ and $\beta_{1,t}$ increase over time. The third of the gradually changing functions is a *cosine function*, where $\beta_{0,t}$ and $\beta_{1,t}$ first increase, then decrease and in the end increase again. The fourth and fifth functions are non-gradually changing, and thus violate the assumption of gradual change of the TV-AR model. The fourth function is a *random walk function*, in which $\beta_{0,t}$ and $\beta_{1,t}$ are generated in such a way that they show random and fast change that can also result in an increase or decrease in the function over a period of time. The fifth function is a *stepwise function*, meaning that $\beta_{0,t}$ and $\beta_{1,t}$ have for a certain period of time a constant value, which then changes abruptly to a higher value.

2. *Low and high maximum values.* Besides the different generating functions, we also compared low and high value settings for the maximum absolute values possible for the time-varying parameters. The maximum absolute values for the low condition for $\beta_{0,t}$ (the intercept) were 1 and for the high condition 1.5. Thus, for example, the peak values for the cosine function were 1 and -1 in the low condition (and 1.5 and -1.5 in the high condition). The maximum absolute value for the low and high condition for $\beta_{1,t}$ (the autoregressive parameter) was set to 0.2 and 0.5, respectively (based on values typically found in psychological studies, see e.g., Rovine & Walls, 2006). Whereas the invariant, linear, cosine and stepwise function are by definition bounded,

a random walk is not, so in order to have a bounded random walk with the above mentioned maximum absolute values we used an adapted version of the formula $\rho a_t / \max_{0 \leq j \leq t} |a_j|$, based on Giraitis et al. (2014). This formula guarantees that the random walk will be bounded between the pre-specified $-\rho$ and ρ. In this formula, a is defined as follows: $a_t - a_{t-1} = \eta_t$. Here, the difference between a_t and a_{t-1} equals η_t, a random number drawn from an independent identically normal distribution. At every time point ρ is multiplied with a_t and then divided by the maximum absolute value of a up to current time point t.

3. *Sample size.* Furthermore, sample sizes (the number of time points, n) were chosen to be comparable to those possible in psychological research: 30, 60, 100, 200, 400 and 1000. This will shed light on the amount of time points needed in order for the TV-AR to give a reliable recovery of the "true" underlying model.[14]

Step 2: Generating the simulation data. To generate the simulation data, we used the TV-AR formula introduced in section 3: $y_t = \beta_{0,t} + \beta_{1,t} y_{t-1} + \varepsilon_t$ (see also step 2 in Figure 5.A). The time-varying intercept $\beta_{0,t}$ and the time-varying autoregressive parameter $\beta_{1,t}$ can be generated after the parameter generating function, the maximum absolute value of the parameters and the sample size have been set. The residuals ε_t are a white noise process. This is simulated by drawing n times (with n being the number of time points) randomly from a standard normal distribution $\mathcal{N}(0, 1)$. Since the model is an autoregressive model with a lagged variable, we had to pre-specify the zeroth observation (y_0), which we drew from a stationary marginal normal distribution:

$$\mathcal{N}(\frac{\beta_0}{1 - \beta_1}, \frac{\sigma_\varepsilon^2}{1 - \beta_{1,t}^2}). \tag{5.A}$$

Marginal means here that the time point is not conditioned on the previous time point (see also the *R*-code). Now all further time points of y_t can be simulated. Note that the generated time series, as can been seen in Figure 5.A, follows the trajectory of the attractor (μ_t).

Step 3: Estimation and evaluation. We used seven different settings for estimating $\beta_{0,t}$ and $\beta_{1,t}$ [15]: 1) a TV-AR model using the default setting (a thin plate regression spline basis using 10 basis functions); 2) a TV-AR model with only a time-varying intercept and a time-invariant autoregressive parameter using the default settings; 3) a TV-AR model with only a time-varying autoregressive parameter using the default settings; 4) a standard time-invariant AR model; and 5) a thin plate regression spline basis using 30 basis functions.

[14] As pointed out by an anonymous reviewer, the local range of change of the cosine function is dependent upon sample size: there is a smaller rate of change for larger sample sizes.

[15] The attractor μ_t was again indirectly derived from the results of $\beta_{0,t}$ and $\beta_{1,t}$ (see Figure 5.A step 3).

Although 10 basis functions is the standard setting in the *mgcv* package, this might not always be enough to capture the wiggliness of a function, especially when a function takes a lot of turns, as is the case with for example the random walk function. Therefore, it is interesting to check whether an increase in basis functions leads to better estimations. However, increasing the number of basis functions requires that there is a large amount of time points. In this simulation, at least 400 time points were needed for increasing the number of basis functions to 30. Thus, we could only compare the difference between 10 and 30 basis functions for the sample sizes $n = 400$ and $n = 1000$ time points. Note that sometimes with already 100 time points it is possible to estimate a TV-AR model with 30 basis functions. However, from 400 time points on, the TV-AR model with 30 basis functions could be fitted for all 1000 replications, whereas with less than 400 time points this was not always the case.

To evaluate the global performance of the TV-AR model, we used the log of the median of the mean squared errors (MSEs) of the $R = 1000$ replications per condition. The MSE for a single time-varying parameter is defined as $\frac{1}{n}\sum_{t=1}^{n}(\hat{\theta}_t - \theta_t)^2$, in which $\hat{\theta}_t$ stands for the estimated value at time point t and θ_t for the true value at time point t. Any of the parameters $\beta_{0,t}$, $\beta_{1,t}$ or μ_t can take the role of θ_t and n stands for the number of time points. In addition, a coverage probability was calculated, which is the proportion of time that the true value is captured by the constructed CIs.[16]

An example is given in step 3 of Figure 5.A. All parameters have been estimated with our TV-AR model after the data were generated. The estimated values $\hat{\theta}_t$ and the true values θ_t of the parameters are represented as the middle solid black and red lines respectively. In this figure, the black and red solid line are close to each other, meaning that the model estimates had a low MSE value and the true underlying function(s) could be estimated well. The estimated CIs corresponding to the dashed lines show that almost everywhere the true values of the function are within these intervals, meaning that also the coverage probability was very high and the TV-AR model could estimate the true underlying model well.

Finally, we evaluated whether we could discriminate between a time-varying and time-invariant model (models 1, 2, 3 and 4 of step 3). We used the AIC, BIC and GCV to select the best fitting model for every replication. Next we calculated how often the correct model was selected by these fit indices. For example, for the cosine generating function condition we calculated how often the BIC correctly indicated that model 1, 2 or 3 (time-varying models) was the best fitting model versus model 4 (time-invariant model). In addition, the type I and type II errors were calculated. Finally, for both the intercept and autoregressive parameter, if applicable, the effective degrees of freedom (edf) and

[16] Although the CIs are given as output for the intercept and the autoregressive parameter, this is not the case for the attractor, since this time-varying parameter is only estimated indirectly. Therefore, for the smooth function of the attractor we calculated the CIs independently following the same procedures as in the *mgcv* package (see also the R-code).

p-values were extracted.

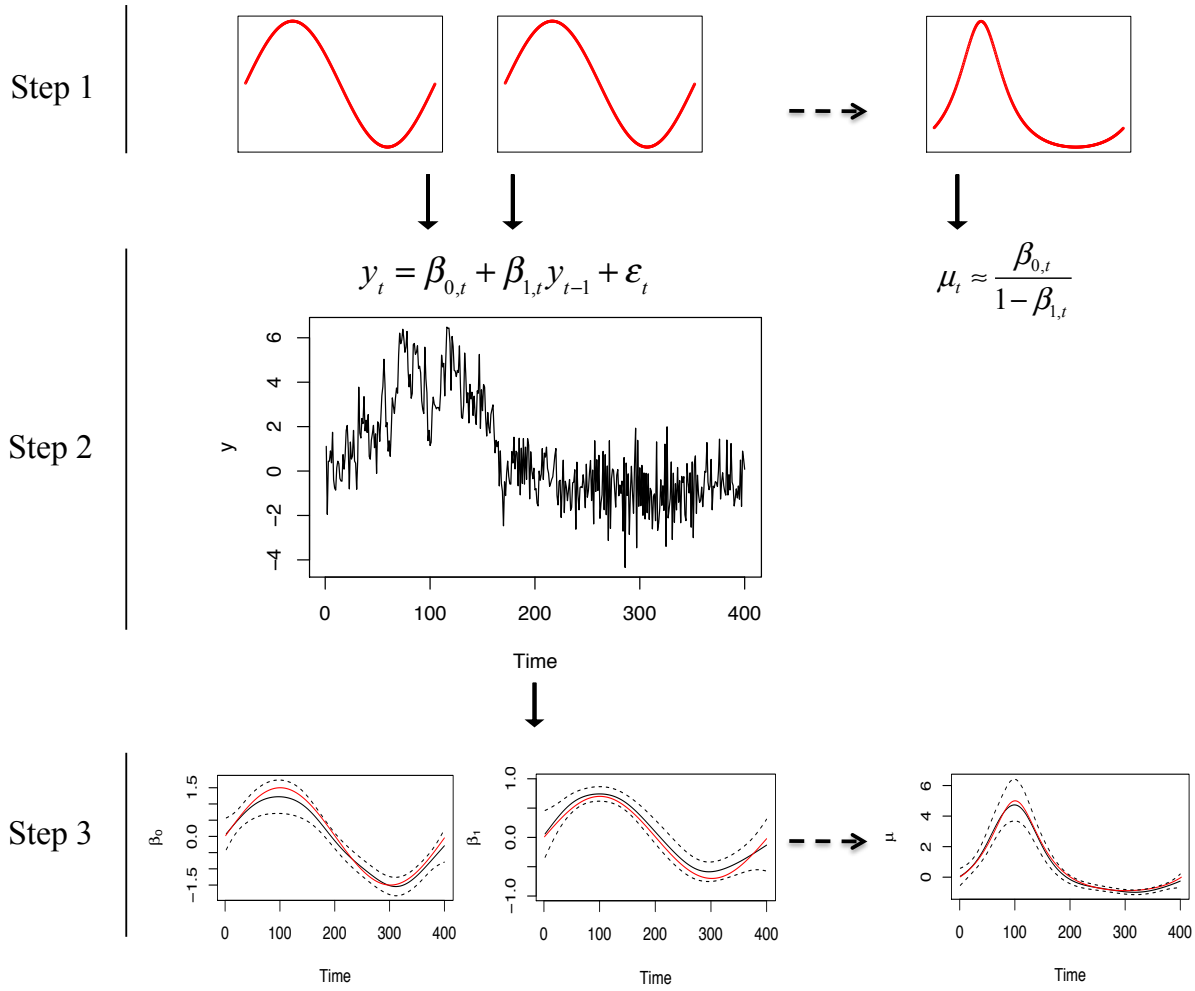


Figure 5.A: *The simulation setup.* The simulation setup consists of three steps. Step 1 represents the simulation conditions, step 2 the generation of the simulation data and step 3 the estimation and evaluation of the TV-AR model.

6 Matching Structural, Effective, and Functional Connectivity: A Comparison Between Structural Equation Modeling and Ancestral Graphs

The brain is a network of brain regions that are connected by anatomical tracts (Rubinov & Sporns, 2010; van den Heuvel & Pol, 2010). Brain connectivity can be studied at a structural (anatomical) and a functional level in a noninvasive way by magnetic resonance imaging (MRI) techniques. At the structural level, connectivity refers to the anatomical links of the brain that are made up of white matter tracts which can be modeled by diffusion tensor imaging (DTI) (Guye, Bartolomei, & Ranjeva, 2008; Johansen-Berg & Behrens, 2006; Rykhlevskaia, Gratton, & Fabiani, 2008; Tournier, Mori, & Leemans, 2011). At the functional level, connectivity reflects statistical associations (e.g., correlations) between regions based on indirect detection of neural activity through blood oxygen level-dependent (BOLD) signals measured with functional MRI (fMRI) (Bullmore & Sporns, 2009; Friston, 2011; He & Evans, 2010)

Connectivity at the functional level can be further divided into functional and effective connectivity. Functional connectivity is defined as the (temporal) correlation between different brain regions, whereas effective connectivity refers to the influence that one brain area exerts on another (Büchel & Friston, 1997; Bullmore & Sporns, 2009; Friston, 2011; Telesford, Simpson, Burdette, Hayasaka, & Laurienti, 2011). In contrast to functional connectivity, where connections are undirected, effective connectivity contains directed connections, implying a causal relationship between brain regions (Zhang et al., 2008; but see Ramsey et al., 2010 for a critical review on causality in effective connectivity in brain networks). Since effective connectivity is more informative, it is preferred to functional connectivity. Several methods have been proposed to examine effective connectivity; for example, structural equation modeling (SEM) (Büchel & Friston, 1997; Gonçalves & Hall, 2003; Mclntosh & Gonzalez-Lima, 1994), dynamic causal modeling (Friston, Harrison, & Penny, 2003; Friston, 2011), and Granger causality analysis (Eichler, 2005; Roebroeck, Formisano, & Goebel, 2005). Among these methods, SEM has been one of the most commonly used (Friston, 2011; Penny, Stephan, Mechelli, & Friston, 2004). In the SEM method, composing a model is always hypothesis–driven, and stan-

dard SEM models contain only directed connections (Mclntosh and Gonzalez-Lima (1994); see also McIntosh, Grady, Haxby, Ungerrleider, and Horwitz (1996) for examples of SEM models with recurrent connections).

A common problem of SEM and most other methods for assessing effective connectivity is that they implicitly assume that all relevant regions are in the model (except for Eichler's method; Eichler (2005)), because, for example, a region was not deemed relevant, or it did not pass a (corrected) threshold. This is a problematic assumption, because these missing regions can result in spurious connections, meaning that although the model indicates a direct connection between area A and B, the connection between the two areas is actually indirect, due to, for example, an unmeasured common cause, area C (Eichler, 2005; Waldorp et al., 2011). In contrast, ancestral graphs (AGs) represent a class of models for effective connectivity that can detect missing regions in a model (Waldorp et al., 2011). Most methods examining functional or effective connectivity use only undirected (functional connectivity) or directed (effective connectivity) connections, whereas AGs can model undirected, directed, and bidirected connections (Richardson & Spirtes, 2002). Intuitively, a directed connection represents effective connectivity, an undirected connection represents functional connectivity, and a bidirected connection can be interpreted as an indirect connection that is due to an unobserved area. Hence, AGs are able to explicitly indicate missing brain regions (Waldorp et al., 2011).

In this study, we compared the conventional SEM method for studying effective connectivity with AG. Five participants were measured using fMRI while performing one of three visual perception tasks. Since brain connectivity patterns are likely to be different between individuals (Horwitz et al., 2005), effective and structural connectivity were estimated for each of the five subjects separately. To estimate effective connectivity, we analyzed fMRI data for six regions of interest (ROIs) of a task in which motion-defined figures were presented. The ROIs used in the current study included areas V1, V2, and V3 taken together, because this is where the vast majority of visual information enters the brain. We also selected area LO in the lateral occipital cortex because of its involvement in shape processing (Grill-Spector, Kourtzi, & Kanwisher, 2001; Malach et al., 1995) and area middle temporal (MT), because it plays a central role in motion processing (Albright, 1984; Albright & Stoner, 1995; J. D. Watson et al., 1993). Finally, we included area inferior temporal (IT) because of its role in object perception and integration of information from lower-tier areas (Tanaka, 1996). For each subject and each condition of the task, the best model was selected for AG and SEM. The determination of whether SEM or AGs was better for estimating effective connecitvity was based on the ability of predicting structural connections found with tractography based on DTI. Connection probabilities were estimated with DTI probabilistic tractography between the same six ROIs as used in the effective connectivity analyses. Thus, DTI tractography was used to examine whether connections found with effective connectivity are also likely to be present at a structural level.

We show that, in general, AGs result in more accurate models than SEM. The reason for this is that missing regions are taken into account when modeling with AG but not when modeling with SEM: AG can be used to explicitly test the assumption of missing regions. If the set of regions is complete, SEM and AG perform about equally well.

6.1 Material and Methods

Subjects

Five healthy subjects (three men; mean age: 27.4 years; range: 24-31 years) without any history of neurological or psychiatric disease participated in the study. Three subjects were right handed, and the other two were left handed as was indicated by the Edinburgh Handedness Inventory (Oldfield, 1971). All experimental procedures were approved by the ethics committee of the Faculty of Psychology of the University of Amsterdam, with all subjects providing written informed consent. Subjects had normal or corrected-to-normal vision.

Task and procedure

Before the actual fMRI experiment, subjects practiced the experiment outside the scanner for 20 min to familiarize themselves with the task. During the fMRI experiment, stimuli were projected on a screen at the end of the scanner. Subjects viewed the screen via a mirror system attached to the MRI head coil. To reduce motion artifacts subjects heads were immobilized using foam pads. Subjects received earplugs and a headphone to decrease scanner noise. The start of a run was triggered by scanner pulses, and stimuli were presented with presentation (Neurobehavioral Systems, Inc.).

Subjects had to discriminate between three stimulus conditions: a *Frame*, a *Stack*, and a *Homogenous* condition (see Figure 6.1), by pressing one of three buttons, each corresponding to a condition. Each stimulus consisted of a displacement of randomly distributed black and white dots, which had the size of a pixel. The displacement happened in one out of four directions: 22.5°, 67.5°, 112.5°, or 157.5°. A stimulus contained three regions: the background, the frame, and the inner region. Stimulus presentation started with the background region (randomly distributed black and white dots), which changed after 100 msec into one of the three stimuli conditions and after another 100 msec into the background again.

In the *Frame* condition, the dots of the frame region moved in a different direction from the background and the region inside the frame. The *Stack* condition was similar to the *Frame* condition, except that the dots within the inner region moved in a different direction from the background as well as from the frame. In the *Homogenous* condition, the dots in the frame and inner region moved

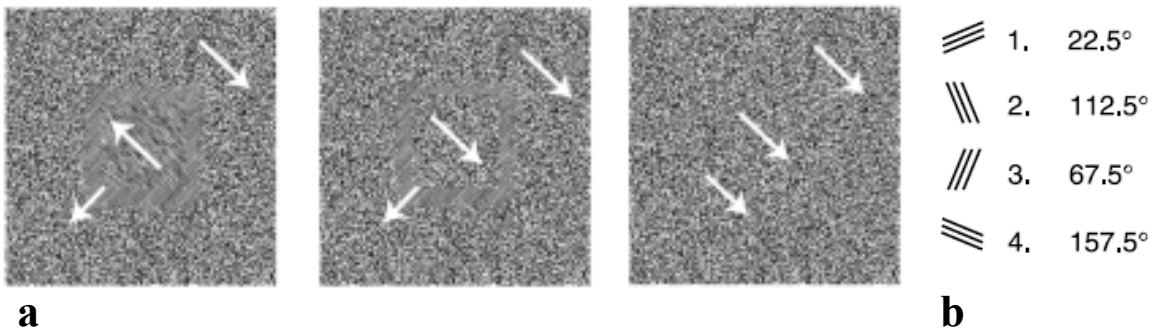


Figure 6.1: The first part **(a)** represents the *Stack*, the *Frame*, and the *Homogenous* condition, respectively. The second part **(b)** represents the four different directions of dot displacement.

homogenously with regard to the background, in which case almost no frame or inner region movement was visible (Scholte, Jolij, Fahrenfort, & Lamme, 2008). Each trial was presented in 300 msec and was followed by an inter-trial interval of 6 sec. To optimize the measured signal, seven stimuli per trial were presented on the screen: three squares on the left, three on the right side of the screen, and one in the middle.

Magnetic resonance imaging

Scans were conducted on a 3T magnetic resonance scanner (Philips Achieva) that was equipped with a 32-channel SENSE head coil. To obtain DTI and fMRI data, three scanning sessions were performed. The first scanning session was used to acquire DTI data, and the last two scanning sessions were used to acquire the fMRI data. Besides the main task, cortical mappers were used.

fMRI acquisition

The experimental setup was an event-related design, meaning that participants were randomly presented as a *Stack*, *Frame*, or *Homogenous* stimulus; while fMRI recordings of the BOLD response were made at regular intervals. The stimuli were presented in a pseudo-random order for 20 times per stimulus type over two runs. One run lasted for approximately 10.6 min and consisted of acquiring 288 volumes (GE-EPI, 200^2 mm field of view [FOV]; 80^2 in-plane resolution; 38 slices, 2.5mm slice thickness; 0.25mm slice spacing; repetition time [TR], 2200 msec; echo time [TE], 29.93 msec; flip angle [FA], 80°, SENSE factor 2). Furthermore, high-resolution T1-weighted anatomical images (T1; turbo field echo, 160^2 mm FOV; 256^2 in-plane resolution; 160 slices, 1mm slice thickness; TR, 8.159 msec; TE, 3.73 msec; FA, 8°) were obtained from each subject for registration purposes.

fMRI preprocessing and analysis

Preprocessing and statistical analyses of the functional data were performed with FEAT (FMRI Expert Analysis Tool) version 5.98, part of FSL (Oxford Center for Functional MRI of the brain (FMRIB) Software Library (`www.fmrib.ox.ac.uk/fsl`) (S. M. Smith et al., 2004)) and Matlab (Mathworks, Inc.). Preprocessing of functional images included head motion correction, slice time correction, brain extraction, spatial smoothing using a Gaussian kernel of Full Width at Half Maximum of 2 mm, and a high-pass cut-off in the temporal domain using a Gaussian kernel with a standard deviation of 100s. As a final preprocessing step, the functional data were aligned to the T1 image of the subject, and the data of each subject were transformed to MNI152 (Montreal Neurological Institute) using FNIRT nonlinear registration. Both functional and structural analyses were done for apriori defined ROIs. Selection of the ROIs was based on a large number of studies on object processing and motion perception (e.g., Albright, 1984; Grill-Spector et al., 2001; Malach et al., 1995; Tanaka, 1996; J. D. Watson et al., 1993). The six ROIs were V123, IT cortex, MT area of the left and right hemisphere, and LO area of the left and right hemisphere. Region V123 was defined according to the Jülich histological atlas (Amunts et al., 2005), and the IT cortex was defined according to the Harvard–Oxford cortical structural atlas (available at `www.cma.mgh.harvard.edu`).

Since the IT cortex has some overlap with the MT and LO areas, the posterior part of the IT was left out of the IT ROI to prevent overlap. Both the Jülich and the Harvard–Oxford atlas are probabilistic, and, therefore, the threshold of a voxel belonging to the V123 or IT ROI was set at a probability of 25%. Since there are large individual differences for the exact location of the MT and LO areas, we used standard functional mappers to localize these brain regions (J. D. Watson et al., 1993). The fMRI settings for these functional mapping scans were the same as described earlier. The data from these mappers were analyzed with a general linear model (GLM), including regressors for each condition.

The functional data were modeled using a GLM at a single subject level using FMRIB's improved linear model with local autocorrelation correction using an AR(1) model (Woolrich, Ripley, Brady, & Smith, 2001). The event onsets of each trial from a specific condition were convolved with a canonical hemodynamic response function (double gamma) to generate the regressors used in the GLM. Results were rendered on Z statistic images thresholded by $Z > 5.3$ with an uncorrected significance threshold of $p = 0.05$. This resulted in the mean parameter estimates (PE) of only the active voxels of each ROI per trial, condition, and subject.

DTI acquisition

In a different session, diffusion-weighted images (TR 6345 msec, TE 76 msec, FA 90°, 120^2 mm FOV, 224^2 in-plane resolution, 60 slices, b = 1000 msec) were acquired along 32, 48, and 64 collinear directions for obtaining detailed DW images. Each series of directions was preceded by acquisition of a non-diffusion-weighted volume for purposes of registration and motion correction. The sum of diffusion-weighted volumes was 575 ($4 * 32, 3 * 48, 2 * 64$). Total acquisition time was 110 min.

DTI preprocessing and analysis

All DTI preprocessing and analyses were conducted using FSL tools (`www.fmrib.ox.ac.uk/fsl`). Diffusion data were corrected for eddy currents and possible head motion (Jenkinson & Smith, 2001). Next, all non-brain data were discarded, and images were aligned to MNI152 standard space. This ensured that the DTI data were in the same space as the functional data. Images from all subjects were visually inspected to confirm a close registration.

To study the structural connectivity between the six ROIs (V123, IT, MTleft, MTright, LOleft, and LOright; see also fMRI section preprocessing and analysis), we used probabilistic fiber tracking by applying the FMRIB Diffusion Toolkit. Subsequently, the BEDPOSTX tool, which runs a Markov Chain Monte–Carlo estimation process, was used to create distributions of diffusion parameters describing the principle water diffusion direction in each voxel (Behrens et al., 2003). For each voxel included in the ROI or seed mask, 5000 streamline samples were taken from the distribution. This resulted in a probabilistic map indicating the connections of each voxel included in the seed mask with the rest of the brain. Next, the probability map was filtered so that only the streamlines connecting the voxels of two different ROIs, the seed ROI and the target ROI, were taken into account. Probabilistic tractography between two ROIs was done in both directions; for example, from V123 to IT and back. We summed up the number of streamlines that left the seed ROI and reached the target ROI. The number of completed streamlines reflects the confidence that a connection exists at a structural level (Tournier et al., 2011; see also Jones, 2010 for a critical review on DTI as a measure of structural connectivity). These derived pathway strengths are then an (indirect) measure of structural connectivity (Jbabdi & Johansen-Berg, 2011). This number was divided by the volume of the seed and target mask to normalize for between-subject variability in area size.

SEM and AG connectivity analysis

Figure 6.2 demonstrates the procedure for SEM and AG connectivity analysis. The first three steps are identical for the SEM and AG methods. As described earlier, the event-related BOLD fMRI data were used as input to a GLM, which resulted in PEs of neural activation for all six ROIs (averaged

over voxels within ROIs) per condition specific trial. Error trials were excluded from the connectivity analysis. Importantly, connectivity analysis in both SEM and AG is based on the replication of the condition-specific trials and not on the time series. In this way, SEM and AG do not depend on the low temporal resolution of time series in fMRI but on the number of replications per condition (Waldorp et al., 2011). Based on the PEs of single trial data, the covariance matrices for each condition and each subject were determined. Since there were three task conditions and five subjects, this resulted in 15 data covariance matrices.

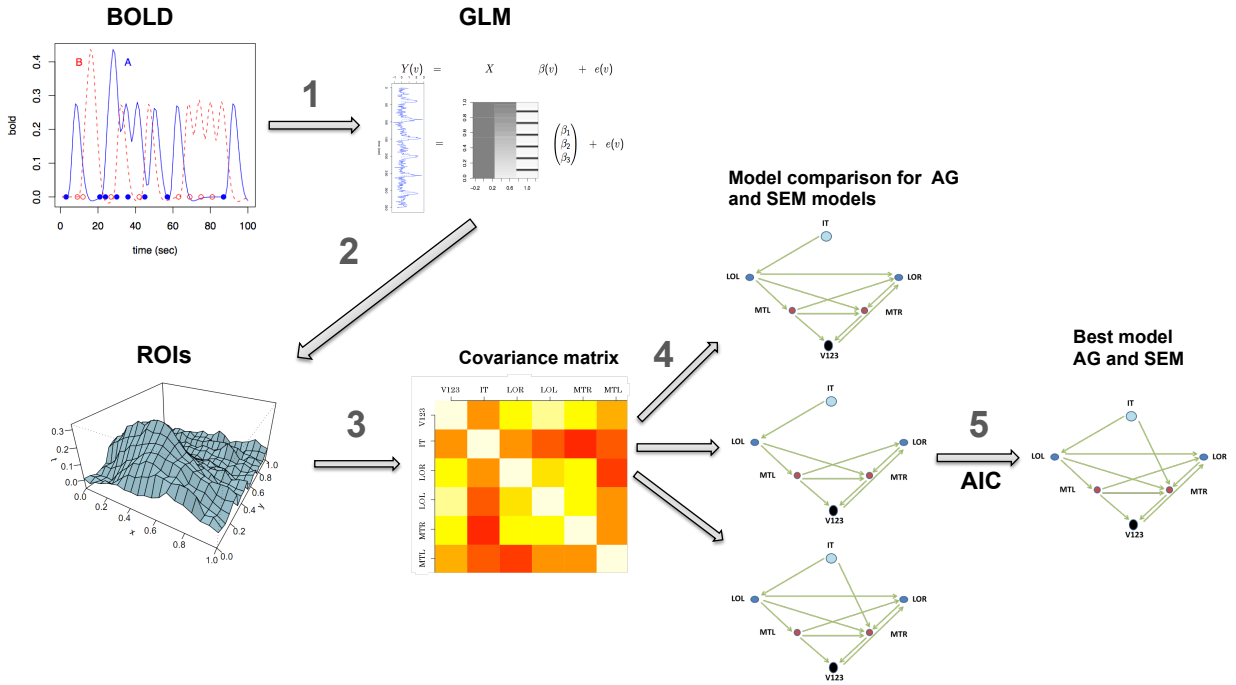


Figure 6.2: Procedure of effective connectivity analysis for AG and SEM. (1) The event-related BOLD fMRI data are used as input to the GLM. (2) This results in the mean parameter estimates of the neural activation of all six ROIs per condition-specific trial. (3) Based on this, the data covariance matrices for each condition and each subject are determined. (4) Based on the covariance matrices, the model fit of different models is compared. (5) The lower the AIC, the better the model fit. Choosing the best model is based on a joint probability of AIC and robustness probability. AG, ancestral graphs; SEM, structural equation model; fMRI, functional magnetic resonance imaging; BOLD, blood oxygen level dependent; GLM, general linear model; AIC, Akaike's information criterion.

In SEM as well as AG, the parameters of the connectivity models, including the path coefficients or path strengths and the error terms, are estimated by minimizing the difference between the observed and estimated covariance matrix using maximum likelihood. The methods differ, however, in the way in which the population error covariance matrix is modeled. While SEM assumes (most often) that the errors of the regressions (effective connectivity) are uncorrelated, AG distinguishes between correlated and uncorrelated errors (Richardson & Spirtes, 2002), providing a way to determine whether there are missing regions. Consequently, in standard SEM, Σ can only be estimated based on directed

connections, regressions, which implies effective connectivity and is indicated by B in SEM (Mclntosh & Gonzalez-Lima, 1994). Furthermore, the error structure, denoted by Φ_ε, is almost always specified as a diagonal covariance matrix, meaning that the estimated error structure in the model is uncorrelated (Gates, Molenaar, Hillary, Ram, & Rovine, 2010; Mclntosh & Gonzalez-Lima, 1994).

In AG, two other connection types can be identified besides directed connections (denoted by B): undirected connections (denoted by Λ) and bidirected connections (denoted by Ω). Directed connections are ordinary regression parameters, implying effective connectivity; while undirected connections are partial covariances (unscaled partial correlations), implying functional connectivity. Bidirected connections refer to the covariance of the residuals from the regressions. Such a covariance implies that there is an unexplained structure in the residuals, meaning that a parameter (a brain region) is missing from the model (Waldorp et al., 2011). It should be noted that bidirected connections in AG are not the same as reciprocal connections in SEM; in AG, undirected edges are used to indicate reciprocal information flow and are, therefore, similar to the reciprocal connections in SEM. The covariance matrix is modeled by SEM and AG, respectively:

$$SEM : \Sigma = \quad (I - B)^{-1}\Phi_\varepsilon(I - B^t)^{-1} \tag{6.1}$$

$$AG : \Sigma = \quad (I - B)^{-1} \begin{pmatrix} \Lambda^{-1} & 0 \\ 0 & \Omega \end{pmatrix} (I - B^t)^{-1} \tag{6.2}$$

where I is the identity matrix, t indicates transposition, and -1 indicates inversion. Thus, the AG method can model both effective (directed connections) and functional connectivity (undirected connections), and it can indicate missing regions (bidirected connections; see Waldorp et al., 2011); whereas standard SEM only models direct effective connections. How well the estimated covariance matrix of the model fits the data covariance matrix is indicated by Akaike's information criterion (AIC; Akaike, 1974), which involves the log-likelihood $L(\theta)$ with q parameters collected in the vector θ for model or graph G_q:

$$AIC(G_q) = \quad -2L(\theta) + 2q. \tag{6.3}$$

The differences between the AICs of different models are often small, which makes it difficult to distinguish between models. To overcome this problem, we used Akaike weights, that is, we normalized the AIC differences and treated them as probabilities (Burnham & Anderson, 2004; Wagenmakers & Farrell, 2004). This was done with the following formula, where W stands for Akaike weights, Δ is the difference between the lowest AIC and a current model in AICi, $exp()$ is the natural exponent,

and R is the number of models in the probability space:

$$W_i = \frac{exp(-\Delta_i/2)}{\sum_{r=1}^{R} exp(-\Delta_r/2)}. \tag{6.4}$$

The probabilities were given for the 20 models ($R = 20$) having the lowest AIC. In order to choose the best model of the 20 models selected with the AIC, we also considered the robustness of the model. The degree of robustness depends on how many connections in the model are also present in the other models. A model has a higher degree of robustness if the connections of the model are in all or most of the other 20 models as well, indicating that the connection is invariant over different configurations in a graph. Thus, besides Akaike weights, we also take robustness into account when selecting the best-fitting model. This leads to the following joint probability of AIC and robustness probability:

$$P(\text{AIC and robustness}|\text{data}) = P(\text{AIC}|\text{data})P(\text{robustness}|\text{data}). \tag{6.5}$$

The model with the highest joint probability was selected as the best-fitting model for each subject and condition.

In SEM, composing a model is always hypothesis–driven, because no data–driven method is available. We, therefore, constructed models for the SEM method based on previous research. The areas that make up the visual cortex are vastly interconnected. One approach to modeling their organization is to presume that processing is both distributed and hierarchical (Felleman & van Essen, 1991). From this perspective, it would be expected that information is relayed from the earliest, very broadly tuned, visual areas (V123) to somewhat higher-tier processing stations devoted to motion signals (MT) and object shape perception (LO) before being finally relayed to pure object processing areas (IT). It is, however, clear that there are many projections between different visual areas (Felleman & van Essen, 1991) and that these, furthermore, not only project from lower-tier to higher-tier areas but also vice versa (Lamme & Roelfsema, 2000).

Based on this, we constructed three models having directed connections that either start from the V123 area going up to the IT area via the MT and LO areas (model 1, see also Fig. 8) or start from the IT area going down to the V123 via LO and/or MT areas (models 2 and 3). These hypothesis–driven models were tested for the different conditions (Homogenous, Frame and Stack condition) with SEM (Mplus version 6.11; Muthén & Muthén, 2012), and for each condition, the model with the highest joint probability was selected as the best-fitting model.

Since there are many projections between different visual areas, it is difficult to predict how exactly the information flows between the visual areas. Thus, in this case, a more explorative or data–driven

method is likely to be beneficial. Furthermore, a data–driven method can lead to new insights on possible connections or directions between the ROIs. Even though there is no data–driven method for SEM, such a data–driven search process has been developed for the AG method in the language *R* (*R* Development Core Team, 2012). To make the comparison between AG and SEM more optimal, we used the models found with (data–driven) AG to be fitted with SEM. This resulted in hypothesis–driven SEM models, data–driven SEM models, and data–driven AG models. AG was compared with both hypothesis–driven and data–driven SEM models. Next, we describe the data–driven search process in greater detail.

Since the number of possible models in the AG method is between 14 million and 1.07 billion, it is impossible to test all models. Instead, we developed a method that can find the best-fitting model without testing all models. In this data–driven approach, we started for each of the three conditions with the six ROIs without any connections. Next, a single directed connection entered the model. The fit of this connection was determined for all pairs of ROIs. The connection having the lowest AIC remained in the model. Subsequent connections are obtained in a similar manner. This procedure continued until adding a new connection did not lead to a lower AIC. The same procedure was executed for a combination of directed, undirected, and bidirected connections. For SEM, the best three models that were found with the data–driven AG procedure and contained only directed edges were fitted. Again, for each condition, the model with the highest joint probability was selected as the best-fitting model. This resulted in 15 models for the AG method and 15 models for the hypothesis- and data–driven SEM method (3 conditions x 5 subjects).

Combining effective and structural connectivity

In order to analyze the performance of both AG and SEM with regard to predicting structural connectivity, the connection strengths of the effective connectivity analyses were correlated with the connection probabilities resulting from probabilistic tractography (probtackx in FSL), which reflect the confidence that a connection exists between brain areas. We tested differences between the different methods: AG, SEM (data–driven SEM or SEM supplemented by AG), and SEMS (standard or hypothesis–driven SEM) using normalized scores. Normalization ensured that different scales of each method were not causing differences.

Scores for each method were normalized per subject and per condition (Euclidean distance) across the 30 connections. We used a Wald-type test for the difference between two parameter vectors of connections, which incorporates the covariance matrix of the connections, such as Hotelling's test for MANOVA. Under the null hypothesis of no difference between the vectors of connections, this test has a chi-square distribution (since the parameters of the connections are approximately normally

distributed).

The covariance matrix was obtained by assuming that it is the same for all subjects and conditions, such that the data from the different subjects and conditions could be pooled. Then, a lasso estimate of the covariance matrix was obtained using the *glasso* function in R (Friedman, Hastie, & Tibshirani, 2008). The degrees of freedom were determined using the Satterthwaite approximation, as is common in linear mixed models. We tested at a Bonferroni corrected level of $0.05/15 = 0.0033$.

6.2 Results

Structural connectivity

To study structural connectivity, probabilistic tractography was performed for six ROIs. Figure 6.3 displays the positions of the ROIs of one subject projected on the MNI 152 brain, whereas ROIs V123 and IT were defined with a probabilistic atlas; ROIs MTleft, MTright, LOleft, and LOright were defined with functional mappers (as described in section "fMRI preprocessing and analysis"). In line with previous research, MT was defined as the activation cluster in IT sulcus, and lateral occipital sulcus and the LO area was defined as the activation cluster between area MT and V123 (Scholte et al., 2008).

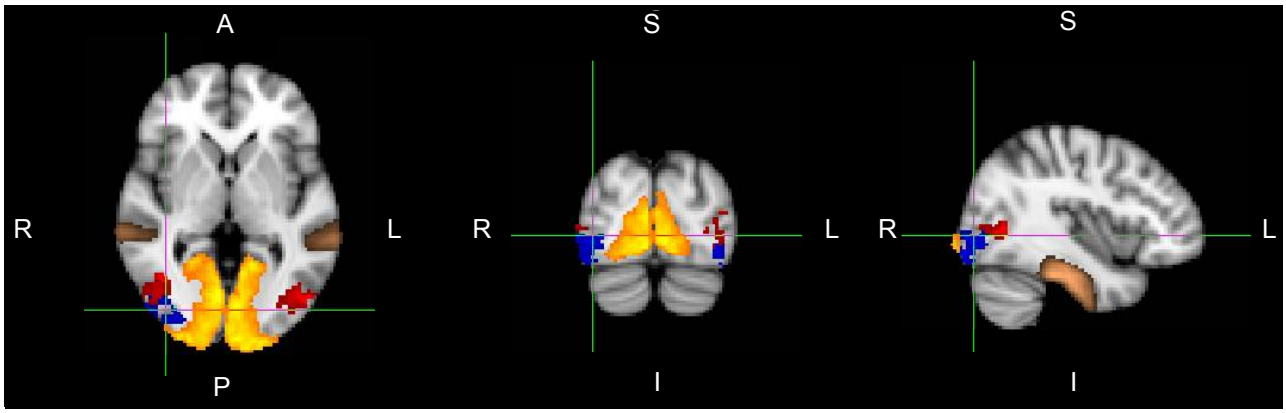


Figure 6.3: The six ROI. The yellow region is V123, the brown region is IT cortex, the red regions are the temporal motion areas, MTleft and the MTright, and the blue regions are the lateral occipital areas, LOleft and LOright. S, superior; I, inferior; R, right; L, left; A, anterior; ROI, regions of interest; MT, middle temporal; IT, inferior temporal.

The pathway strength between two ROIs was derived from the number of completed paths between two ROIs. In Figure 6.4, an example of probabilistic tractography is presented. Tracking starts in this example from ROI MTleft. The two target ROIs presented in the figure are MTright and the LOleft. As the figure shows, almost no tracts from the MTleft area reach the MTright area, whereas a large number of tracts from the MTleft area reach the LOleft area.

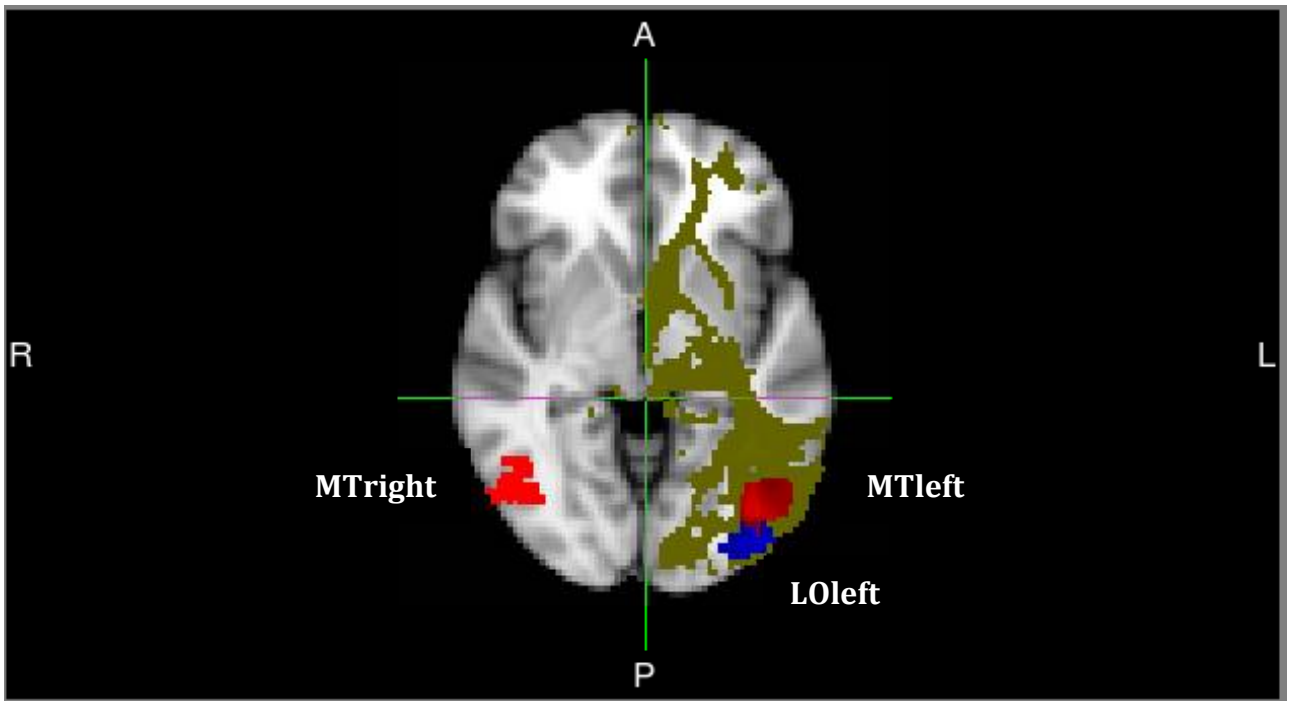


Figure 6.4: An example of probabilistic tractography as displayed in standard MNI space. The red areas indicate the MT areas, and the blue area indicates the LO area. Tracking starts in this example from ROI MTleft. The figure shows that more tracts are going from MTleft to LOleft than to MTright. P, posterior; R, right; L, left; A, anterior.

In line with previous findings (Kaiser & Hilgetag, 2004), connectivity between ROIs decreases with distance between ROIs (Figure 6.5). For example, the connection probability between LOleft and V123 is smaller than the connection probability between LOleft and MTleft. Thus, it is less likely that there is a direct structural connection between LOleft and V123 than between LOleft and MTleft.

Effective connectivity

Effective connectivity analysis using either SEM or AG is based on the replication of the condition-specific trials of the motion perception task (see Figures 6.6 and 6.7). The motion perception task comprised three conditions (*Homogenous*, *Frame* and *Stack*), each consisting of 40 trials. On average, subjects responded correctly to 92.0% of the trials in the Homogenous condition (36.8 trials with a standard deviation (SD) of 2.59), to 86.5% in the Frame condition (34.6 trials with an SD of 5), and to 76.0% in the Stack condition (30.4 trials with an SD of 8.23).

An example of the best model for the AG and the data–driven and hypothesis–driven SEM models for the first subject of the *Homogenous* condition is displayed in Figure 6.8. In this particular case, the AG model includes both bidirected connections (displayed in dashed orange arrows) and directed

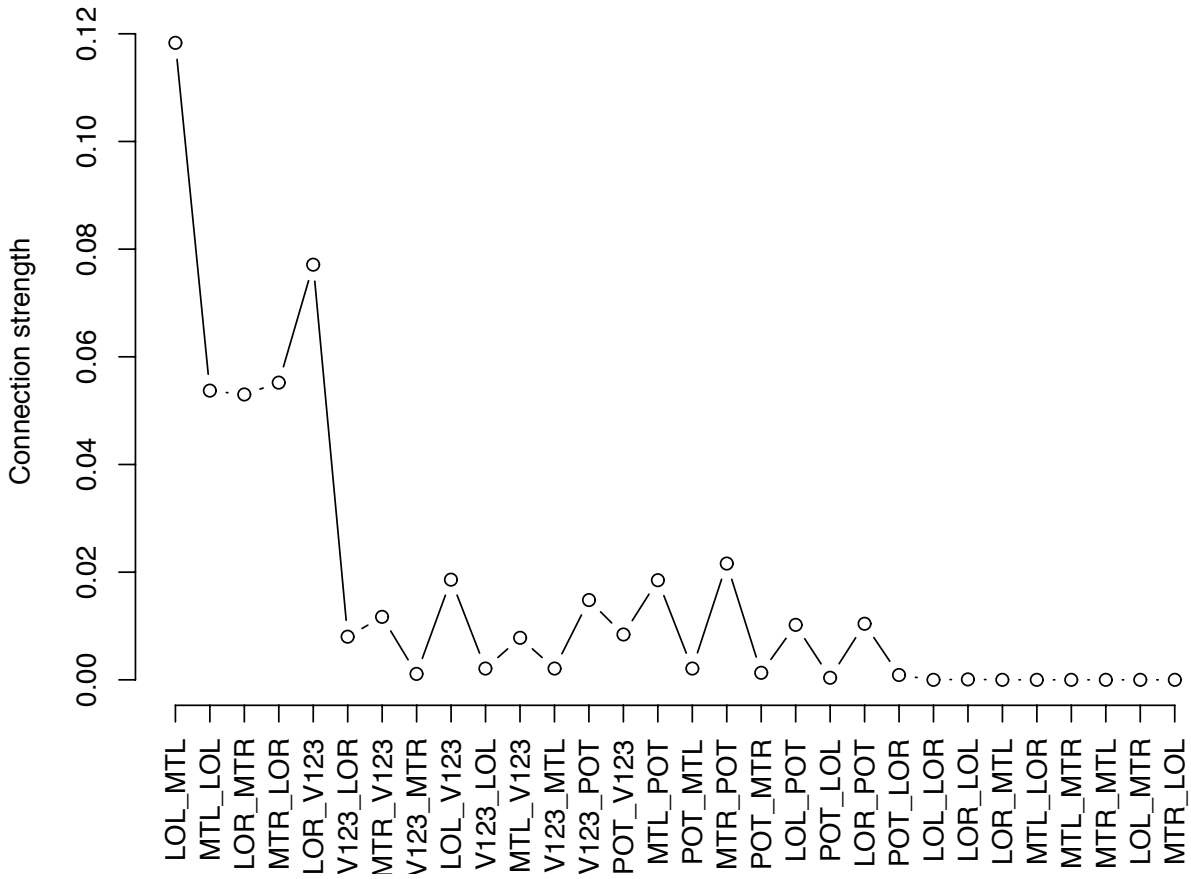


Figure 6.5: Connection probabilities of DTI connectivity analysis as a function of pairs of ROIs with increasing (Euclidean) distance between ROIs from left to right. Connection probabilities are averaged over all subjects. LL, LOleft; ML, MTleft; LR, LOright; MR, MTright; V, V123. DTI, diffusion tensor images.

connections (displayed in green solid arrows), whereas the SEM model includes only directed connections. The bidirected connections have a strength of zero, indicating that there is a missing region causing a correlation but no direct connection between the two ROIs.

Validating the AG method using SEM and DTI

To examine whether the AG method can predict structural connectivity equally well or even better than the conventional SEM method, the standardized connection strengths of both the AG and SEM models were correlated with the standardized connection probabilities from tractography (see Figure 6.9). Standardization was performed for each subject and condition separately (compare the method of Urbach and Kutas (2002)). The standardized connection strengths have a scale from 0 to 1.

We tested differences between the different methods: AG, SEM, and SEMS (hypothsis–driven) with

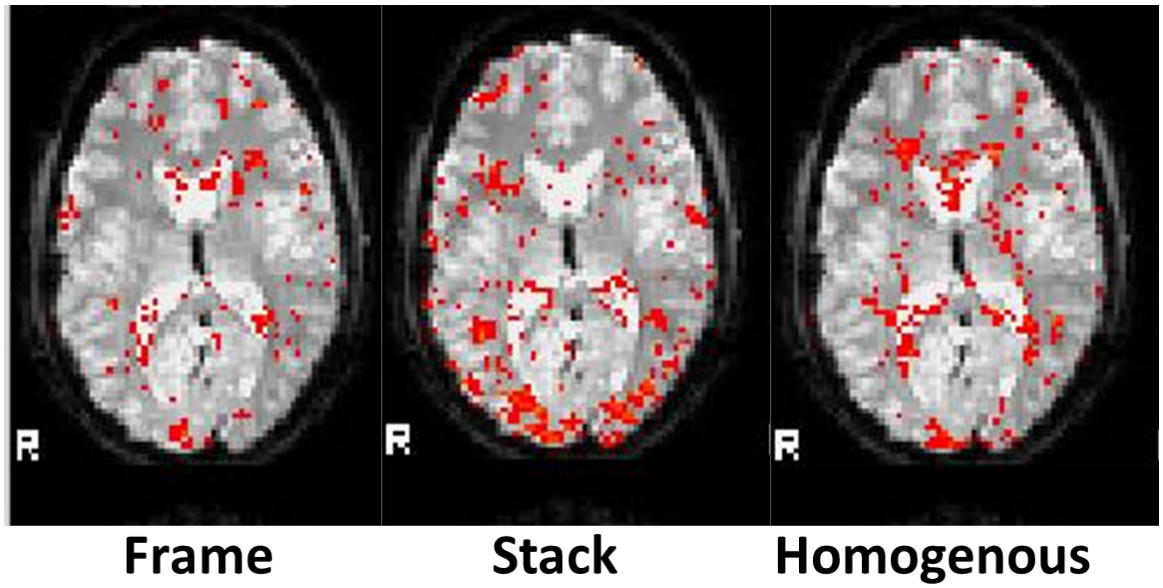


Figure 6.6: The fMRI data of one trial per condition for one subject. For each condition, the 20th trial of the first run is shown. The fMRI images are images of correctly performed trials. R, right

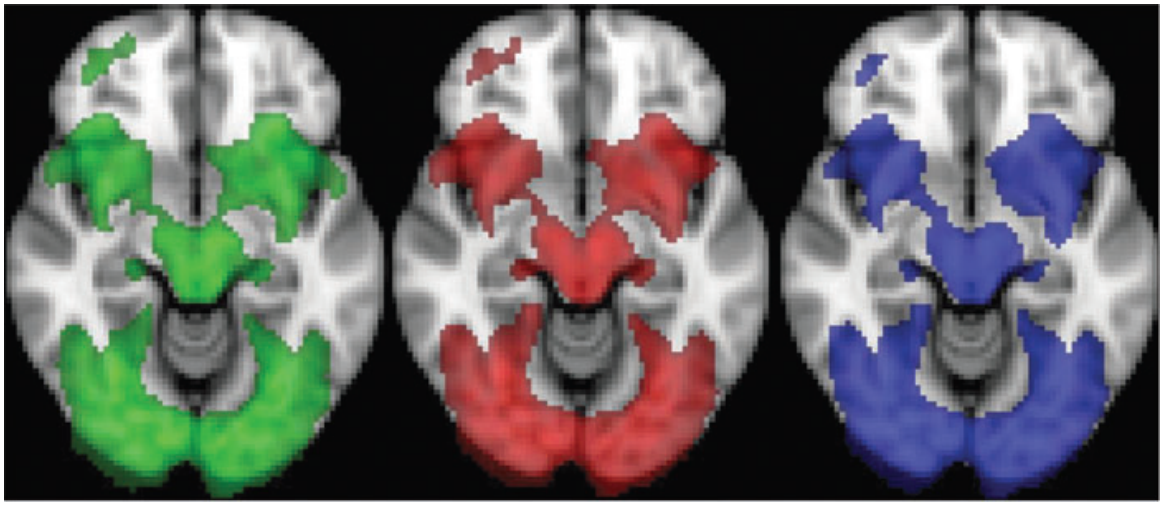


Figure 6.7: From left to right, the figure shows the activation pattern of, respectively, *Homogenous*, *Frame* and *Stack*) condition over all runs (within a subject fixed, over subjects mixed).

the normalized scores using a Wald type that has a chi-square distribution under the null hypothesis of no difference. We tested individual and condition-specific tests at 0.05 and subsequent tests for each individual and condition separately (15 in total) at a Bonferroni–corrected level of $0.05/15 = 0.0033$. Overall tests revealed that there were no differences between AG and SEM (χ^2 [10.71] = 8.56, $p = 0.63824$). However, the difference between AG and SEMS was significant (χ^2 [16.60] = 28.99, $p = 0.0300$), and also the difference between SEM and SEMS was significant (χ^2 [17.09] = 29.87, $p = 0.0282$). This indicates that AG and SEM perform equally well, but SEMS performs worse. The

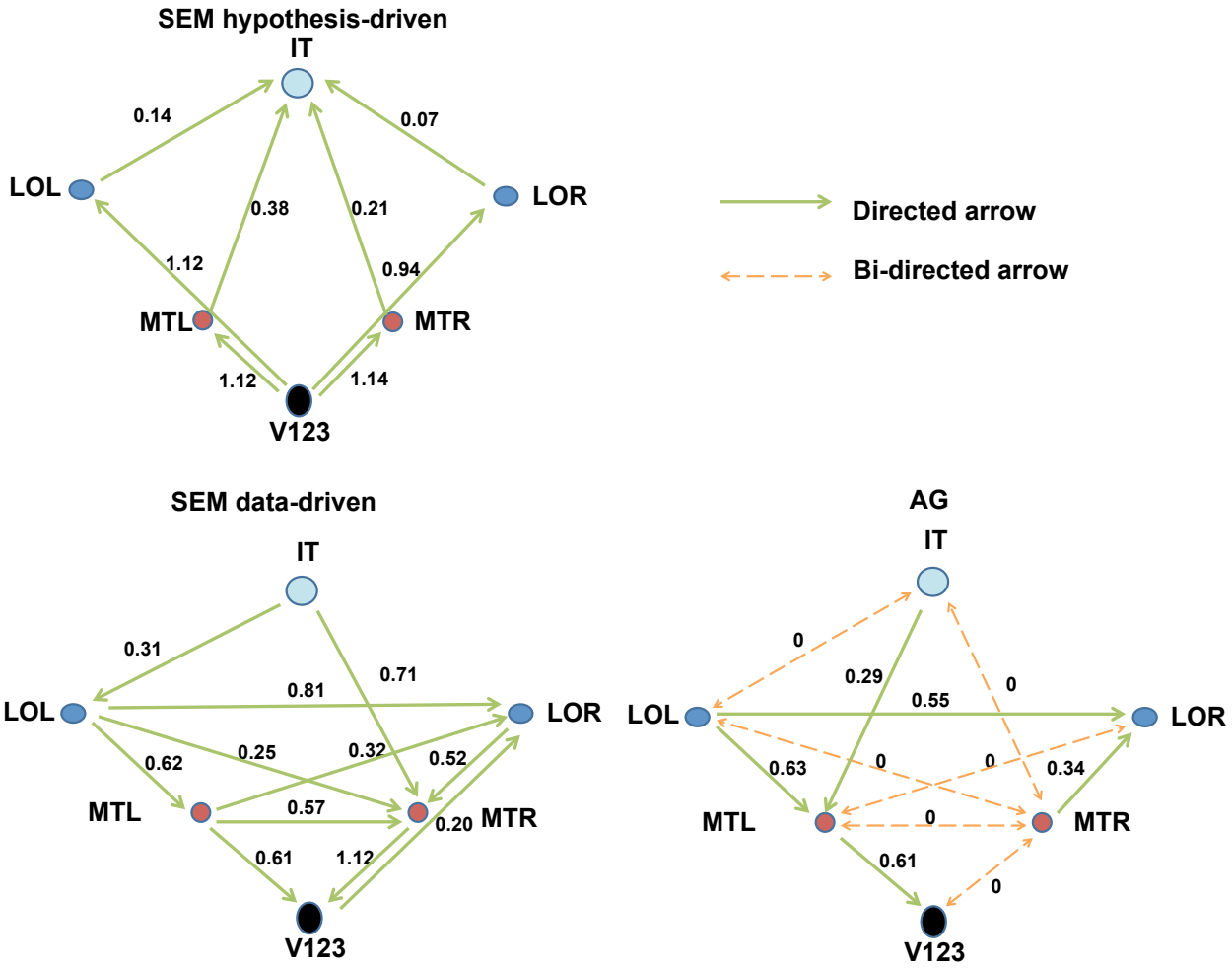


Figure 6.8: The best-fitting models for the first subject in the *Homogenous* condition. Orange dotted connections are bidirected connections and indicate that there is a missing region. Green solid connections are directed connections. Besides the connections themselves, the unstandardized strengths of the connections are also presented in the figure.

results for the individual and condition-specific tests are in Table 6.1.

Most important are the differences between SEM and AG for subject 1 in the homogeneous and stack condition. For this subject in the AG model, six bidirected edges were obtained, indicating no direct connection. This corresponded well with the DTI values. SEM and SEMS, on the other hand, had nonzero coefficients for these connections, making its relation with DTI poor (see also Figure 6.8). In the stack condition, the AG model obtained five undirected connections, indicating mutual influence (or at least no direction could be estimated reliably). This also corresponded well with DTI, but not with SEM or SEMS. These results are in line with the results from the correlations.

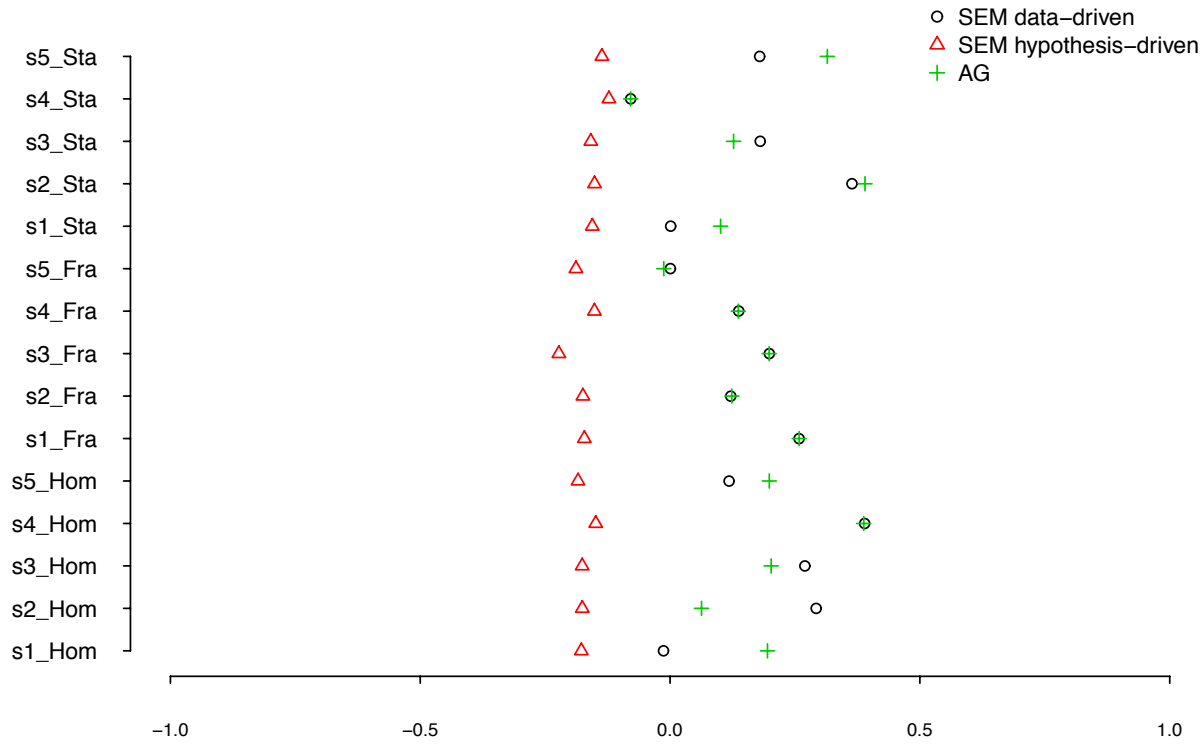


Figure 6.9: Correlations of AG, SEM, and SEMS methods with DTI values. s, subject; Hom, homogenous condition; Fra, frame condition; Sta, stack condition; SEMS, SEM standard (hypothesis–driven).

6.3 Discussion

In this study, we examined AGs for studying effective connectivity. AG was compared with the conventional SEM. We compared the more explorative or data–driven AG method with both the standard hypothesis–driven and a data–driven SEM method. For the data–driven SEM method, we used the models found with AG, as no data–driven method is currently at hand for SEM. We used the data–driven AG and SEM and the hypothesis–driven SEM methods to estimate the connection strength between six ROIs of the visual cortex based on fMRI data of a motion perception task. The achieved effective connection strengths between the ROIs of all methods were correlated with connection probabilities derived from the DTI analysis to compare the performance between AG and SEM methods.

Results indicated that the least accurate models were the models of the hypothesis–driven SEM method. The hypothesis– driven SEM method performed worse than both the data–driven SEM and

Table 6.1: Significance tests of differences between methods of analysis for 30 connections, where AG is ancestral graph, SEM is structural equation modeling (data–driven through AG), and SEMS is SEM standard (hypothsis–driven). The degrees of freedom (df) were computed for each method separately using the Satterthwaite approximation. Starred p-values are significant at level $0.05/15 = 0.0033$

	AG-SEM		*AG-SEMS*		*SEM-SEMS*	
	$df = 10.71$		$df = 16.60$		$df = 17.09$	
$Subj - cond$	χ^2	p	χ^2	p	χ^2	p
S1-hom	35.04	0.0002*	53.40	0.0000*	57.69	0.0000*
S1-frame	0.00	1.0000	26.23	0.0620	26.24	0.0721
S1-stack	35.87	0.0001*	36.84	0.0029*	37.47	0.0030*
S2-hom	13.58	0.2378	40.57	0.0009*	41.55	0.0008*
S2-frame	0.00	1.0000	28.60	0.0334	28.66	0.0389
S2-stack	6.01	0.8573	30.01	0.0226	36.44	0.0042
S3-hom	13.61	0.2361	17.25	0.4099	21.66	0.2022
S3-frame	0.00	1.0000	17.00	0.4266	17.00	0.4602
S3-stack	5.52	0.8910	28.99	0.0300	28.05	0.0457
S4-hom	0.00	1.0000	29.06	0.0295	29.06	0.0350
S4-frame	0.01	1.0000	21.90	0.1709	21.73	0.1995
S4-stack	0.00	1.0000	31.51	0.0148	31.56	0.0176
S5-hom	0.84	1.0000	23.74	0.1134	23.81	0.1276
S5-frame	9.84	0.5187	26.42	0.0590	23.08	0.1500
S5-stack	8.03	0.6872	23.30	0.1255	24.02	0.1217

the AG method. This is probably due to the complicated structure of projections between different visual areas (Felleman & van Essen, 1991), going not only from lower-tier to higher-tier areas but also vice versa (Lamme & Roelfsema, 2000). At this moment, too little is known about this exact information flow between the different regions, which makes it logical that an explorative model has a better fit than a theory-based model.

A comparison of the data–driven SEM and AG methods showed that, in general, the AG and the SEM method predicted structural connectivity equally well. We performed the correlational analyses for each subject and condition separately. Only in the Homogenous and Stack condition of the first subject, the AG method predicted structural connectivity significantly better than the SEM method. In the Homogenous condition and the Stack condition of this subject, the AG models contained a lot of bidirected or undirected connections, respectively. This seems to indicate that whenever there are possibly missing regions, as indicated by the bidirected connections, the AG method outperforms SEM.

Furthermore, it is beneficial for SEM to use models that are based on a selection from AG. Using AG to compose models for SEM has led to models that usually would not be found with SEM. Thus, AG can lead to more informative and accurate models of brain networks in future connectivity research.

6.4 Conclusion

This study showed that AG is a fruitful method to study effective connectivity. In contrast to conventional methods to study effective connectivity, such as SEM, AG can detect, besides directed connections, whether there are undirected connections, indicating mutual influence (or at least no direction could be estimated reliably) and bidirected connections, indicating that there is a missing region causing a correlation but no direct connection between the two ROIs. In particular, the ability to detect missing regions is a unique feature of AG that leads to network models with fewer spurious connections.

7 Heating up the measurement debate: What psychologists can learn from the history of physics

Current theories of psychological measurement are largely disconnected from discussions of measurement in the natural sciences. There are extensive debates on measurement in both domains, but attempts to bring them together are rare.[1] In most textbooks, manuals, or monographs on psychological measurement (e.g., AERA, APA, & NCME, 2014; Borsboom, 2005; Kline, 2000; McDonald, 1999), physical measurement appears (if at all) only in the form of simplified standard examples, such as weight and length, which are not analyzed in any serious detail, and are often used just as contrast cases to psychological measurement.

Indeed, there are substantial differences between psychological and physical measurement: human participants are capable of learning and thus may react very differently at different time points, psychological measurements usually are sum scores made up of item responses, the results may have social and ethical implications, there is a difference between intra- and inter-individual measurements, and so on (Gigerenzer, 1987; McDonald, 1999; Messick, 1989). However, our approach in this paper is to focus on the similarities, and to look for aspects and episodes in the history of physical measurement that are relevant for psychological measurement, more specifically for the debate on validity.

Validity is arguably the most fundamental and controversial issue in psychological measurement (Lissitz, 2009). In the latest edition of the Standards for Educational and Psychological Testing, validity is the first topic discussed, and is characterized as the most fundamental consideration in developing tests and evaluating tests (AERA et al., 2014, p. 11). According to the classic definition, validity refers to the extent to which the test or instrument measures what it is intended to measure (Kline, 2000, p. 17; McDonald, 1999, p. 197), but in contemporary validity literature, accounts of validity differ greatly, and there is no agreement even on how validity should be defined (see Newton and Shaw (2013) for a state-of-the-art overview).

Some of the most prominent approaches to validity are Messick's (1989) unified treatment of validity (and its various refinements), where the focus is on the adequacy of inferences that psychologists make based on test scores; Kane's (2001, 2006, 2013) argument- based approach, where validation

[1] Some notable exceptions are Humphry (2011, 2013), Michell (1999), Rasch (1980).

consists in giving evidence-backed arguments for interpretations of test scores; and the causal approach, where the idea is that in a valid measurement the thing to be measured should cause the measurement outcome (Borsboom, 2005; Borsboom, Mellenbergh, & van Heerden, 2004; Markus & Borsboom, 2013).[2] In this article, we take as the starting point the core idea that validity concerns the extent to which the instrument or test measures what it is intended to measure, but our conclusions have relevance independently of how validity is defined.[3]

Natural scientists do not use the same terminology as psychologists or psychometricians, and do not talk of the validity of measurements, but this does not mean that issues related to validity do not arise in physics – as we will show in this paper, in the history of physics it has often been unclear whether the instruments are measuring what they are intended to measure. In this article, we will focus on a real and detailed example (temperature) from the history of science, and establish connections and parallels between physical and psychological measurement.[4] More specifically, our novel contribution is to show that looking at the history of measurement in physics can lead to new insights and viewpoints for the validity debate in psychology.

We have chosen temperature as our case study because temperature measurement has a long and rich history that is well documented, and has been analyzed in detail by historians and philosophers of science (e.g., Chang, 2004; Sherry, 2011). Furthermore, temperature is a representative example of a physical attribute that can be measured in various ways, and that is easily understandable without any background in physics. As we will show, there are surprising parallels between temperature measurement in the first half of the 19th century and the current situation in psychological research practice. In that period, the focus was on making temperature measurements increasingly precise, consistent, and mutually compatible, without engaging in theoretical work on the nature of temperature. In a similar way, current practice in psychological measurement focuses mainly on criteria such as reliability, generalizability, and correlation with other measures, and far less on theories concerning the psychological attributes measured, or the causal processes underlying the measurements (Borsboom, 2005; Hubley, Zhu, Sasaki, & Gadermann, 2013; Markus & Borsboom, 2013).[5]

As we will show, in temperature measurement this atheoretical approach was insufficient, and substantial scientific progress was made only when the measurements were linked to theory. At a very general level, this supports the views in psychology that emphasize the crucial importance of

[2]Interestingly, Hood (2009) argues that the causal approach is in fact compatible with Messick's (1989) approach to validity, and Newton and Shaw (2013) argue that Kane's argument-based account is compatible with the causal account. Thus, it may be that the different approaches are compatible and focus on different aspects of validity.

[3]If validity is defined in a different sense, so that our arguments do not, strictly speaking, concern validity, they are still relevant for psychological measurement in general.

[4]In this article, we understand psychological measurement to cover all kinds of measurements done in the various fields of psychology, ranging from intelligence tests to measurements of the capacity of short–term memory.

[5]It should be noted that this concerns research and measurement *practice* in psychology. Theoretically oriented psychologists have discussed the importance of theory throughout the 20th century and up to this day.

theory for measurement and validity. However, our main point is to analyze three more concrete conclusions that can be drawn from the physical case, and that are relevant for the validity debate in psychology. First of all, studying the causal mechanisms underlying the measurements can be crucial for evaluating whether the measurements are valid. Secondly, psychologists would benefit from focusing more on the robustness of measurements. Robustness refers here to the idea that if there are several independent ways of measuring something, this increases our confidence in the measurements.[6] Finally, we argue that it is possible to make good science based on (relatively) bad measurements, and that the explanatory success of science can contribute to justifying the validity of measurements.

As an important terminological remark, the expressions "mechanism" and "causal" in this paper should be understood very broadly. By "mechanism", we mean roughly a set of components that are organized together to perform a function (Bechtel, 2008). This covers not only physical mechanisms, but also cognitive and biological mechanisms that need not be deterministic. Similarly, "causal" and "causation" should be understood here in terms of difference-making and potential manipulation and control (Woodward, 2003), and not in terms of exclusively deterministic or physical causation. These broad conceptions of causation and mechanism are compatible with the possibility that the mind or the brain is fundamentally probabilistic (cf. Gigerenzer, 1987).

The structure of this article is as follows. In the next section, we will briefly sketch the relevant cases from the history of temperature measurement. In the following section, we will relate this to the current situation in psychological measurement, and discuss in detail three insights from the history of temperature measurement that are relevant for the validity debate in psychology. In the final section, we will discuss open issues and briefly return to the general topic of measurement in psychology and physics.

7.1 A brief history of temperature measurement

In this section, we will briefly go through some key episodes in the history of temperature measurement. The main focus will be on the atheoretical approach to measurement that reached its high point in the work of Henri Victor Regnault. As we will argue, this approach resulted in very precise and comparable temperature measurements, but fell short for various reasons. Most importantly, it did not result in an increased understanding of what temperature is, and it did not help in assessing what happens in new circumstances when the validity of measurements is unclear. Furthermore, we point

[6]The term "robustness" is ambiguous, and can refer to different things in different contexts. For example, Markus and Borsboom (2013) use it to characterize the stability of causal relationships, and in statistics it refers to measures that are resistant to deviations and errors. We use the term in order to make a connection to the long tradition in philosophy of science (going back at least to Wimsatt, 1981), where robustness has been discussed in the same sense as in this paper.

out that the high degree of precision, consistency, and comparability that Regnault was aiming at is not even necessary for making valid measurements, and that theoretical progress can provide indirect evidence for the validity of measurements.

Let us start with the early days of temperature measurement (see Barnett, 1956 for historical details; the following overview is mostly based on Chang, 2004, pp. 39–56). In the 16th century, researchers such as Galileo started making attempts to develop instruments (thermoscopes) to be able to measure phenomena of heat and cold. Based on subjective sensations of warm and cold, it was discovered early on that liquids (and air) tend to expand when they are heated. Thus, a liquid in a closed glass tube (or any closed vessel) will expand as it gets warmer, and contract as it gets colder. This principle was the starting point for measuring heat and cold. What this means is that the measurement of temperature was not originally based on physical theory, but started from subjective experiences and a simple empirical regularity.

As Chang (2004, pp. 51–52, 159) has pointed out, the improvement of the precision and consistency of temperature measurements proceeded iteratively without much influence from theoretical developments. The simple thermoscopes described in the previous paragraph made it possible to find phenomena that are relatively constant in temperature (such as boiling), and these could be used as fixed points for measurements. Based on this, it was possible to divide the interval between two fixed points into units, resulting in a numerical temperature scale, which allowed for more precise measurements. These numerical thermometers could then be improved in terms of various empirical criteria: they could be made more precise in the sense that they produce more fine-grained readings, more consistent (or reliable) in the sense that they produce the same result in the same circumstances, more comparable in the sense that any two particular thermometers of the same type function in the same way, and more robust in the sense that different types of thermometers give the same results. In this way, measurements could be improved to a high degree, independently of theoretical developments (see also Choppin, 1985).

The culmination of this approach of improving thermometers based on empirical criteria was the work of the French scientist Henri Victor Regnault (1810-1878). Regnault shunned all theoretical speculation about the nature of temperature and emphasized the importance of rigorous testing with a minimal amount of assumptions (Barnett, 1956, pp. 333–341; Chang, 2004, pp. 74–84). Thus, Regnault's approach was anti-theoretical to the extreme. He collected a vast amount of data based on meticulously precise measurements, and used different constructions of instruments to make sure the results were robust (Chang, 2004, p. 175). In the end, Regnault successfully constructed highly precise and comparable gas thermometers, the measurements of which differed from each other only by less than 0.1% (i.e., if one thermometer recorded a temperature of 70°C, the measurements of the same conditions by the other thermometers fell within the range 69.93–70.07°C; Chang, 2004, p. 81).

However, even though Regnault was able to make such extremely precise and consistent measurements, his approach had some fundamental shortcomings. First of all, although an atheoretical approach can guarantee that measurements are consistent and comparable in controlled conditions, such an approach falls short when the conditions are new or unknown. For example, in the 18th century, the best thermometers available were mercury thermometers (invented by Fahrenheit). They had been extensively tested and used only in conditions that naturally occur or that are easy to produce in a laboratory, but it was unclear whether they would continue to provide valid measurements in other circumstances, such as extreme heat or cold. In fact, they did not, as is nicely illustrated by the following story (described in Chang, 2004, pp. 105–118).[7]

In 1733, the Russian scientist Johann Georg Gmelin set out to explore the eastern stretches of Siberia, and on his journey experienced freezing conditions of unexpected harshness. The mercury thermometer that Gmelin was using indicated a temperature of -120°F (-84.4°C). Gmelin was happy to accept this reading as roughly accurate, as indeed it had been very cold, but others were skeptical. Nothing even close to that temperature had ever been recorded on earth. It seemed more likely that the thermometer was no longer providing valid measurements of temperature. This initiated a heated scientific debate, and a new research project: although many of the properties of mercury were well understood, its freezing point was unknown, which various scientists now set out to discover. Only decades later it was established that mercury freezes around -40°C. Like most substances, mercury becomes much denser when frozen, resulting in lower levels of mercury in the thermometer. Thus, the mercury of Gmelins thermometer had frozen, and the temperature had been far less severe than -84.4°C.

The crucial point here is that, in contrast to what Gmelin thought, the comparability and consistency that had been established for mercury thermometers provided no justification for believing in the results, because the conditions were novel and untested. In a similar way, the precision, consistency, and comparability of Regnault's thermometers was only established for limited conditions and a limited part of the temperature scale. Furthermore, it was not possible to resolve the issue of whether the measurements continued to be reliable based on only empirical criteria. Any other mercury thermometer would have also frozen in conditions of extreme cold. In order to solve the problem, scientists had to study how the measurement instrument actually works, that is, the causal mechanism that results in the measurement outcome, and this in turn required theoretical advances (i.e., discovering that a substance like mercury can freeze, and that it will contract when frozen).

A second limitation of the atheoretical approach of Regnault was that this approach did not result in increased understanding of what temperature is, or in new connections with other areas of physics.

[7]This episode took place before Regnaults time, but we describe it here because it gives a vivid illustration of the limits of focusing just on precision and reliability, and thus also the limits of Regnaults approach.

One concrete implication of this was that the temperature scale itself remained just as arbitrary as it had been before Regnaults efforts: The fixed points of the scale(s) were conventions based on practical considerations, and there was no plausible theoretical definition for what it means for temperature to change by one degree. Only after a connection was made to theory did it become possible to formulate an objective definition for what a change of one degree of temperature means, and to calculate the absolute zero (Chang, 2004, pp. 159-197). This was achieved by Thompson (also known as Lord Kelvin, 1824-1907), who connected temperature to the thermodynamic notions of work and heat. This also allowed making temperature measurements more robust: while Regnaults measurements were all in the end based on simple gas laws, the connection to thermodynamics made it possible to derive temperature values also from thermodynamic equations. Eventually, other theoretical advances also resulted in new kinds of instruments for measuring temperature, such as resistance thermometers, which are based on the principle that the electrical resistance of some materials increases with rising temperature.

As a third point, in order to achieve valid measurements and scientific progress, the kind of high degree of precision, consistency, and comparability that the atheoretical approach aims at is not necessary – to put it simply, it is possible to do good science based on relatively bad measurements. To illustrate this, we can again move back in the history of temperature measurement, and consider an episode that took place before Regnault's time: Joseph Black's (1728-1799) discovery of the theory of latent heat (the heat that a substance can absorb or release without changing in temperature; Sherry, 2011). Black's theory amounted to a great scientific advance that marked the beginning of the science of thermodynamics. Interestingly, the measurement instruments at Black's disposal were mercury thermometers that by Regnault's standards (or contemporary standards) would not have counted as very consistent, precise, or mutually compatible. However, this did not stop Black from theorizing about heat and temperature and testing his hypotheses with temperature measurements. Even with these imperfect measurements, Black was able to quantify the notion of latent heat, and with his theory of latent heat he could provide novel explanations to a broad range of phenomena, including the melting of ice and freezing of water.

This case also illustrates another related point: theoretical progress can contribute to the validity of measurements retroactively, or in hindsight. The theory of latent heat was built on the assumption that mercury thermometers provide valid (although imprecise) measurements of temperature. The predictions and explanations based on the theory of latent heat were extremely successful. Thus, Black and his contemporaries had good reasons to believe that the original hypothesis concerning the validity of the measurements was correct (Sherry, 2011). Of course, it was in principle possible that the theory happened to be correct in spite of the temperature measurements being completely invalid, but this would have been almost miraculous: the far more likely explanation was that the

measurements were in fact valid, at least in the sense that they were roughly measuring some real quantity (Sherry, 2011). Thus, theoretical progress and success can contribute to indirectly justifying the validity of measurements.

7.2 Lessons for measurement in psychology

The most general moral that we can draw from the above is that theory is crucially important for measurement and validity. While an atheoretical approach, where the aim is to make measurements better on purely empirical standards (such as reliability and invariance), will result in measurements that are consistent and comparable under a limited range of conditions (corresponding to Regnaults achievements), it will not guarantee the validity of measurements or lead to significant scientific progress.

In many ways, the situation in psychological research practice resembles the situation of temperature measurement in the late 18th and early 19th centuries: the focus is on criteria such as reliability and invariance, and on correlational and purely empirical studies, at the expense of theory-building or theoretical speculation. The standard approach in psychometric modeling is to find statistical models that fit the data, which can be done independently of theoretical assumptions concerning the thing that is measured (Markus & Borsboom, 2013, p. 43). Assessments of validity in practice most often amount to evaluating the internal structure of the test, or correlating the results with external variables and seeing whether the correlations are in the right direction (Borsboom, Cramer, Kievit, Scholten, & Franić, 2009; Borsboom et al., 2004; Hubley et al., 2013). Furthermore, just like the temperature scales in Regnaults time, psychological scales and units lack any clear theoretical foundation, and there is no clear understanding of the nature of the attributes measured (Humphry, 2011).

The temperature story illustrates the limits of such an atheoretical approach: it will result in measurements that are consistent and precise under a limited range of conditions (corresponding to Regnaults achievements), but it will not guarantee the validity of measurements or lead to significant scientific progress. For advances on these fronts, theory is required.

This point as such is not novel – the importance of theory has been widely discussed in the debates on validity in psychology, starting from the classic paper by Cronbach and Meehl (1955) and going on up to this day (e.g., Borsboom, 2005; Embretson, 1983; Embretson & Gorin, 2001; Kane, 2013; Markus & Borsboom, 2013; Messick, 1989; Newton & Shaw, 2013). Here we have provided new evidence for the perils of neglecting theory. Furthermore, in the validity literature, the views on the role and importance of theory for establishing or assessing validity vary greatly, and the considerations in the previous section clearly support accounts that place theory at the very core of assessing validity (e.g., Borsboom et al., 2004; Embretson, 1983, 1998; Embretson & Gorin, 2001).

However, our main point in this section is to show how the three more specific issues we picked up from the history of temperature measurement are relevant for the validity debate. Our first main point in the previous section was that understanding the causal mechanism underlying the measurement instrument is essential for assessing validity. This was evident in the story of Gmelin's frozen thermometer: in order to determine whether Gmelin's measurements were valid measurements of temperature, scientists had to study what exactly happens in the mechanism of the thermometer in extreme temperatures. Thus, theoretical understanding of the causal mechanism of measurement seems to be crucial for assessing the validity of measurements, especially in novel circumstances. More generally, in the philosophical literature on measurement, the focus nowadays is on understanding and modeling the measurement process and the (causal) functioning of the measurement instrument (Chang, 2004; Frigerio, Giordani, & Mari, 2010; Kyburg, 1992; Tal, 2013; Trout, 1998).

This suggests that understanding the causal mechanisms underlying measurements should be crucial also for assessing psychological validity, in line with the causal account of validity defended by Borsboom and colleagues (Borsboom, 2005; Borsboom, Cramer, et al., 2009; Borsboom et al., 2004; Markus & Borsboom, 2013). However, we do not agree with these authors that validity only concerns the question of whether the attribute to be measured actually exists and causes the variations in the measurement outcomes (e.g., Borsboom et al., 2004, p. 106). As will become clear below, we believe that there are also many other important aspects to validity.

One obvious problem that arises when the approach of studying the causal mechanisms of measurement is applied to psychology is whether we can actually study the relevant mechanisms. The most prominent attempt at this is found in Susan Embretson's groundbreaking work. According to Embretson's (1983, 1998, 2004) process-oriented account of validity, traditional assessments of construct validity (i.e., comparing the test scores to relevant external variables) need to be supplemented with studies of the cognitive processes and strategies that participants use to respond to test items, based on state-of-the-art cognitive psychology. When this approach is applied in practice, cognitive theory influences both test construction and the measurement models (such as item response theory, IRT, models): the items selected for the test are based on cognitive theory, and the models include parameters representing the cognitive demands of the item (Embretson, 1983, 2004; Tatsuoka, 1987, 1990).[8] For example, a test for assessing abstract reasoning was created based on processing theory, and the IRT model was combined with a cognitive model, including parameters such as working memory load and perceptual processing (Embretson, 1998).

However, as important as these studies are, including cognitive parameters in measurement models is still very far from describing the causal mechanisms underlying the measurement process. Ideally,

[8] A similar approach to incorporating cognitive theory into test development is Mislevy's Evidence Centered Design (ECD; Mislevy, Steinberg, & Almond, 2002).

there should be models that describe the steps in the causal process that start with the attribute intended to measure and end with the measurement outcome. It may be that the reason why this is not generally attempted, or not even seen as a goal, is that the causal mechanisms in psychology are thought to be so complex that figuring out the causally relevant components is practically impossible (Trendler, 2009). We acknowledge that the challenges may seem daunting at present, but do not believe that the situation is hopeless with regard to the future: great progress has been made in recent decades in discovering mental mechanisms (Bechtel, 2008), an issue to which we return below.

The second important insight for validity (and psychological measurement in general) that we draw from the physical case is the principle of robustness. This is a method or principle that pervades all of science and has also many other names: mutual grounding, mutual compatibility, overdetermination, triangulation, diverse testing, and so on (Chang, 1995, 2004; Eronen, 2012; Hacking, 1983; Tal, 2011; Trout, 1998; Wimsatt, 1981, 2007). The basic idea is that if there are several independent ways of achieving the same result, this increases our confidence in the result. This can also be expressed as the following mathematical principle: if there are several (independent) ways of measuring something, the probability that all of them happen to go wrong is a product of the individual probabilities of going wrong, and this product becomes increasingly tiny as more and more independent ways are added (Wimsatt, 1981).

The idea of independence is crucial for robustness. There is no uncontroversial or widely accepted account of the exact nature of the required independence, but certain key features can be spelled out (Nederbragt, 2012; Stegenga, 2012; Stroebe & Strack, 2014; Wimsatt, 2007). First of all, it is obvious that statistical independence is not what is required: different ways of measuring temperature will be statistically correlated, even when they are in other important respects independent (e.g., two thermometers based on different physical principles, such as a mercury thermometer and a radiation thermometer). The idea is rather that the different ways of measuring should partly rely on different theoretical assumptions, different physical processes, or different experimental setups. What is necessary is that any problematic or unconfirmed assumptions should not be shared by the different ways (Stegenga, 2012). Different approaches or ways of measuring are fully independent only if they rely on different assumptions and different parts of theory (such as mercury thermometers and radiation thermometers). It is clear that independence is a matter of degree, and not a none-or-all property: two different mercury thermometers are less independent from each other than a mercury thermometer and a radiation thermometer.

Robustness itself is also a matter of degree, corresponding to the number and independence of the different ways of measuring. Once a high degree of robustness is reached, we can be confident in the measurements, and conversely, if measurements are not robust or robust to a low degree, we should approach them with healthy skepticism. This principle can be applied to measurements, attributes,

properties, and entities, but it is important to keep in mind that it is a fallible epistemic principle, and not a guarantee of truth or reality (see Eronen, 2012 and Wimsatt, 2007 for more).

An increase in the degree of robustness is evident in the history of temperature measurement, as outlined in the previous section. If several different thermometers give the same reading, it is more likely that the reading is correct than when only one thermometer is used. However, applying multiple instruments based on the same theoretical principle leads to a low degree of robustness, because their independence from each other is very limited. If the theory on which the instruments are based turns out to be false, the fact that multiple instruments were used becomes irrelevant.[9] Also, the robustness of Regnaults measurements was limited, because all the thermometers he used were implicitly based on the same gas laws, and no further connection to theory was made. A higher degree of robustness was only reached through theoretical developments, which led eventually to new types of instruments for measuring temperature, relying on different areas of physics, such as radiation thermometers and resistance thermometers.

The idea of robustness (although different terms are used) also has a tradition in psychometrics, going all the way back to the classic article by Cronbach and Meehl (1955) and the multitrait-multimethod matrix approach of D. T. Campbell and Fiske (1959). In contemporary validity theory, the idea comes up in the context of convergent validity, which refers to evidence from other measures that are intended to assess the same or a similar attribute (AERA et al., 2014, pp. 16-17). However, in practice, assessing convergent validity usually amounts to calculating correlation coefficients between measures that are expected to be related, possibly supplemented with factor analysis or principal component analysis (see, e.g., Clapham, 2004; Duckworth & Kern, 2011). No special attention is paid to the independence of measures, or to deriving the result from theory. Thus, convergent validity can be seen as a weak form of robustness. Furthermore, in psychometrics convergent validity is usually briefly mentioned as one possible source of evidence for validity, while in the natural sciences robustness is central to the validity of measurements (Chang, 1995, 2004; Wimsatt, 2007).

It is clear that the degree of robustness of the measures and constructs in contemporary psychology varies greatly. For example, it could be argued that intelligence measurements (IQ scores) are robust to a low degree, because although the results of different intelligence tests are highly correlated, intelligence scores are not based on any widely accepted theory, and the independence of the various tests can be questioned (van der Maas et al., 2006). An example of a domain where psycho-

[9] A classic example of this is the bacterial mesosome (Culp, 1994; Wimsatt, 2007, p. 381, note 3). This entity appeared in various experiments studying bacteria and was initially thought to be a new kind of cellular organelle. Because independent research groups using different experimental setups could detect the bacterial mesosomes, the results that indicated their existence seemed to be robust. However, it later turned out that all of the different experimental setups were using the same fixation methods for preparing the samples, and the bacterial mesosomes were merely artifacts of the preparation methods. Thus, the various experimental setups were – in a crucially important respect – not independent from each other, and the robustness of the results was illusory.

logical measurements have a higher degree of robustness would be short-term memory. The capacity of short-term memory can be measured in a broad range of different types of (independent) experimental setups: imposing an information overload, preventing long-term memory access, examining performance discontinuities in memory tasks, mathematically modeling memory performance, and so on (Cowan, 2001; Jonides et al., 2008). In any case, we believe that the debate on validity would benefit from a closer analysis of robustness, and more specifically the independence of measurements.

The third main point for validity that we draw from the history of physics can be summarized as follows: relatively bad measurements can result in good science, and scientific progress can justify the validity of measurements in hindsight. This was discussed in the last part of the previous section. Instead of waiting for better measurement techniques, theoretically oriented scientists such as Black made the working hypothesis that the imperfect measurement instruments (mercury thermometers) at their disposal were consistent and valid enough, and formulated theories and explanations based on that working hypothesis (see also Choppin, 1985). Those theories turned out to be very successful. This success gave the scientists more reason and justification for believing that the working hypothesis was true, and that the measurements were valid in the sense that a real and scientifically important quantity was being measured.[10] Note that this does not involve any vicious circularity in the sense that measurements are validated by theory, and the theory is validated by measurements. The pattern is rather this: what increases confidence in the validity of measurements is the success of the theories that are based on them, and what justifies the success of those theories is their explanatory and predictive power. Testing the latter need not involve the same types of measurements whose validity is in question.

We certainly do not want to claim that this is the only way of establishing the validity of measurements. The point is rather that this is one way of contributing to validity arguments or justifying validity claims. To the best of our knowledge, this has not been explicitly discussed in current validity literature (although interestingly Coleman, 1964, pp. 70-73 makes a similar point in the context of measurement in sociology). Consequences are generally regarded as one source of evidence for validity, but this refers to the consequences of test use in practice, and not consequences for theory and science in general (see, e.g., Messick, 1989; Newton & Shaw, 2013).

This point also has implications for Joel Michell's arguments against psychological measurement. Michell (1997, 1999, 2000, 2013) has argued that psychologists are treating the attributes they are measuring as quantitative, without having even attempted to show that they fulfil the requirements

[10]There are also numerous cases in the history of science where the working hypothesis was not confirmed, and the measurements turned out to have been invalid. Consider phlogiston: scientists in the 17th and the 18th centuries assumed that they were measuring quantities of phlogiston in combustible things. However, explanations and predictions based on the phlogiston theory were fundamentally problematic, and eventually the theory was replaced by oxygen-based theories of combustion. Thus, the validity of phlogiston measurements was disconfirmed by later developments in science.

for being quantitative in structure (such as additivity, i.e., that there is a meaningful way of adding up quantities of the attribute). Thus, psychological measurement currently lacks any foundation, and as long as it has not been shown that psychological attributes are in fact quantitative, psychometrics is a pathological science (see also Trendler, 2009).

However, as Sherry (2011, pp. 515–517) has pointed out, the history of thermometry suggests that this shortcoming may be less devastating (or "pathological") for psychology than Michell thinks. Black and his contemporaries did not have any conclusive arguments for the quantitative nature of temperature either, and they had no conception of the actual nature of heat and temperature (which were only discovered in the late 19th century; see also Choppin, 1985). Instead, they made the working hypothesis that temperature is measurable and that mercury thermometers provide roughly valid measurements of it, and built their theories based on this assumption. Considering the success of those theories, in retrospect it is clear that making that hypothesis was a crucially important and justified move. Thus, it is plausible that psychologists can also make a similar working hypothesis, which will then be confirmed or disconfirmed by later developments in science (see also Humphry, 2011).

Michell could respond that while this strategy works in physics, there is no reason to expect it to work in psychology. In Black's case, the theories based on temperature measurements were very successful, and were soon broadly accepted, but psychology has so far failed to produce theories of significant scope or explanatory power, and current psychological theories are not rich and detailed enough to provide serious tests for the hypothesis that psychological measurements are valid (see, e.g., Michell, 2004). Thus, perhaps we cannot expect in psychology the kind of progress that led to the vindication of temperature measurements.

In our view, this is a question that can only be resolved by the eventual development and progress of psychology as a science, and in this regard we are far less skeptical than Michell. At the moment, no overarching theories of the kind developed by Black, Thomson, or Maxwell in the case of temperature are foreseeable in psychology, but this should not be seen as discouraging: theories in psychology and the life sciences in general tend to be more local than in physics (Bechtel, 2008; Bechtel & Richardson, 1993; Kyngdon, 2013). For example, there is no (and likely never will be) single overarching theory of biology, but a broad range of theories concerning natural selection, gene expression, development, ecology, and so on. In a similar way, instead of one unified theory of human psychology, there will be increasingly precise theories or models of perception, language learning, problem solving, and so on.

In fact, it may be that theoretical development in psychology is hampered by the implicit assumption that theories are simply better the more general they are. For example, based on the mirror neuron mechanism, psychologists have proposed far-reaching theories of social cognition (Gallese, Rochat, Cossu, & Sinigaglia, 2009). It may be better to focus first on local mechanisms and their

specific and limited roles and functions, for example the role of mirror neurons in perception of goal-directed behavior (Spaulding, 2013), and to make these local theories and explanations as elaborate and plausible as possible.

Indeed, in many areas of psychological measurement, such as memory, we already find relatively well-developed local theories, for example in the case of short-term memory mentioned above. There are theories concerning the interplay of short-term memory, long-term memory, focus of attention, and perception (Cowan, 2001; Jonides et al., 2008). These theories predict and explain what happens, for example, when participants are asked to report the elements in a visual array that is presented in a flash: very roughly, their focus of attention has a limited capacity, and thus participants only report the elements they attended to (Cowan, 2001). Other examples of promising and local psychological theories are provided by Kyngdon (2013).

7.3 Concluding remarks

In this article, we have discussed the history of the measurement of temperature and its methodological relevance for psychology, particularly the debate on validity. We started by going through some important episodes in the history of temperature measurement, and pointed out that there is a parallel between the hyper-empirical approach of Regnault in the 19th century and the atheoretical attitude that is still common in psychological research practice. We also argued that this approach was in the end insufficient in temperature measurement, and that it will likely be insufficient in psychology as well. In short, it does not lead to increased understanding of the phenomena or attributes measured and does not lead to important scientific advances, and the high reliability and consistency that it strives for is not even necessary for valid measurements.

We then looked at more concrete ways in which the validity of psychological measurements could be improved. Interestingly, all of these three ways are closely related to theory. Our first point was that assessing measurements in novel situations requires theoretical understanding of the causal mechanism underlying the measurement process. In the case of robustness, determining the degree of robustness depends on theories and models about the experimental setups, measuring instruments, and the things being measured. Our last point explicitly concerned theories: the validity of measurements can be indirectly justified or established based on the development of successful theories that build on the measurements. Thus, this article can be seen as continuing the long tradition of emphasizing the importance of theory for measurement and validity.

Although we have argued that there are parallels between physical and psychological measurement, we do not want to deny that there are also substantial differences, as we acknowledged in the introduction. However, we believe that the differences are a matter of degree, and not as categorical as is often

supposed. For example, although properties such as length or weight can be measured in a relatively direct and straightforward way, the same does not apply to phenomena such as the weak nuclear force or the background radiation of the universe. Such phenomena (which includes most phenomena studied in contemporary physics) can be measured only indirectly, and have no straightforward operationalizations (Kyburg, 1984). To return to the example of temperature, it is worth mentioning that even nowadays temperature measurement faces considerable practical and conceptual problems and is far from trivial. For example, a monograph on temperature measurement by Quinn (1990) uses hundreds of pages to discuss various issues and complications in measuring temperature.

In our view, the main differences between physical attributes and psychological attributes are that (a) most psychological attributes have not been embedded into any successful and widely accepted theory (see also Sijtsma, 2012), and relatedly, (b) there is no solid theoretical foundation for the units, ratios, and scales for psychological attributes (see also Humphry, 2011, 2013). However, as we have shown, the situation in early temperature measurement was not much better. Thus, we do not think that these differences rule out the possibility of measuring psychological attributes; rather, they emphasize the importance of developing better psychological theories.

In sum, we believe that the existing discussions of this topic have focused too little on the similarities between physical and psychological measurement, and we do not think that the differences are so fundamental that they prevent drawing interesting parallels and looking to physical sciences for insights. More generally, we believe that the methodology and history of physical measurement can be valuable to psychologists, as we hope to have shown in this paper in the context of the validity debate. We do not claim that psychology will necessarily develop in the same way as physics has developed, but rather that psychologists should not think that the history and theory of measurement in physics and other natural sciences is irrelevant to psychology. In conclusion, we hope that this article can contribute to opening new pathways for studying psychological measurement.

8 Discussion

Now that we have discussed different approaches to networks and measurement in psychology, we will take a bird's-eye view on this topic in this final chapter. First, we will consider whether the grounds for using a network approach in psychopathology are really different from the grounds for using networks in other fields. As we have seen, in psychopathology, emotion and personality research the network approach is motivated by the need for an alternative to the latent variable model, but as will become clear, this issue is not as transparent as it may seem at first. Second, we will scrutinize the fundament on which most networks in this thesis are built upon: the VAR model. Finally, this critical examination will lead back to the original title question: Dynamical networks in psychology – more than a pretty picture?

8.1 Networks versus latent variable models

Network analyses are popular throughout all scientific disciplines, including psychology. However, as has become clear in this thesis, only recently the network approach has also found its way to psychopathology, emotion and personality research. What is unique to this recent network approach is the theoretical rationale for the need of networks in this field of research. In Chapters 1 to 4, following the reasoning of Borsboom, Cramer and colleagues, contrasting network models to latent variable model has been the main motivation for the network approach. This motivation is nicely summarized by Cramer and colleagues as follows: "The pivotal problem of comorbidity research lies in the psychometric foundation it rests on, that is, latent variable theory, in which a mental disorder is viewed as a latent variable that causes a constellation of symptoms. ... We argue that such a latent variable perspective encounters serious problems in the study of comorbidity, and offer a radically different conceptualization in terms of a network approach ..." (Cramer et al., 2010, p.137).

This criticism has been taken to heart, and especially in clinical research, an exponential increase in network research is apparent. Networks of depressive disorder (Cramer, Borsboom, et al., 2012; Fried et al., 2014; Fried & Nesse, 2014; Fried et al., 2015; Robinaugh et al., 2014), autism spectrum disorder (Anderson, 2015; Ruzzano et al., 2015), post traumatic stress disorder (McNally, 2012; McNally et al., 2015), personality traits (Cramer, van der Sluis, et al., 2012a) and diagnostic assessment tools

for mental disorders (e.g, BDI, DSM and ICD manual; Boschloo et al., 2015; Bringmann et al., 2015; Borsboom et al., 2011; Tio, Epskamp, Noordhof, & Borsboom, in press) have been introduced, all contrasting this new perspective to the latent variable approach. Additionally, several articles not only conceptually contrast networks with the latent variable model, but also empirically test the necessity for a network approach. For example, Cramer, Borsboom, et al. (2012) and Fried et al. (2014) tested whether risk factors for developing depression such as stressful life events directly influenced the correlation between symptoms, or whether this influence was indirect via a latent variable. They found that the network model, which assumes that risk factors influence the correlation between depressive symptoms directly, had a better fit than a latent variable model, concluding that the former perspective is preferable to the latter.

However, beyond this embracement of the critique of latent variable models, some researchers have cast doubt on the necessity to counterpose the network perspective to the latent variable approach (see for example: Ashton & Lee, 2012; Danks, Fancsali, Glymour, & Scheines, 2010; Haig & Vertue, 2010; Humphry & McGrane, 2010; Krueger, DeYoung, & Markon, 2010; Markus, 2010; McFarland & Malta, 2010; Molenaar, 2010). Comments have been made that the common cause model that Cramer and colleagues criticize is not a model actually used by personality or psychopathology researchers (Ashton & Lee, 2012), that the common cause or latent variable model where symptoms or variables are not allowed to influence each other is a straw man as it is also possible to allow for such influences (Danks et al., 2010), and that network models and latent variable models are mathematically equivalent (Molenaar, 2010). In general, it seems that the distinction between theoretical and psychometric issues has become blurred in this debate. In the next paragraphs, we will try to disentangle the two in order to get a better understanding of the relationship between latent variable and network models.

The common cause approach

The starting point for the debate on networks and latent variables is the question why symptoms (or items) of a certain disorder such as depression (a latent construct) tend to covary more with each other than with symptoms of, for instance, schizophrenia. The dominant answer, according to Borsboom and Cramer (2013), is that symptoms tend to covary because there is a common cause to them, the disorder itself. However, according to Borsboom and Cramer (2013) this common cause perspective to mental disorders is problematic (see also introduction). Cramer and colleagues argue that the problem of the dominant view lies in its psychometric foundation (the latent variable model), but also give a conceptual argument: Psychiatric disorders and their symptoms are related as a whole to its parts, while in a common cause perspective, symptoms are not seen as (part of) the disorder, but as mere indicators of and caused by the disorder (Cramer et al., 2010). Thus, in the common cause

approach disorders and their symptoms are implicitly (and mistakenly) assumed to be conceptually independent or distinct.

This common cause perspective can be, for example, found in research searching for biological and/or genetic mechanisms of mental diseases (Borsboom & Cramer, 2013; Cramer, van der Sluis, et al., 2012a). A famous example is the serotonin hypothesis of depression. According to this theory, depression is caused by neurochemical alterations in the nervous system (Lacasse & Leo, 2005). Following this reasoning, researchers have tried, unsuccessfully, to induce depression by, for example, depleting serotonin levels (Heninger, Delgado, & Charney, 1996). The general lack of success of this approach has been assigned to its simplistic nature: the cause of depression cannot be just a simple neurotransmitter defect, but some more complex biological mechanisms and processes. In studies such as these, a common cause perspective is implicitly taken, as symptoms are assumed to be caused by depression, and no special attention is given to specific symptoms nor their interactions, the goal being to tackle the disorder itself and not merely it symptoms (Patten, 2015).

A common cause perspective seems plausible for diseases such as HIV, where the disorder does cause symptoms, and can even exist without symptoms. Thus, regarding medical diseases like HIV or brain tumor, the disease and its symptoms are conceptually distinct. For psychiatric disorders such as depression, however, the symptoms make up the disorder, as it is implausible that someone could be diagnosed with depression without any symptoms. Thus, depression and its symptoms seem not to be conceptually distinct. For this reason, it is incorrect to say that symptoms are caused by the disorder, just as it is incorrect to state that "being a heavy smoker causes one to smoke 20 cigarettes a day" (J. Campbell, 2010, p. 70). Following this reasoning, it can be argued that in order to make progress in understanding psychiatric disorders such as depression, symptoms and their interaction should be the focus of research, as the disorder *is* (made up of) its symptoms (Cramer, 2012; Cramer, van der Sluis, et al., 2012b; Fried, 2015; Fried & Nesse, 2015).

Don't blame the model

As shown in the previous section, the common cause perspective is a genuine problem in psychological theorizing, as it conceptualizes psychiatric disorders (and perhaps other psychological constructs) in a way that seems implausible. Additionally, we pointed out above that the root of this problem can be traced to the problematic assumption of conceptual distinctiveness, as psychiatric disorders *are* (composed of) their symptoms and thus cannot be the causes of the symptoms. A reasoning more difficult to follow is the psychometric connotation given to this, in essence, conceptual problem. According to Cramer and colleagues, the common cause hypothesis (or disease model) is wedded to one psychometric interpretation, namely the latent variable *model* (see e.g., Cramer et al., 2010; Schmittmann et al.,

2013). Even more so, the common cause hypothesis commits one to a very restrictive latent variable model, a factor model that has the assumption of local independence. Moreover, by just applying a latent variable model with local independence to your data, you implicitly take the common cause perspective (Cramer, van der Sluis, et al., 2012a). In this way, Cramer and colleagues argue that the common cause hypothesis and the latent variable model are two sides of the same coin. Additionally, if one does not assume that a common cause structure gives rise to the covariance between the variables, then, according to Cramer et al. (2010) a network model is the psychometric model of choice.

Although appealing, this reasoning falls short. As Cramer et al. (2010) argue that taking the latent variable model automatically leads to the common cause perspective, they interpret latent variable models as causal models. However, there are other ways of interpreting them. In essence, a latent variable is a variable that is itself unobserved but is inferred from or operationalized by observed variables, for example, through a latent variable model (Jöreskog & Goldberger, 1975). Quite generally, a latent variable model or factor model can be seen as an efficient way of representing or capturing correlations in the data. As some variables tend to covary, the latent variable model can be used to statistically summarize this covariance by the factor(s) of the model (Jonas & Markon, 2016). Thus, the theoretical interpretation of the covariance between the variables and the factor itself is up to the researcher, and not something the statistical model itself can "tell" (Borsboom, Mellenbergh, & Van Heerden, 2003, p. 206). One can apply a latent variable model without necessarily assuming that the factor is causing the observed variables. Therefore, the claim that using the latent variable model automatically leads to the common cause perspective is not warranted, as one can also interpret a latent variable model as a parsimonious representation of the covariance matrix.

Moreover, even if you interpret latent variable models causally, this does not necessarily lead to the common cause perspective as defined by Cramer and colleagues. As pointed out also by Cramer and colleagues, a restrictive latent variable model assuming local independence appears in directed acyclic graphs when there is an unobserved common cause in the graph (Cramer, van der Sluis, et al., 2012a; Pearl, 2000). A standard example is a model with two observed variables: having a cough and having yellow finger nails. These two variables are correlated, but not because they are causally connected. The correlation is due to an unobserved common cause: smoking. In the same vein, consider the ancestral graphs that were discussed in chapter 6. In this framework, a bidirectional arrow is an indication of a missing variable not taken into account in the network, in other words an unobserved (latent) common cause. However, in the context of causal graphs, the latent variable is not given or described beforehand, but is simply some unknown and unobserved variable. In the common cause perspective as characterized by Cramer and colleagues, the latent variable that accounts for the correlations between observed variables (symptoms) is already a priori assumed to be the psychiatric disorder. Thus, the way in which the latent variable model is applied is very different in these two

contexts (see also Bollen, 2002). In the common cause perspective, it is intended to represent a *predefined* cause (the disorder), while in the context of causal modeling, it indicates the presence of an *unknown* cause or variable(s) that is not included in the model. For this reason, applying a causally interpreted latent variable model does not as such lead to the common cause perspective, which is an additional and independent conceptual assumption.

Furthermore, you can also make the problematic assumptions of the common cause perspective when not using a latent variable model, but for example a network model. According to Cramer and colleagues, one important aspect of the common cause perspective is that the focus is on the underlying "disorder itself" instead of symptoms and their interactions (Borsboom & Cramer, 2013, p. 92). However, taking again the example of major depressive disorder, one could also construct the following simple network with three nodes: 1) sum score of a depression scale (e.g., BDI-II) that characterizes the severity of depression for an individual, 2) genes that code for serotonin, and 3) brain areas in which serotonin pathways are thought to be located. Moreover, let us assume that all nodes are connected, and thus some kind of interaction is apparent between all the nodes. Following the serotonin example of the previous section, one could then based on the network edges make a statement such as "depression is more likely to happen when certain serotonin genes are present". According to Cramer and colleagues, when "correlating latent variables (by their sum score proxy) with all sort of (non-)biological phenomena ... [one does] grant the latent variable a status that comes undeniably close to reification" (Cramer, van der Sluis, et al., 2012a, p. 453). Thus, here we have a network model that also includes the "disorder itself" as a real component and implicitly assumes a central aspect of the common cause perspective that Cramer and colleagues criticize.

In general, the argumentation of Cramer and colleagues results in confusion as it does not separate the conceptual (common cause hypothesis as the conceptual distinctiveness hypothesis) from the psychometric (common causes in latent variable, graph and network models). Consider, for example, the following statement: "In a strict psychometric sense, a latent variable model does not allow for many direct relations since the majority of covariance between symptoms needs to be explained by the common cause" (Cramer et al., 2010, p.139). However, the reason why these direct relations between the variables are not allowed seems to be a conceptual rather than a psychometric reason. Indeed, Cramer and colleagues point out that "... technically [it is] not a problem to fit a one-factor model in which certain items are allowed to correlate, in addition to and independent of the relation that they share via the latent factor" (Cramer, van der Sluis, et al., 2012a, p. 452). The problem according to them is that "the more such correlations are allowed to exist in the model, the less convincing is the case for an underlying trait that explains the majority of covariance between the items" (ibid.). In other words, if you are a true believer in the the common cause perspective, you would, according to Cramer and colleagues, expect a strict one-factor model with conditional independence to hold.

However, you are technically, statistically and psychometrically allowed to include direct relationships between symptoms in your latent variable model (Danks et al., 2010). It is just not in line with the conceptual idea that depression is really the cause of its symptoms and that the symptoms are mere indicators of depression.

Conclusion

First of all, Cramer and colleagues have pointed out an important conceptual problem that deserves more research: The common cause hypothesis or rather the conceptual distinctiveness assumption. With this in mind, we should be aware that psychiatric disorders are likely to be conceptually different from medical diseases.[1] Second, it was shown that this kind of problematic thinking about psychiatric disorders (i.e., the conceptual distinctiveness problem) is not a problem of using latent variable models as such. Therefore, latent variable models, even restrained ones assuming local independence, do not have to be shunned. Even more so, latent variable models are mathematically equivalent to network models, in the sense that they have an equal number of free parameters and goodness of fit to the data (Molenaar, van Rijn, & Hamaker, 2007; Molenaar, 2010; Epskamp, Maris, Waldorp, & Borsboom, in press). For example, a one-factor model can be transformed to a mathematically equivalent network where the edges represent regression relationships between the observed variables and the latent variable is transformed out of the network (Molenaar et al., 2007, p. 189).

Based on the considerations above, it is not useful or necessary to pit latent variable models against network models, for example trying to empirically find out whether the disease model or the network model fits the data better. Instead, the focus should be on how the variables you are interested in can be modeled in a sound way. If you are, as is the case here, interested in modeling the interaction between variables or symptoms, you should not shun the use of latent variable models, but incorporate them in a network model. In a symptom network, there is likely to be overlap between variables: for example, one might question how conceptually distinct the BDI items fatigue and lack of energy really are. Furthermore, it is plausible that measurement error occurs when measuring the variables (Markus, 2010; Molenaar, 2010). Using latent variable models can lead to reduction of symptom variables where necessary, while at the same time controlling for measurement error (McFarland & Malta, 2010).[2]

[1] However, even for medical disorders, where it assumed that the common cause perspective is warranted, it is not clear that interactions between the symptoms or even interactions from the symptoms to the disease itself are irrelevant or spurious. For example, even though treating cancer completely would likely lead to the vanishing of the symptoms, the symptom interaction might still have relevance on its own, as it could influence (e.g., worsen) the recovery from cancer. For instance, if the interaction between fatigue and fever (two possible symptoms of cancer) is ignored, and the patient is not getting enough rest, the cancer might worsen as the immune system gets weaker. In this sense, symptom interaction also in medical disorders is not something that can be ignored, which makes the conceptual difference between medical and psychiatric diseases again more blurry and in need of further research.

[2] Such models are similar to dynamic factor models, which have already been successfully used in the literature (e.g.,

Finally, the rationale for using network models in psychopathology, emotion or personality research does not seem that different from other fields of science. We do not need networks because there has to be an alternative to latent variable models, but because we are interested in the interaction between variables, and network models are a useful tool for studying this.

8.2 VAR: Very Awful Regressions?

Although the network approach in general is very appealing and plausible, in practice a network is only as good as the model it is based on. In this thesis, the VAR model is used to infer most networks. Thus, in order to determine whether the networks presented in this thesis are more than just pretty visualizations, we will now take a critical look at the VAR model. VAR models have their origin in econometrics, and have not been uncontroversial there. For example, Harvey (1997, p. 199) claims: "To many econometricians, VAR stands for 'Very Awful Regression'."

Before the use of VAR models took off, structural simultaneous equations models were the standard. In simultaneous equation systems left-hand side variables can also appear as right-hand side variables in other equations of the system (Brandt & Williams, 2007). Thus, there is instantaneous feedback from the output side of the system to the input side. As this system is not identified, researchers must place a "structure" or a priori restrictions on the system inspired by theory, for example, removing instantaneous feedback in the system by specifying upfront that some of the coefficients are zero (see Kennedy, 2003, Chapter 11).

An example of such a model is the structural equation form of VAR (SVAR) (Brandt & Williams, 2007):

$$y_{1,t} = g_{10} - a_{12}y_{2,t} + \gamma_{11}y_{1,t-1} + \gamma_{12}y_{2,t-1} + u_{1,t} \tag{8.1}$$

$$y_{2,t} = g_{20} - a_{21}y_{1,t} + \gamma_{21}y_{1,t-1} + \gamma_{22}y_{2,t-1} + u_{2,t} \tag{8.2}$$

In a SVAR model, the variables $(y_{1,t}, y_{2,t})$ are not only specified by their temporal dynamics (regressed on their lagged values $y_{1,t-1}, y_{2,t-1}$ respectively), but also contemporaneous dynamics are modeled (through the coefficients a). This creates instantaneous feedback as $y_{2,t}$ is determined by $y_{1,t}$ and vice versa, and thus a priori restrictions are needed to identify the system. Contrary to the innovations in a VAR model, the innovations in a SVAR model are not allowed to be correlated across equations, in other words, contemporaneously correlated. Note furthermore that due to the inclusion of contemporaneous effects, the parameters representing the intercept (g), lagged effects (γ) and innovations (u) have a different interpretation as in the standard VAR equations 1.1 and 1.2.

Molenaar, 1987; Ferrer, Widaman, Card, Selig, & Little, 2008).

Structural models were criticized by, for example, Sims (1980), who claimed that the restrictions to identify such models are incredible and inappropriate. He found support in the fact that such theoretical structural models were out-forecasted by atheoretical models or reduced form models such as the univariate form of VAR (i.e., AR; Kennedy, 2003). Therefore, Sims advocated the standard VAR model, which can be seen as the reduced form of a SVAR model. In the standard or reduced VAR model used in this thesis, no a priori restrictions are necessary as only lagged effects are included, banning the contemporaneous effects to the innovation matrix. As we often do not know beforehand which variables are important to understand a certain process, Sims suggested including all possibly relevant variables and as many lags as possible. This leads to an overparameterized VAR, creating inevitably multicollinearity between the regressors, and therefore it is advised to pare down the model through model selection. Note that as the reduced form has always less model parameters (no contemporaneous effects) than the SVAR model, there is no one to one mapping between the structural and the reduced model. Instead, a VAR model is mathematically equivalent to multiple SVAR models (Brandt & Williams, 2007).

Besides its a-theoretical nature, most of the controversy related to the VAR model arises from the difficulties of interpreting its parameters. Probably the start of the confusion was the invention of *Granger causality*, which can be empirically tested in a VAR model. Granger causality is a data driven approach and starts with the Hume inspired notion that causes must precede their effects, implying that time is essential when studying causality (Granger, 1969; Hoover, 1993). Granger causality can be characterized as follows: Variable X is a Granger-cause of variable Y exactly when prediction of Y at time t is improved by taking into account all past values of X, in addition to all other past information (Kennedy, 2003, p.63). Aside from the problem that it is impracticable to have all relevant past information included in a VAR model, Granger causality is not sufficient to infer causality. For example, Christmas shopping is likely to Granger-cause or predict Christmas very accurately, but this does not mean that Christmas shopping causes Christmas. Even when nobody would do any shopping, Christmas would still arrive. Thus, in the best case, Granger causality just implies predictability or precedence, which quite possibly has little to do with causality (Leamer, 1985).

Furthermore, in economics and also in psychology, a lot of voices have been raised to additionally study contemporaneous effects and contemporaneous causality. Contemporaneous effects and especially contemporaneous causality as it is represented in a SVAR model were seen by Granger as artefacts and unreal as causality is, according to Granger, a temporal notion. The seemingly contemporaneous or instantaneous causality would disappear if there were no omitted causal variables, or if the relevant variables were measured continuously or near continuously (Hoover, 1993). However, often such continuous data is not available, and some variables, such as alcohol consumption, are simply not continuous processes, making it unlikely that contemporaneous effects can be ruled out in

economics and psychology.

Unfortunately, even though SVAR and VAR are mathematically equivalent, they have a very different physical interpretation (Lütkepohl, 2007, p. 48). Thus, if one assumes that the contemporaneous effects have to be explicitly modeled, a VAR model is merely a starting point, and again identification restrictions on the error structure are necessary, which require some sort of theory. Sims, for example, suggested all kinds of statistical tricks such as orthogonalization of the innovations (e.g., a Choleski decomposition) to get from a VAR model to a SVAR model, ironically bringing back the "incredible identification restrictions" he had criticized before (Kennedy, 2003, p. 306).

To circumvent the issue of problematic identification restrictions, data driven methods to estimate SVAR models have been recently developed. In fMRI research, the Group Iterative Multiple Model Estimation (GIMME) has been developed, using Structural Equation Modeling techniques and modeling both contemporaneous and lagged effects (Gates & Molenaar, 2012). This method has been recently extended to be also applicable to, for example, daily diary data (Beltz, Wright, Sprague, & Molenaar, in press). In econometrics, a graph-theoretic approach, similar to the method in Chapter 6 of this thesis, is used to infer a SVAR model (Hoover, 2005). Graph theory uses mathematical techniques to draw conclusions based on the probability distributions of variables. For example, graph theory can be used to determine if variables A and B are directly related or whether they are actually independent given a third variable C, a common cause (Pearl, 2000; see also the example on common causes in the previous section). In essence, graph theory uses simple relationships of probabilistic dependence and independence, and interestingly time is often not taken into account to infer the direction of edges in the graph. When applied to the SVAR method, first a VAR model is estimated and then a graph theory algorithm is used to find the best fitting SVAR model in a data driven way (Hoover, 2005).

Besides the issue of whether a VAR or SVAR model is preferable, other issues has been raised that are problematic for the interpretation of the VAR models currently used in psychological research. For example, although it is advised to use a large number of variables in a VAR model, as is done in network research, there is still a lack of model selection afterwards, leaving the model overparameterized. Using techniques to pare down the model with Least Absolute Shrinkage and Selection Operator (LASSO) methods would be a step forward (Tibshirani, 1996). Paring down the model, however, is likely not without problems, as it is suggested in the econometric literature that the standard ordinary least square method used in this thesis can only be applied when the right hand side of every VAR equation contains the same lagged variables (Enders, 2008). This suggests that in order to select more parsimonious models, different techniques such as Bayesian or SEM based (multilevel) VAR techniques, in which equations can be estimated simultaneously, may be required (Gates & Molenaar, 2012; Schuurman et al., in press).

Furthermore, in order to compare edges in the network and between networks, several issues need

to be taken into account: standardization, measurement error and unique versus shared variance. If, for example, measurement error is not accounted for, edges might be underestimated or overestimated (Cooley & LeRoy, 1985; Schuurman et al., 2015). All of these issues have been mentioned before in this thesis and progress has been made in order to resolve them (Bulteel et al., in press; Schuurman et al., 2015, 2016), but as most of these techniques are still fairly new or under development, these issues are not always accounted for in practice.

Another important goal in the network analyses of this thesis has been to unite idiographic and nomothetic approaches (Molenaar, 2004; Nesselroade & Molenaar, 2010; Steele, Ferrer, & Nesselroade, 2013). On the one hand, progress has been made by developing a multilevel VAR model in which individual and group effects can be modeled at once. However, a multilevel VAR model is still quite restrictive, as the individual (random) effects are assumed to come from a multivariate normal distribution. This might not always be a plausible assumption. In Chapter 5, for example, it was shown that individual differences in temporal dynamics can become quite complicated as the process under study might be non-stationary, and thus the differences in dynamics between individuals also occur over time. In this case, just one variable (valence) was taken into account, but in networks where sometimes 21 variables are studied, the differences between individuals, especially when the process under study is non-stationary, might be much larger than what we find when fitting a restrictive multilevel VAR model to the data. A solution would be to fit a (time-varying) VAR model for every individual to get an estimate of the differences between individuals and to see if a (stationary) multivariate normal distribution is tenable (see also Molenaar, Beltz, Gates, & Wilson, 2016). Another option would be to use a multilevel VAR model with a non-parametric distribution for the random effects, or the GIMME approach, in which individual and a group networks are made without assuming a specific distribution for individual networks (Beltz et al., in press).

In the end, VAR models seem to be not that awful at all. Indeed, causal interpretation is something one should stay far from, but this problem is not restricted to the VAR model. With just observational data and without a developed theory, no model, including the SVAR model, will give you a reliable causal interpretation of the mechanism under study. Econometricians do agree that VAR models can be used for prediction (Cooley & LeRoy, 1985; Harvey, 1997; Leamer, 1985) and that is something that can be useful for psychologists too. Although you might find relations like Christmas shopping predicting Christmas, or spurious relations due to missing variables or a misspecification of the model, a VAR model can be a starting point for generating and testing hypotheses. Following from this, it is important to not only make more advanced and complex modeling techniques, but to also try to gather experimental data. As the world is becoming full of new techniques like smart phones, continuous streaming of individuals' lives should not be a too far off possibility. In this way, natural experiments (e.g., when somebody receives bad news) could be captured immediately. In general, we should go

back to studying simpler processes on which we already have good intuitions or developed theories (see also Chapter 7 of this thesis), and try to develop real life experiments in order to learn about causality *and* the interpretation of our models.

8.3 Dynamical networks in psychology: More than a pretty picture?

A conclusion that follows from the previous sections is that the VAR model can be a fruitful basis for constructing networks. The question that is left then is whether these networks are something beyond the visualization of the parameters of a VAR model. First of all, we should realize from the first part of the discussion that, as with any model, network models are not always the answer. Networks are often simple and powerful representations, and this abstraction, where not everything is and can be visualized, is both a virtue and a vice. Considering the dynamic networks studied in this thesis, it can be noticed that the intercept and the mean of a VAR model were not visualized in the networks. In general, the mean does not find its way into network representations as the idea behind a network approach is to zoom in on the interaction between variables. It is this focus on interaction, however, that is not always necessary, as sometimes the answer to a problem lies in changes in the mean or in the individual elements of a network.

That a network model is not always the answer to a problem can be seen in the work of Duijn et al. (2014). These researchers studied the disruption of criminal networks and showed, interestingly, that to break down such networks a network model as such was not necessary. Instead of analyzing the interaction between criminals or variables, studying the "substitutability" of an individual criminal was the effective method to disrupt the criminal network, because the most specialized criminals were the most difficult to substitute once removed from the network. As the interaction between nodes or criminals was not crucial here, a network approach was not necessary to influence the system. Similarly in psychology, to help patients in practice, a network model may sometimes be unnecessary, as the plain increase of a certain symptom might already give enough valuable information to help and treat individuals with a disorder. For example, in elderly individuals, just the symptom sleep disturbance seems to be a good and simple predictor of a full blown depression (Livingston, Blizard, & Mann, 1993).

Assuming that the interaction between variables is relevant for solving the research question of interest, one might still wonder if a network model is anything more than a visualization of, in this case, the parameters of a VAR model. Indeed, one could argue that it goes beyond visualization, as now *network analyses* such as centrality analyses have become a part of our statistical toolbox. However, some network analyses, especially the centrality analyses, seem to be rather instable and a

good interpretation of them is still lacking. For example, in Chapter 2 betweenness centrality analyses regarding worry and neuroticism could not be replicated and in Chapter 3 many of the centrality analyses did not replicate in both datasets. Furthermore, centrality measures such as out-strength are interpreted according to how they are used in social networks, but it is not at all clear if a high out-strength (or any other centrality measure) can be interpreted in the same way in psychology, especially since the parameters they are based on are already difficult to interpret without the context of a network.

Moreover, even in social networks centrality measures cannot always be interpreted as initially thought of. Consider again the research on criminal networks (Duijn et al., 2014). In these networks, nodes (or criminals) with the highest centrality were not the criminals with the most influence nor the most important ones, and thus attacking the most central criminal did not lead to a disruption of the network (Duijn et al., 2014; Firmani, Italiano, & Laura, 2014). Interpretations of centrality measures in psychological networks are further complicated by the fact that we do not know for sure if the nodes are really distinct entities, as in social networks, where individuals are a priori distinguishable. Thus, before these issues are addressed, the use of centrality analyses in especially clinical psychological networks seems problematic.

What appear to be more robust are the global or overall measures such as the density analyses done in Chapter 3. Although also not entirely unproblematic, it seems that there is a clear idea of how to interpret the density measure, and there are already some simple and plausible theories behind it which can be directly tested. One example is the theory that individuals with depression get more stuck in negative emotions, and therefore have stronger connections or predictability between emotions over time (Pe et al., 2015). The visualization is in this case also a very powerful tool to get a better grip on the theoretical idea. Density analysis thus seems to be a promising line of research that has been directly inspired by the network approach and can be further tested.

So are dynamical networks in psychology more than a pretty picture? They have opened up a whole new research paradigm, generating new questions and ideas. However, the importance of the kind of networks used in this thesis should not be overestimated: as is clear from the above, the VAR model they rely on is not without flaws, even though progress has been made in this thesis regarding extensions of VAR models. Furthermore, the network approach should not become a chant, as sometimes putting variables in a network and looking at their interactions is unnecessary and only leads to extra complications. So yes, dynamical networks are arguably more than just a pretty picture, even though the real proof is yet to come.

References

aan het Rot, M., Hogenelst, K., & Schoevers, R. A. (2012). Mood disorders in everyday life: A systematic review of experience sampling and ecological momentary assessment studies. *Clinical Psychology Review*, *32*(6), 510–523.

Adolph, K. E., Robinson, S. R., Young, J. W., & Gill-Alvarez, F. (2008). What is the shape of developmental change? *Psychological Review*, *115*(3), 527–543.

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing.* Washington, DC: Author.

Ahmed, A., & Xing, E. P. (2009). Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, *106*(29), 11878–11883.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Albright, T. D. (1984). Direction and orientation selectivity of neurons in visual area MT of the macaque. *Journal of Neurophysiology*, *52*(6), 1106–1130.

Albright, T. D., & Stoner, G. R. (1995). Visual motion perception. *Proceedings of the National Academy of Sciences*, *92*(7), 2433–2440.

Amunts, K., Kedo, O., Kindler, M., Pieperhoff, P., Mohlberg, H., Shah, N., ... Zilles, K. (2005). Cytoarchitectonic mapping of the human amygdala, hippocampal region and entorhinal cortex: Intersubject variability and probability maps. *Anatomy and Embryology*, *210*(5-6), 343–352.

Anderson, G. M. (2015). Autism biomarkers: challenges, pitfalls and possibilities. *Journal of Autism and Developmental Disorders*, *45*(4), 1103–1113.

APA. (2000). *Diagnostic and statistical manual of mental disorders (4th ed., Text Revision).* Washington, DC: Author.

Arnau, R. C., Meagher, M. W., Norris, M. P., & Bramson, R. (2001). Psychometric evaluation of the Beck Depression Inventory-II with primary care medical patients. *Health Psychology*, *20*(2), 112–119.

Arnold, B. C., & Strauss, D. (1991). Pseudolikelihood estimation: some examples. *Sankhyā: The Indian Journal of Statistics, Series B*, *53*, 233–243.

Ashton, M. C., & Lee, K. (2012). On models of personality structure. *European Journal of Personality*, *26*(4), 433–434.

Baas, M., De Dreu, C. K., & Nijstad, B. A. (2008). A meta-analysis of 25 years of mood-creativity research: Hedonic tone, activation, or regulatory focus? *Psychological Bulletin*, *134*(6), 779–806.

Baltagi, B. (2008). *Econometric analysis of panel data*. Chichester: John Wiley & Sons.

Barabási, A.-L. (2011). The network takeover. *Nature Physics*, *8*(1), 14.

Barlow, D. H., Sauer-Zavala, S., Carl, J. R., Bullis, J. R., & Ellard, K. K. (2014). The nature, diagnosis, and treatment of neuroticism back to the future. *Clinical Psychological Science*, *2*(3), 344–365.

Barnett, M. K. (1956). The development of thermometry and the temperature concept. *Osiris*, *12*, 269–341.

Barrat, A., Barthelemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, *101*(11), 3747–3752.

Barrett, L. F. (1998). Discrete emotions or dimensions? The role of valence focus and arousal focus. *Cognition and Emotion*, *12*(4), 579–599.

Basner, M., Dinges, D. F., Mollicone, D., Ecker, A., Jones, C. W., Hyder, E. C., ... Sutton, J. (2013). Mars 520-d mission simulation reveals protracted crew hypokinesis and alterations of sleep duration and timing. *Proceedings of the National Academy of Sciences*, *110*(7), 2635–2640.

Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using s4 classes [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=lme4` (R package version 0.999999-0)

Bechtel, W. C. (2008). *Mental Mechanisms*. London: Routledge.

Bechtel, W. C., & Richardson, R. C. (1993). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton, NJ: Princeton University Press.

Beck, A. (1964). Thinking and depression II. Theory and Therapy. *Archives of General Psychiatry*, *10*(6), 561–571.

Beck, A. (1979). *Cognitive therapy of depression*. New York: Guilford.

Beck, A., Steer, R. A., Ball, R., & Ranieri, W. F. (1996). Comparison of Beck Depression Inventories-IA and -II in psychiatric outpatients. *Journal of Personality Assessment*, *67*(3), 588–597.

Beck, A., Steer, R. A., & Brown, G. K. (1996). Beck depression inventory-II [Computer software manual]. San Antonio, TX: Psychological Corporation.

Beck, A., Ward, C., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561–571.

Behrens, T., Woolrich, M., Jenkinson, M., Johansen-Berg, H., Nunes, R., Clare, S., ... Smith, S.

(2003). Characterization and propagation of uncertainty in diffusion-weighted MR imaging. *Magnetic Resonance in Medicine*, *50*(5), 1077–1088.

Belsley, D. A., & Kuh, E. (1973). Time-varying parameter structures: An overview. *Annals of Economic and Social Measurement*, *2*(4), 375–379.

Beltz, A. M., Wright, A. G., Sprague, B. N., & Molenaar, P. C. M. (in press). Bridging the nomothetic and idiographic approaches to the analysis of clinical data. *Assessment*.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *57*, 289–300.

Bhar, S. S., Gelfand, L. A., Schmid, S. P., Gallop, R., DeRubeis, R. J., Hollon, S. D., ... Beck, A. T. (2008). Sequence of improvement in depressive symptoms across cognitive therapy and pharmacotherapy. *Journal of Affective Disorders*, *110*(1), 161–166.

Bisconti, T. L., Bergeman, C., & Boker, S. M. (2004). Emotional well-being in recently bereaved widows: A dynamical systems approach. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *59*(4), 158–167.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics Reports*, *424*(4), 175–308.

Boker, S. M., Molenaar, P., & Nesselroade, J. R. (2009). Issues in intraindividual variability: Individual differences in equilibria and dynamics over multiple time scales. *Psychology and Aging*, *24*(4), 858–862.

Boker, S. M., Rotondo, J. L., Xu, M., & King, K. (2002). Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological Methods*, *7*(3), 338–355.

Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, *54*(1), 579–616.

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. New York, NY: Guilford Press.

Bollen, K. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, *53*(1), 605–634.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*(2), 305.

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, *27*(1), 55–71.

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science*, *323*(5916), 892–895.

Borkovec, T., Ray, W. J., & Stober, J. (1998). Worry: A cognitive phenomenon intimately linked to

affective, physiological, and interpersonal behavioral processes. *Cognitive Therapy and Research*, *22*(6), 561–576.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

Borsboom, D. (2008). Latent variable theory. *Measurement: Interdisciplinarity Research and Perspectives*, *6*, 25–53.

Borsboom, D., & Cramer, A. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*, 91–121.

Borsboom, D., Cramer, A., Schmittmann, V. D., Epskamp, S., & Waldorp, L. J. (2011). The small world of psychopathology. *PloS One*, *6*(11), e27407.

Borsboom, D., Cramer, A. O., Kievit, R. A., Scholten, A. Z., & Franić, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 135–170). Charlotte, NC: Information Age Publishing.

Borsboom, D., Kievit, R. A., Cervone, D., & Hood, S. B. (2009). The two disciplines of scientific psychology, or: The disunity of psychology as a working hypothesis. In J. Valsiner, P. C. M. Molenaar, M. C. D. P. Lyra, & N. Chaudary (Eds.), *Dynamic process methodology in the social and developmental sciences* (pp. 67–97). New York: Springer.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071.

Bos, E. H., & Wanders, R. B. (2016). Group-level symptom networks in depression. *JAMA Psychiatry*, *73*(4), 411.

Boschloo, L., van Borkulo, C. D., Rhemtulla, M., Keyes, K. M., Borsboom, D., & Schoevers, R. A. (2015). The network structure of symptoms of the diagnostic and statistical manual of mental disorders. *PloS One*, *10*(9), e0137621.

Box, J., Jenkins, G., & Reinsel, G. (1994). *Time series analysis: Forecasting and control*. Englewood Cliffs, NJ: Prentice Hal.

Brandt, P. T., & Williams, J. T. (2007). *Multiple Time Series Models. Series: Quantitative Applications in the Social Sciences.* Thousand Oaks, CA: Sage Publications Inc.

Bringmann, L. F., Hamaker, E. L., Vigo, D. E., Aubert, A., Borsboom, D., & Tuerlinckx, F. (in press). Changing dynamics: Time-varying autoregressive models using generalized additive modeling. *Psychological Methods*.

Bringmann, L. F., Lemmens, L. H., Huibers, M. J., Borsboom, D., & Tuerlinckx, F. (2015). Revealing the dynamic network structure of the Beck Depression Inventory-II. *Psychological Medicine*,

45(04), 747–757.

Bringmann, L. F., Scholte, H. S., & Waldorp, L. J. (2013). Matching structural, effective, and functional connectivity: A comparison between structural equation modeling and ancestral graphs. *Brain Connectivity*, *3*(4), 375–385.

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.

Bronfenbrenner, U. (1986). Ecology of the family as a context for human development: Research perspectives. *Developmental Psychology*, *22*(6), 723–742.

Brosschot, J. F., Gerin, W., & Thayer, J. F. (2006). The perseverative cognition hypothesis: A review of worry, prolonged stress-related physiological activation, and health. *Journal of Psychosomatic Research*, *60*(2), 113–124.

Büchel, C., & Friston, K. (1997). Modulation of connectivity in visual pathways by attention: cortical interactions evaluated with structural equation modelling and fmri. *Cerebral Cortex*, *7*(8), 768–778.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, *10*(3), 186–198.

Bulteel, K., Tuerlinckx, F., Brose, A., & Ceulemans, E. (in press). Using raw var regression coefficients to build networks can be misleading. *Multivariate Behavioral Research*.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological Methods and Research*, *33*(2), 261–304.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105.

Campbell, J. (2010). Independence of variables in mental causation. *Philosophical Issues*, *20*(1), 64–79.

Chang, H. (1995). Circularity and reliability in measurement. *Perspectives on Science*, *3*, 153–172.

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford: Oxford University Press.

Chatfield, C. (2003). *The analysis of time series: An introduction* (Fifth ed.). Boca Raton, FL: Chapman and Hall/CRC.

Choppin, B. (1985). Lessons for psychometrics from thermometry. *Evaluation in Education*, *9*(1), 9–12.

Chow, S.-M., Zu, J., Shifren, K., & Zhang, G. (2011). Dynamic factor analysis models with time-varying parameters. *Multivariate Behavioral Research*, *46*(2), 303–339.

Clapham, M. M. (2004). The convergent validity of the Torrance tests of creative thinking and

creativity interest inventories. *Educational and Psychological Measurement*, *64*(5), 828–841.

Coleman, J. S. (1964). *Introduction to mathematical sociology.* New York, NY: Free Press.

Cooley, T. F., & LeRoy, S. F. (1985). Atheoretical macroeconometrics: A critique. *Journal of Monetary Economics*, *16*(3), 283–308.

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.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87-114.

Cramer, A. (2012). Why the item "23 + 1" is not in a depression questionnaire: Validity from a network perspective. *Measurement: Interdisciplinary Research and Perspective*, *10*(1-2), 50–54.

Cramer, A., Borsboom, D., Aggen, S., & Kendler, K. (2012). The pathoplasticity of dysphoric episodes: Differential impact of stressful life events on the pattern of depressive symptom inter-correlations. *Psychological Medicine*, *42*, 957–965.

Cramer, A., van der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., ... Borsboom, D. (2012a). Dimensions of normal personality as networks in search of equilibrium: You can't like parties if you don't like people. *European Journal of Personality*, *26*(4), 414–431.

Cramer, A., van der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., ... Borsboom, D. (2012b). Measurable like temperature or mereological like flocking? On the nature of personality traits. *European Journal of Personality*, *26*(4), 451–459.

Cramer, A., Waldorp, L. J., van der Maas, H. L., & Borsboom, D. (2010). Comorbidity: A network perspective. *Behavioral and Brain Sciences*, *33*(2-3), 137–150.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281.

Csikszentmihalyi, M., & Larson, R. (2014). Validity and reliability of the experience-sampling method. In M. Csikszentmihalyi (Ed.), *Flow and the foundations of positive psychology* (pp. 35–54). New York: Springer.

Cuijpers, P., van Straten, A., Andersson, G., & van Oppen, P. (2008). Psychotherapy for depression in adults: A meta-analysis of comparative outcome studies. *Journal of Consulting and Clinical Psychology*, *76*(6), 909–922.

Culp, S. (1994). Defending robustness: The bacterial mesosome as a test case. *PSA 1994: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *1*, 46–57.

Czsikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience sample method. *Journal of Nervous and Mental Disease*, *175*, 526–536.

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *The Annals of Statistics*,

25(1), 1–37.

Danks, D., Fancsali, S., Glymour, C., & Scheines, R. (2010). Comorbid science? *Behavioral and Brain Sciences*, *33*(2-3), 153–155.

Deboeck, P. R. (2013). Dynamical systems and models of continuous time. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in psychology: Vol. 2: Statistical analysis* (pp. 411–431). Oxford, England: Oxford University Press.

de Haan-Rietdijk, S., Gottman, J. M., Bergeman, C. S., & Hamaker, E. L. (2014). Get over it! A multilevel threshold autoregressive model for state-dependent affect regulation. *Psychometrika*, *81*, 217–241.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Asociation*, *74*(366a), 427–431.

Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, *45*(3), 259–268.

Duijn, P. A., Kashirin, V., & Sloot, P. M. (2014). The relative ineffectiveness of criminal network disruption. *Scientific Reports*, *4*, 4238.

Ebner-Priemer, U. W., Eid, M., Kleindienst, N., Stabenow, S., & Trull, T. J. (2009). Analytic strategies for understanding affective (in)stability and other dynamic processes in psychopathology. *Journal of Abnormal Psychology*, *118*(1), 195–202.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. New York: Chapman and Hall/CRC.

Eichler, M. (2005). A graphical approach for evaluating effective connectivity in neural systems. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *360*(1457), 953–967.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*(1), 179–197.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*(3), 380–396.

Embretson, S. E. (2004). The second century of ability testing: Some predictions and speculations. *Measurement: Interdisciplinary Research and Perspectives*, *2*(1), 1–32.

Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*(4), 343–368.

Enders, W. (2008). *Applied econometric time series*. New York, NY: John Wiley & Sons.

Epskamp, S., Cramer, A., 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.

Epskamp, S., Maris, G., Waldorp, L. J., & Borsboom, D. (in press). Network psychometrics. In *Handbook of Psychometrics.* New York, NY: John Wiley & Sons.

Eronen, M. I. (2012). Pluralistic physicalism and the causal exclusion argument. *European Journal for Philosophy of Science*, *2*(2), 219–232.

Fan, J., & Yao, Q. (2003). *Nonlinear time series: Nonparametric and parametric methods.* New York, NY: Springer.

Faraway, J. J. (2006). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models.* Boca Raton, FL: Chapman and Hall/CRC.

Felleman, D. J., & van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*(1), 1–47.

Ferrer, E., & Nesselroade, J. R. (2003). Modeling affective processes in dyadic relations via dynamic factor analysis. *Emotion*, *3*(4), 344-360.

Ferrer, E., Widaman, K. F., Card, N., Selig, J., & Little, T. (2008). Dynamic factor analysis of dyadic affective processes with inter-group differences. In N. Card, J. Selig, & T. Little (Eds.), *Modeling dyadic and interdependent data in the developmental and behavioral sciences* (pp. 107–137). Psychology Press. Hillsdale, NJ.

Fieuws, S., & Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, *62*(2), 424–431.

Firmani, D., Italiano, G. F., & Laura, L. (2014). The (not so) critical nodes of criminal networks. In L. M. Aiello & D. McFralan (Eds.), *Social Informatics* (pp. 87–96). New York, NY: Springer.

Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (2008). *Longitudinal data analysis.* Boca Raton, FL: Chapman and Hall/CRC.

Fleeson, W. (2001). Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, *80*(6), 1011–1027.

Fleeson, W. (2004). Moving personality beyond the person-situation debate the challenge and the opportunity of within-person variability. *Current Directions in Psychological Science*, *13*(2), 83–87.

Fournier, J. C., DeRubeis, R. J., Hollon, S. D., Gallop, R., Shelton, R. C., & Amsterdam, J. D. (2013). Differential change in specific depressive symptoms during antidepressant medication or cognitive therapy. *Behaviour Research and Therapy*, *51*(7), 392–398.

Fredrickson, B. L., & Joiner, T. (2002). Positive emotions trigger upward spirals toward emotional well-being. *Psychological Science*, *13*(2), 172–175.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239.

Frewen, P. A., Allen, S. L., Lanius, R. A., & Neufeld, R. W. (2012). Perceived causal relations

novel methodology for assessing client attributions about causal associations between variables including symptoms and functional impairment. *Assessment*, *19*(4), 480–493.

Frewen, P. A., Schmittmann, V. D., Bringmann, L. F., & Borsboom, D. (2013). Perceived causal relations between anxiety, posttraumatic stress and depression: Extension to moderation, mediation, and network analysis. *European Journal of Psychotraumatology*, *4*, 20656.

Fried, E. I. (2015). Problematic assumptions have slowed down depression research: Why symptoms, not syndromes are the way forward. *Frontiers in Psychology*, *6*, 1–11.

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., 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.

Fried, E. I., & Nesse, R. M. (2014). The impact of individual depressive symptoms on impairment of psychosocial functioning. *PLoS One*, *9*(2), e90311.

Fried, E. I., & Nesse, R. M. (2015). Depression sum-scores dont add up: why analyzing specific depression symptoms is essential. *BMC Medicine*, *13*, 1–11.

Fried, E. I., Nesse, R. M., Zivin, K., Guille, C., & Sen, S. (2014). Depression is more than the sum score of its parts: Individual DSM symptoms have different risk factors. *Psychological Medicine*, *44*(10), 2067–2076.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441.

Frigerio, A., Giordani, A., & Mari, L. (2010). Outline of a general model of measurement. *Synthese*, *175*(2), 123–149.

Friston, K. J. (2011). Functional and effective connectivity: a review. *Brain Connectivity*, *1*(1), 13–36.

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, *19*(4), 1273–1302.

Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, *21*(11), 1129–1164.

Funatogawa, I., Funatogawa, T., & Ohashi, Y. (2007). An autoregressive linear mixed effects model for the analysis of longitudinal data which show profiles approaching asymptotes. *Statistics in Medicine*, *26*(9), 2113–2130.

Gallese, V., Rochat, M., Cossu, G., & Sinigaglia, C. (2009). Motor cognition and its role in the phylogeny and ontogeny of action understanding. *Developmental Psychology*, *45*(1), 103–113.

Gates, K. M., & Molenaar, P. C. M. (2012). Group search algorithm recovers effective connectivity maps for individuals in homogeneous and heterogeneous samples. *NeuroImage*, *63*(1), 310–319.

Gates, K. M., Molenaar, P. C. M., Hillary, F. G., Ram, N., & Rovine, M. J. (2010). Automatic search for fMRI connectivity mapping: An alternative to Granger causality testing using formal equivalences among SEM path modeling, VAR, and unified SEM. *NeuroImage*, *50*(3), 1118–1125.

Gather, U., Imhoff, M., & Fried, R. (2002). Graphical models for multivariate time series from intensive care monitoring. *Statistics in Medicine*, *21*(18), 2685–2701.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Geschwind, N., Peeters, F., Drukker, M., Van Os, J., & Wichers, M. (2011). Mindfulness training increases momentary positive emotions and reward experience in adults vulnerable to depression: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *79*(5), 618–628.

Gigerenzer, G. (1987). Probabilistic thinking and the fight against subjectivity. In L. Krüger, G. Gigerenzer, & M. S. Morgan (Eds.), *The probabilistic revolution: Vol. 2. ideas in the sciences* (pp. 11–33). Cambridge, MA: MIT Press.

Giraitis, L., Kapetanios, G., & Yates, T. (2014). Inference on stochastic time-varying coefficient models. *Journal of Econometrics*, *179*(1), 46–65.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), 7821–7826.

Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, *21*(2), 215–223.

Gonçalves, M. S., & Hall, D. A. (2003). Connectivity analysis with structural equation modelling: An example of the effects of voxel selection. *NeuroImage*, *20*(3), 1455–1467.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, *37*, 424–438.

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, *41*(10), 1409–1422.

Gruber, J., Eidelman, P., & Harvey, A. G. (2008). Transdiagnostic emotion regulation processes in bipolar disorder and insomnia. *Behaviour Research and Therapy*, *46*(9), 1096–1100.

Gu, G. (2002). *Smoothing spline anova models*. New York: Springer.

Guye, M., Bartolomei, F., & Ranjeva, J.-P. (2008). Imaging structural and functional connectivity: Towards a unified definition of human brain organization? *Current Opinion in Neurology*, *21*(4),

393–403.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. New York, NY: Cambridge University Press.

Haig, B. D., & Vertue, F. M. (2010). Extending the network perspective on comorbidity. *Behavioral and Brain Sciences*, *33*(2-3), 158–158.

Hamaker, E. L. (2009). Using information criteria to determine the number of regimes in threshold autoregressive models. *Journal of Mathematical Psychology*, *53*(6), 518–529.

Hamaker, E. L. (2012). Why researchers should think within-person: A paradigmatic rationale. In M. R. Mehl & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 43–61). New York, NY: Guilford.

Hamaker, E. L., Ceulemans, E., Grasman, R., & Tuerlinckx, F. (2015). Modeling affect dynamics: State of the art and future challenges. *Emotion Review*, *7*(4), 316–322.

Hamaker, E. L., & Dolan, C. V. (2009). Idiographic data analysis: Quantitative methods from simple to advanced. In J. Valsiner, P. C. M. Molenaar, M. Lyra, & N. Chaudhary (Eds.), *Dynamic process methodology in the social and developmental sciences* (pp. 191–216). New York, NY: Springer-Verlag.

Hamaker, E. L., Dolan, C. V., & Molenaar, P. C. M. (2003). ARMA-based SEM when the number of time points T exceeds the number of cases N: Raw data maximum likelihood. *Structural Equation Modeling*, *10*(3), 352–379.

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. (2014). To center or not to center? Investigating inertia with a multilevel autoregressive model. *Frontiers in Psychology*, *5*. doi: 10.3389/fpsyg.2014.01492.

Hamaker, E. L., Grasman, R. P. P. P., & Kamphuis, J. H. (2010). Regime-switching models to study psychological processes. In P. C. M. Molenaar & K. Newell (Eds.), *Individual pathways of change* (pp. 155–168). Washington, DC: American Psychological Association.

Hamilton, J. D. (1994). *Time series analysis*. Princeton, NJ: Princeton university press.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*, *23*(1), 56–62.

Hardeveld, F., Spijker, J., De Graaf, R., Nolen, W., & Beekman, A. (2010). Prevalence and predictors of recurrence of major depressive disorder in the adult population. *Acta Psychiatrica Scandinavica*, *122*(3), 184–191.

Härdle, W., Lütkepohl, H., & Chen, R. (1997). A review of nonparametric time series analysis. *International Statistical Review*, *65*(1), 49–72.

Harvey, A. (1997). Trends, cycles and autoregressions. *The Economic Journal*, *107*(440), 192–201.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. Boca Raton, FL: Chapman and Hall/CRC.

He, Y., & Evans, A. (2010). Graph theoretical modeling of brain connectivity. *Current Opinion in Neurology*, *23*(4), 341–350.

Hendrickx, D. M., Hendriks, M. M., Eilers, P. H., Smilde, A. K., & Hoefsloot, H. C. (2011). Reverse engineering of metabolic networks, a critical assessment. *Molecular Biosystems*, *7*(2), 511–520.

Heninger, G., Delgado, P., & Charney, D. (1996). The revised monoamine theory of depression: a modulatory role for monoamines, based on new findings from monoamine depletion experiments in humans. *Pharmacopsychiatry*, *29*(1), 2–11.

Hertzog, C., & Nesselroade, J. R. (2003). Assessing psychological change in adulthood: An overview of methodological issues. *Psychology and Aging*, *18*(4), 639–657.

Hoekstra, H., Ormel, J., & De Fruyt, F. (1996). *Handleiding NEO-PI–R en NEO-FFI Big Five persoonlijkheidsvragenlijsten [Dutch manual for NEO-PI–R and NEOFFI Big Five personality questionnaires]*. Lisse, the Netherlands: Swets and Zeitlinger.

Hofmans, J., Kuppens, P., & Allik, J. (2008). Is short in length short in content? An examination of the domain representation of the Ten Item Personality Inventory scales in Dutch language. *Personality and Individual Differences*, *45*(8), 750–755.

Hollon, S. D., & Ponniah, K. (2010). A review of empirically supported psychological therapies for mood disorders in adults. *Depression and Anxiety*, *27*(10), 891–932.

Hood, S. B. (2009). Validity in psychological testing and scientific realism. *Theory and Psychology*, *19*(4), 451–473.

Hoover, K. D. (1993). Causality and temporal order in macroeconomics or why even economists don't know how to get causes from probabilities. *The British Journal for the Philosophy of Science*, *44*(4), 693–710.

Hoover, K. D. (2005). Automatic inference of the contemporaneous causal order of a system of equations. *Econometric Theory*, *21*(01), 69–77.

Horváth, C., & Wieringa, J. E. (2008). Pooling data for the analysis of dynamic marketing systems. *Statistica Neerlandica*, *62*(2), 208–229.

Horwitz, B., Warner, B., Fitzer, J., Tagamets, M.-A., Husain, F. T., & Long, T. W. (2005). Investigating the neural basis for functional and effective connectivity. Application to fMRI. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *360*(1457), 1093–1108.

Hubley, A. M., Zhu, S. M., Sasaki, A., & Gadermann, A. M. (2013). Synthesis of validation practices in two assessment journals: Psychological Assessment and the European Journal of Psychological Assessment. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 193–213). Springer.

Humphry, S. M. (2011). The role of the unit in physics and psychometrics. *Measurement*, *9*(1), 1–24.

Humphry, S. M. (2013). Understanding measurement in light of its origins. *Frontiers in Psychology*, *4*. doi: 10.3389/fpsyg.2013.00113.

Humphry, S. M., & McGrane, J. A. (2010). Is there a contradiction between the network and latent variable perspectives? *Behavioral and Brain Sciences*, *33*(2-3), 160–161.

Jahng, S., Wood, P. K., & Trull, T. J. (2008). Analysis of affective instability in ecological momentary assessment: Indices using successive difference and group comparison via multilevel modeling. *Psychological Methods*, *13*(4), 354–375.

Jbabdi, S., & Johansen-Berg, H. (2011). Tractography: Where do we go from here? *Brain Connectivity*, *1*(3), 169–183.

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, *5*(2), 143–156.

Johansen-Berg, H., & Behrens, T. E. (2006). Just pretty pictures? What diffusion tractography can add in clinical neuroscience. *Current Opinion in Neurology*, *19*(4), 379–385.

Jonas, K. G., & Markon, K. E. (2016). A descriptivist approach to trait conceptualization and inference. *Psychological Review*, *123*(1), 90–96.

Jones, D. K. (2010). Challenges and limitations of quantifying brain connectivity in vivo with diffusion MRI. *Imaging in Medicine*, *2*(3), 341–355.

Jongerling, J., Laurenceau, J.-P., & Hamaker, E. L. (2015). A multilevel AR (1) model: Allowing for inter-individual differences in trait-scores, inertia, and innovation variance. *Multivariate Behavioral Research*, *50*(3), 334–349.

Jonides, J., Lewis, R. L., Nee, D. E., Lustig, C. A., Berman, M. G., & Moore, K. S. (2008). The mind and brain of short-term memory. *Annual Review of Psychology*, *59*, 193-224.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351a), 631–639.

Kaiser, M., & Hilgetag, C. C. (2004). Modelling the development of cortical systems networks. *Neurocomputing*, *58*, 297–302.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Keele, L. J. (2008). *Semiparametric regression for the social sciences*. Chichester, England: John

Wiley & Sons.

Kendler, K. (2012). Levels of explanation in psychiatric and substance use disorders: implications for the development of an etiologically based nosology. *Molecular Psychiatry*, *17*(1), 11–21.

Kendler, K., Zachar, P., & Craver, C. (2011). What kinds of things are psychiatric disorders? *Psychological Medicine*, *41*(06), 1143–1150.

Kennedy, P. (2003). *A guide to econometrics*. Cambridge: MIT press.

Kessler, R., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., ... Wang, P. S. (2003). The epidemiology of major depressive disorder: Results from the National Comorbidity Survey Replication (NCS-R). *JAMA Psychiatry*, *289*(23), 3095–3105.

Kievit, R. A., Frankenhuis, W. E., Waldorp, L. J., & Borsboom, D. (2013). Simpson's paradox in psychological science: a practical guide. *Frontiers in Psychology*, *4*, 513.

Kim, C.-J., & Nelson, C. R. (1999). *State-space models with regime switching: classical and gibbs-sampling approaches with applications*. Cambridge: MIT press.

Kitagawa, G., & Gersch, W. (1985). A smoothness priors time-varying AR coefficient modeling of nonstationary covariance time series. *IEEE Transactions on Automatic Control*, *30*(1), 48–56.

Klerks, P. (2001). The network paradigm applied to criminal organizations: Theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the netherlands. *Connections*, *24*(3), 53–65.

Kline, P. (2000). *Handbook of psychological testing*. London: Routledge.

Koop, G. (2012). Using VARs and TVP-VARs with many macroeconomic variables. *Central European Journal of Economic Modelling and Econometrics*, *4*(3), 143–167.

Koval, P., Brose, A., Pe, M. L., Houben, M., Erbas, Y. I., Champagne, D., & Kuppens, P. (in press). Emotional inertia and external events: The roles of exposure, reactivity, and recovery. *Emotion*.

Koval, P., & Kuppens, P. (2012). Changing emotion dynamics: Individual differences in the effect of anticipatory social stress on emotional inertia. *Emotion*, *12*(2), 256–267.

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*(8), 1412–1427.

Koval, P., Pe, M. L., Meers, K., & Kuppens, P. (2013). Affect dynamics in relation to depressive symptoms: Variable, unstable or inert? *Emotion*, *13*(6), 1132.

Krueger, R. F., DeYoung, C. G., & Markon, K. E. (2010). Toward scientifically useful quantitative models of psychopathology: The importance of a comparative approach. *Behavioral and Brain Sciences*, *33*(2-3), 163–164.

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychological Science*, *21*, 984–991.

Kuppens, P., Champagne, D., & Tuerlinckx, F. (2012). The dynamic interplay between appraisal and core affect in daily life. *Frontiers in Psychology*, *3*, 1–8.

Kuppens, P., Oravecz, Z., & Tuerlinckx, F. (2010). Feelings change: accounting for individual differences in the temporal dynamics of affect. *Journal of Personality and Social Psychology*, *99*(6), 1042–1060.

Kuppens, P., Stouten, J., & Mesquita, B. (2009). Individual differences in emotion components and dynamics: Introduction to the special issue. *Cognition and Emotion*, *23*(7), 1249–1258.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, *54*(1), 159–178.

Kyburg, H. E. (1984). *Theory and measurement*. Cambridge: Cambridge University Press.

Kyburg, H. E. (1992). Measuring errors of measurement. In C. W. Savage & P. Ehrlich (Eds.), *Philosophical and foundational issues in measurement theory* (pp. 75–91). Hillsdale, NJ: Lawrence Erlbaum.

Kyngdon, A. (2013). Descriptive theories of behaviour may allow for the scientific measurement of psychological attributes. *Theory and Psychology*, *23*, 227–250.

Lacasse, J. R., & Leo, J. (2005). Serotonin and depression: A disconnect between the advertisements and the scientific literature. *PLoS Medicine*, *2*(12), e392.

Lamers, F., de Jonge, P., Nolen, W. A., Smit, J. H., Zitman, F. G., Beekman, A. T., & Penninx, B. W. (2010). Identifying depressive subtypes in a large cohort study: Results from the Netherlands Study of Depression and Anxiety (NESDA). *The Journal of Clinical Psychiatry*, *71*(12), 1582–1589.

Lamers, F., Rhebergen, D., Merikangas, K., De Jonge, P., Beekman, A., & Penninx, B. (2012). Stability and transitions of depressive subtypes over a 2-year follow-up. *Psychological Medicine*, *42*(10), 2083–2093.

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, *23*(11), 571–579.

Larsen, J. T., McGraw, A. P., & Cacioppo, J. T. (2001). Can people feel happy and sad at the same time? *Journal of Personality and Social Psychology*, *81*(4), 684–696.

Larsen, R., & Diener, E. (1992). Promises and problems with the circumplex model of emotion. In M. Clark (Ed.), *Review of personality and social psychology (vol. 13): Emotion* (pp. 25–59). Newbury Park, CA: Sage.

Lavie, P. (2001). Sleep-wake as a biological rhythm. *Annual Review of Psychology*, *52*(1), 277–303.

Leamer, E. E. (1985). Vector autoregressions for causal inference? *Carnegie-Rochester Conference Series on Public Policy*, *22*, 255–304.

Lemmens, L., Arntz, A., Peeters, F., Hollon, S., Roefs, A., & Huibers, M. (2015). Clinical effectiveness of cognitive therapy vs. interpersonal psychotherapy for depression: Results of a randomized controlled trial. *Psychological Medicine*, *45*(10), 2095–2110.

Lemmens, L., Arntz, A., Peeters, F., Hollon, S. D., Roefs, A., & Huibers, M. J. (2011). Effectiveness, relapse prevention and mechanisms of change of cognitive therapy vs. interpersonal therapy for depression: Study protocol for a randomised controlled trial. *Trials*, *12*(1), 150–162.

Lissitz, R. W. (Ed.). (2009). *The concept of validity: Revisions, new directions, and applications*. Charlotte, NC: Information Age.

Livingston, G., Blizard, B., & Mann, A. (1993). Does sleep disturbance predict depression in elderly people? A study in inner london. *British Journal of General Practice*, *43*(376), 445–448.

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.

Lütkepohl, H. (2007). *New introduction to multiple time series analysis*. New York: Springer.

Malach, R., Reppas, J., Benson, R., Kwong, K., Jiang, H., Kennedy, W., ... Tootell, R. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences*, *92*(18), 8135–8139.

Markus, K. A. (2010). Questions about networks, measurement, and causation. *Behavioral and Brain Sciences*, *33*(2-3), 164–165.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. London: Routledge.

Mason, W. A., Conrey, F. R., & Smith, E. R. (2007). Situating social influence processes: Dynamic, multidirectional flows of influence within social networks. *Personality and Social Psychology Review*, *11*(3), 279–300.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, *60*, 577–605.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

McFarland, D. J., & Malta, L. S. (2010). Symptoms as latent variables. *Behavioral and Brain Sciences*, *33*(2-3), 165–166.

McIntosh, A., Grady, C., Haxby, J., Ungerrleider, L., & Horwitz, B. (1996). Changes in limbic and prefrontal functional interactions in a working memory task for faces. *Cerebral Cortex*, *6*(4), 571–584.

McKeown, G. J., & Sneddon, I. (2014). Modeling continuous self-report measures of perceived emotion using generalized additive mixed models. *Psychological Methods*, *19*(1), 155-174.

McLaughlin, K. A., Borkovec, T. D., & Sibrava, N. J. (2007). The effects of worry and rumination on affect states and cognitive activity. *Behavior Therapy*, *38*(1), 23–38.

Mclntosh, A., & Gonzalez-Lima, F. (1994). Structural equation modeling and its application to network analysis in functional brain imaging. *Human Brain Mapping*, *2*(1-2), 2–22.

McNally, R. J. (2012). The ontology of posttraumatic stress disorder: Natural kind, social construction, or causal system? *Clinical Psychology: Science and Practice*, *19*(3), 220–228.

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.

Mehl, M. R., & Conner, T. S. E. (2012). *Handbook of research methods for studying daily life*. New York, NY: Guilford Press.

Meney, I., Waterhouse, J., Atkinson, G., Reilly, T., & Davenne, D. (1998). The effect of one night's sleep deprivation on temperature, mood, and physical performance in subjects with different amounts of habitual physical activity. *Chronobiology international*, *15*(4), 349–363.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*(3), 355–383.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge: Cambridge University Press.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory and Psychology*, *10*(5), 639–667.

Michell, J. (2004). Item response models, pathological science and the shape of error: Reply to Borsboom and Mellenbergh. *Theory and Psychology*, *14*(1), 121–129.

Michell, J. (2013). Constructs, inferences, and mental measurement. *New Ideas in Psychology*, *31*(1), 13–21.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*(1), 3–67.

Moberly, N. J., & Watkins, E. R. (2008). Ruminative self-focus and negative affect: An experience sampling study. *Journal of Abnormal Psychology*, *117*(2), 314–323.

Molenaar, P. C. M. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, *50*(2), 181–202.

Molenaar, P. C. M. (1987). Dynamic assessment and adaptive optimization of the psychotherapeutic process. *Behavioral Assessment*, *9*, 389–416.

Molenaar, P. C. M. (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. M. (2010). Latent variable models are network models. *Behavioral and Brain*

Sciences, *33*(2-3), 166.

Molenaar, P. C. M. (2015). The future of analysis of intraindividual variation. In M. Diehl, K. Hooker, & M. J. Sliwinski (Eds.), *Handbook of intraindividual variability across the life span* (pp. 343–356). New York, NY: Routledge, Taylor and Francis.

Molenaar, P. C. M., Beltz, A. M., Gates, K. M., & Wilson, S. J. (2016). State space modeling of time-varying contemporaneous and lagged relations in connectivity maps. *NeuroImage*, *125*, 791–802.

Molenaar, P. C. M., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science*, *18*(2), 112–117.

Molenaar, P. C. M., De Gooijer, J. G., & Schmitz, B. (1992). Dynamic factor analysis of nonstationary multivariate time series. *Psychometrika*, *57*(3), 333–349.

Molenaar, P. C. M., & Newell, K. M. (2003). Direct fit of a theoretical model of phase transition in oscillatory finger motions. *British Journal of Mathematical and Statistical Psychology*, *56*(2), 199–214.

Molenaar, P. C. M., Sinclair, K. O., Rovine, M. J., Ram, N., & Corneal, S. E. (2009). Analyzing developmental processes on an individual level using nonstationary time series modeling. *Developmental psychology*, *45*(1), 260–271.

Molenaar, P. C. M., van Rijn, P., & Hamaker, E. L. (2007). A new class of sem model equivalences and its implications. In S. M. Boker & J. M. Wenger (Eds.), *Data analytic techniques for dynamical systems* (pp. 189–211). Hillsdale, NJ: Erlbaum.

Mumtaz, H., & Surico, P. (2009). Time-varying yield curve dynamics and monetary policy. *Journal of Applied Econometrics*, *24*(6), 895–913.

Muris, P., Roelofs, J., Meesters, C., & Boomsma, P. (2004). Rumination and worry in nonclinical adolescents. *Cognitive Therapy and Research*, *28*(4), 539–554.

Muthén, L. K., & Muthén, B. O. (2012). Mplus User's Guide [Computer software manual]. Los Angeles, CA: Muthén & Muthén.

Nederbragt, H. (2012). Multiple derivability and the reliability and stabilization of theories. In L. Soler, E. Trizio, T. Nickles, & W. C. Wimsatt (Eds.), *Characterizing the robustness of science: After the practice turn in the philosophy of science* (pp. 121–145). Dordrecht: Springer.

Nesselroade, J. R. (2004). Intraindividual variability and short-term change. *Gerontology*, *50*(1), 44–47.

Nesselroade, J. R., & Molenaar, P. C. M. (2010). Emphasizing intraindividual variability in the study of development over the life span. In R. M. Lerner & W. F. Overton (Eds.), *The handbook of life-span development* (pp. 30–54). Hoboken, NJ: Wiley.

Nesselroade, J. R., & Ram, N. (2004). Studying intraindividual variability: What we have learned

that will help us understand lives in context. *Research in Human Development*, *1*(1-2), 9–29.

Newman, M. (2010). *Networks: An introduction*. Oxford: Oxford University Press.

Newman, M., Barabási, A.-L., & Watts, D. J. (2006). *The structure and dynamics of networks*. Princeton: Princeton University Press.

Newton, P., & Shaw, S. (2013). *Validity in educational and psychological assessment*. London: Sage.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*, 1–32.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113.

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., & Vandekerckhove, J. (2011). A hierarchical latent stochastic differential equation model for affective dynamics. *Psychological Methods*, *16*(4), 468.

Paltiel, A. D., Weinstein, M. C., Kimmel, A. D., Seage III, G. R., Losina, E., Zhang, H., ... Walensky, R. P. (2005). Expanded screening for HIV in the United States: An analysis of cost-effectiveness. *New England Journal of Medicine*, *352*(6), 586–595.

Patten, S. (2015). Medical models and metaphors for depression. *Epidemiology and psychiatric sciences*, *24*(4), 303–308.

Pe, M. L., Kircanski, K., Thompson, R. J., Bringmann, L. F., Tuerlinckx, F., Mestdagh, M., ... Gotlib, I. H. (2015). Emotion-network density in major depressive disorder. *Clinical Psychological Science*, *3*(2), 292–300.

Pe, M. L., Koval, P., & Kuppens, P. (2013). Executive well-being: Updating of positive stimuli in working memory is associated with subjective well-being. *Cognition*, *126*(2), 335–340.

Pe, M. L., & Kuppens, P. (2012). The dynamic interplay between emotions in daily life: Augmentation, blunting, and the role of appraisal overlap. *Emotion*, *12*(6), 1320.

Pe, M. L., Raes, F., Koval, P., Brans, K., Verduyn, P., & Kuppens, P. (2013). Interference resolution moderates the impact of rumination and reappraisal on affective experiences in daily life. *Cognition and Emotion*, *27*(3), 492–501.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Penny, W. D., Stephan, K., Mechelli, A., & Friston, K. (2004). Comparing dynamic causal models. *NeuroImage*, *22*(3), 1157–1172.

Pfaff, B. (2008). *Analysis of integrated and cointegrated time series with R*. New York: Springer.

Pole, A., West, M., & Harrison, J. (1994). *Applied Bayesian forecasting and time series analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Pons, P., & Latapy, M. (2005). Computing communities in large networks using random walks. In *Computer and Information Sciences-ISCIS* (pp. 284–293). Berlin: Springer.

Posner, M. I., & Rothbart, M. K. (2007). Research on attention networks as a model for the integration of psychological science. *Annual Review of Psychology*, *58*, 1–23.

Prado, R. (2010). Characterization of latent structure in brain signals. In S.-M. Chow, E. Ferrer, & F. Hsieh (Eds.), *Statistical methods for modeling human dynamics: An interdisciplinary dialogue* (pp. 123–153). New York, NY: Routledge, Taylor and Francis.

Quinn, T. (1990). *Monographs in physical measurement: Temperature.* London: Academic Press.

R Core Team. (2012). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Ramsey, J. D., Hanson, S. J., Hanson, C., Halchenko, Y. O., Poldrack, R. A., & Glymour, C. (2010). Six problems for causal inference from fMRI. *NeuroImage*, *49*(2), 1545–1558.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual review of clinical psychology*, *5*, 27–48.

Reisenzein, R. (1994). Pleasure-arousal theory and the intensity of emotions. *Journal of Personality and Social Psychology*, *67*(3), 525–539.

Rhemtulla, M., Fried, E. I., Aggen, S. H., Tuerlinckx, F., Kendler, K. S., & Borsboom, D. (2016). Network analysis of substance abuse and dependence symptoms. *Drug and Alcohol Dependence*, *161*, 230–237.

Richardson, T., & Spirtes, P. (2002). Ancestral graph Markov models. *Annals of Statistics*, *30*, 962–1030.

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.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15*(3), 351–357.

Roebroeck, A., Formisano, E., & Goebel, R. (2005). Mapping directed influence over the brain using Granger causality and fMRI. *NeuroImage*, *25*(1), 230–242.

Rosmalen, J. G., Wenting, A. M., Roest, A. M., de Jonge, P., & Bos, E. H. (2012). Revealing causal heterogeneity using time series analysis of ambulatory assessments: Application to the association between depression and physical activity after myocardial infarction. *Psychosomatic Medicine*, *74*(4), 377–386.

Rosvall, M., & Bergstrom, C. T. (2010). Mapping change in large networks. *PloS One*, *5*(1), e8694.

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). Oxford, England: Oxford University Press.

Rubinov, M., & Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, *52*(3), 1059–1069.

Rush, A. J., Kovacs, M., Beck, A., Weissenburger, J., & Hollon, S. (1981). Differential effects of cognitive therapy and pharmacotherapy on depressive symptoms. *Journal of Affective Disorders*, *3*(3), 221–229.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*, 1161–1178.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145–172.

Russell, J. A., & Carroll, J. M. (1999). On the bipolarity of positive and negative affect. *Psychological Bulletin*, *125*(1), 3–30.

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, *57*(3), 493–502.

Ruzzano, L., Borsboom, D., & Geurts, H. M. (2015). Repetitive behaviors in autism and obsessive–compulsive disorder: New perspectives from a network analysis. *Journal of autism and developmental disorders*, *45*(1), 192–202.

Rykhlevskaia, E., Gratton, G., & Fabiani, M. (2008). Combining structural and functional neuroimaging data for studying brain connectivity: A review. *Psychophysiology*, *45*(2), 173–187.

Schmid, C. H. (2001). Marginal and dynamic regression models for longitudinal data. *Statistics in Medicine*, *20*(21), 3295–3311.

Schmittmann, V. D., Cramer, A., 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.

Schmitz, B., & Skinner, E. (1993). Perceived control, effort, and academic performance: Interindividual, intraindividual, and multivariate time-series analyses. *Journal of Personality and Social Psychology*, *64*(6), 1010–1028.

Scholte, H. S., Jolij, J., Fahrenfort, J. J., & Lamme, V. A. (2008). Feedforward and recurrent processing in scene segmentation: Electroencephalography and functional magnetic resonance imaging. *Journal of Cognitive Neuroscience*, *20*(11), 2097–2109.

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*, 206–261.

Schuurman, N. K., Grasman, R. P. P. P., & Hamaker, E. L. (in press). A comparison of inverse-Wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivariate Behavioral Research*.

Schuurman, N. K., Houtveen, J. H., & Hamaker, E. L. (2015). Incorporating measurement error in n=1 psychological autoregressive modeling. *Frontiers in Psychology*, *6*.

Schwartz, J. E., & Stone, A. A. (1998). Strategies for analyzing ecological momentary assessment data. *Health Psychology*, *17*(1), 6.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Scollon, C. N., Prieto, C.-K., & Diener, E. (2003). Experience sampling: Promises and pitfalls, strength and weaknesses. *Journal of Happiness Studies*, *4*, 5–34.

Segerstrom, S. C., Tsao, J. C., Alden, L. E., & Craske, M. G. (2000). Worry and rumination: Repetitive thought as a concomitant and predictor of negative mood. *Cognitive Therapy and Research*, *24*(6), 671–688.

Shadish, W. R., Zuur, A. F., & Sullivan, K. J. (2014). Using generalized additive (mixed) models to analyze single case designs. *Journal of School Psychology*, *52*(2), 149–178.

Sherry, D. (2011). Thermoscopes, thermometers, and the foundations of measurement. *Studies in History and Philosophy of Science Part A*, *42*(4), 509–524.

Shiffman, S., & Stone, A. A. (1998). Ecological momentary assessment: A new tool for behavioral medicine research. In D. Krantz & A. Baum (Eds.), *Technology and methods in behavioral medicine* (pp. 117–131). Mahwah, NJ: Lawrence Erlbaum Associates.

Shumway, R. H., & Stoffer, D. S. (2010). *Time series analysis and its applications: With R examples*. New York, NY: Springer.

Sijtsma, K. (2012). Psychological measurement between physics and statistics. *Theory and Psychology*, *22*, 768–809.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, *48*(1), 1–48.

Smith, C. A., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology*, *48*(4), 813–838.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., ... Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, *23*, S208–S219.

Snijders, T., & Bosker, R. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage Publications.

Song, H., & Ferrer, E. (2012). Bayesian estimation of random coefficient dynamic factor models. *Multivariate Behavioral Research*, *47*(1), 26–60.

Spaulding, S. (2013). Mirror neurons and social cognition. *Mind and Language*, *28*(2), 233–257.

Sporns, O. (2011). *Networks of the brain*. Cambridge, MA: MIT press.

Steele, J. S., Ferrer, E., & Nesselroade, J. R. (2013). An idiographic approach to estimating models of dyadic interactions with differential equations. *Psychometrika*, 1–26.

Steer, R. A., Ball, R., Ranieri, W. F., & Beck, A. (1999). Dimensions of the Beck Depression Inventory-II in clinically depressed outpatients. *Journal of Clinical Psychology*, *55*(1), 117–128.

Stegenga, J. (2012). Rerum concordia discors: Robustness and discordant multimodal evidence. In L. Soler, E. Trizio, T. Nickles, & W. C. Wimsatt (Eds.), *Characterizing the robustness of science: After the practice turn in the philosophy of science* (pp. 207–226). Dordrecht: Springer.

Stewart, J. G., & Harkness, K. L. (2012). Symptom specificity in the acute treatment of major depressive disorder: A re-analysis of the treatment of depression collaborative research program. *Journal of Affective Disorders*, *137*(1), 87–97.

Stone, A. A., & Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavorial medicine. *Annals of Behavioral Medicine*, *16*, 199–202.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*(1), 59–71.

Stroe-Kunold, E., Wesche, D., Friederich, H.-C., Herzog, W., Zastrow, A., & Wild, B. (2012). Temporal relationships of emotional avoidance in a patient with anorexia nervosa: A time series analysis. *The International Journal of Psychiatry in Medicine*, *44*(1), 53–62.

Sullivan, K. J., Shadish, W. R., & Steiner, P. M. (2015). An introduction to modeling longitudinal data with generalized additive models: Applications to single-case designs. *Psychological Methods*, *20*(1), 26–42.

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.

Suls, J., & Martin, R. (2005). The daily life of the garden-variety neurotic: Reactivity, stressor exposure, mood spillover, and maladaptive coping. *Journal of Personality*, *73*(6), 1485–1510.

Tafforin, C. (2013). The Mars-500 crew in daily life activities: An ethological study. *Acta Astronautica*, *91*, 69–76.

Tal, E. (2011). How accurate is the standard second? *Philosophy of Science*, *78*(5), 1082–1096.

Tal, E. (2013). Old and new problems in philosophy of measurement. *Philosophy Compass*, *8*(12), 1159–1173.

Tan, X., Shiyko, M. P., Li, R., Li, Y., & Dierker, L. (2012). A time-varying effect model for intensive longitudinal data. *Psychological Methods*, *17*(1), 61–77.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, *19*(1), 109–139.

Tarvainen, M. P., Georgiadis, S. D., Ranta-aho, P. O., & Karjalainen, P. A. (2006). Time-varying analysis of heart rate variability signals with a Kalman smoother algorithm. *Physiological Measurement*, *27*(3), 225–239.

Tarvainen, M. P., Hiltunen, J. K., Ranta-aho, P. O., & Karjalainen, P. A. (2004). Estimation of nonstationary EEG with Kalman smoother approach: An application to event-related synchronization. *IEEE Transactions on Biomedical Engineering*, *51*(3), 516–524.

Tatsuoka, K. K. (1987). Validation of cognitive sensitivity for item response curves. *Journal of Educational Measurement*, *24*(3), 233–245.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Erlbaum.

Telesford, Q. K., Simpson, S. L., Burdette, J. H., Hayasaka, S., & Laurienti, P. J. (2011). The brain as a complex system: Using network science as a tool for understanding the brain. *Brain Connectivity*, *1*(4), 295–308.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*, 267–288.

Tio, P., Epskamp, S., Noordhof, A., & Borsboom, D. (in press). Mapping the manuals of madness: Comparing the ICD-10 and DSM-IV-TR using a network approach. *International Journal of Methods in Psychiatric Research*.

Tournier, J.-D., Mori, S., & Leemans, A. (2011). Diffusion tensor imaging and beyond. *Magnetic Resonance in Medicine*, *65*(6), 1532–1556.

Travers, J., & Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, *32*, 425–443.

Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory and Psychology*, *19*(5), 579–599.

Trout, J. D. (1998). *Measuring the intentional world*. Oxford: Oxford University Press.

Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology*, *9*, 151–176.

Tschacher, W., & Ramseyer, F. (2009). Modeling psychotherapy process by time-series panel analysis. *Psychotherapy Research*, *19*(4-5), 469–481.

Tschacher, W., Zorn, P., & Ramseyer, F. (2012). Change mechanisms of schema-centered group psychotherapy with personality disorder patients. *PloS One*, *7*(6), e39687.

Tuerlinckx, F., Rijmen, F., Verbeke, G., & Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, *59*(2), 225–255.

Urbach, T. P., & Kutas, M. (2002). The intractability of scaling scalp distributions to infer neuroelectric sources. *Psychophysiology*, *39*(6), 791–808.

van Borkulo, C. D., Borsboom, D., & Schoevers, R. A. (2016). Group-level symptom networks in depression: Reply. *JAMA Psychiatry*, *73*(4), 411–412.

van Borkulo, C. D., Boschloo, L., Borsboom, D., Penninx, B. W., Waldorp, L. J., & Schoevers, R. A. (2015). Association of symptom network structure with the course of longitudinal depression. *JAMA Psychiatry*, *72*(12), 1219–1226.

van de Leemput, I. A., Wichers, M., Cramer, A., Borsboom, D., Tuerlinckx, F., Kuppens, P., ... Scheffer, M. (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 den Heuvel, M. P., & Pol, H. E. H. (2010). Exploring the brain network: A review on resting-state fMRI functional connectivity. *European Neuropsychopharmacology*, *20*(8), 519–534.

van der Does, A. (2002). *Manual of the Dutch Version of the Beck Depression Inventory (BDI-II-NL).* Harcourt Publishers: Amsterdam.

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.

Vansteelandt, K., van Mechelen, I., & Nezlek, J. B. (2005). The co-occurrence of emotions in daily life: A multilevel approach. *Journal of Research in Personality*, *39*(3), 325–335.

van Steen, M. (2010). *Graph theory and complex networks: An introduction.* Amsterdam: van Steen M.

Velicer, W. F., & Fava, J. L. (2003). Time series analysis. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology, vol. 2* (pp. 581–606). Hoboken, NJ: Wiley.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data.* New York: Springer.

Verbeke, G., Spiessens, B., & Lesaffre, E. (2001). Conditional linear mixed models. *The American Statistician*, *55*(1), 25–34.

Vigo, D. E., Tuerlinckx, F., Ogrinz, B., Wan, L., Simonelli, G., Bersenev, E., ... Aubert, A. E. (2013). Circadian rhythm of autonomic cardiovascular control during Mars500 simulated mission to Mars. *Aviation, Space, and Environmental Medicine*, *84*(10), 1023–1028.

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*(2), 176–192.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin and Review*, *11*(1), 192–196.

Waldorp, L., Christoffels, I., & van de Ven, V. (2011). Effective connectivity of fMRI data using ancestral graph theory: Dealing with missing regions. *NeuroImage*, *54*(4), 2695–2705.

Walls, T. A., & Schafer, J. L. (2006). *Models for intensive longitudinal data*. Oxford, England: Oxford University Press.

Wang, L. P., Hamaker, E. L., & Bergeman, C. S. (2012). Investigating inter-individual differences in short-term intra-individual variability. *Psychological Methods*, *17*(4), 567–581.

Wang, Y., Jing, X., Lv, K., Wu, B., Bai, Y., Luo, Y., ... Li, Y. (2014). During the long way to mars: Effects of 520 days of confinement (Mars500) on the assessment of affective stimuli and stage alteration in mood and plasma hormone levels. *PloS One*, *9*(4), e87087.

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin*, *98*(2), 219–235.

Watson, D., Wiese, D., Vaidya, J., & Tellegen, A. (1999). The two general activation systems of affect: Structural findings, evolutionary considerations, and psychobiological evidence. *Journal of Personality and Social Psychology*, *76*(5), 820–838.

Watson, J. D., Myers, R., Frackowiak, R. S., Hajnal, J. V., Woods, R. P., Mazziotta, J. C., ... Zeki, S. (1993). Area V5 of the human brain: Evidence from a combined study using positron emission tomography and magnetic resonance imaging. *Cerebral Cortex*, *3*(2), 79–94.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of small-worldnetworks. *Nature*, *393*(6684), 440–442.

West, M., Prado, R., & Krystal, A. D. (1999). Evaluation and comparison of EEG traces: Latent structure in nonstationary time series. *Journal of the American Statistical Association*, *94*(446), 375–387.

WHO. (2001). *The world health report 2001: Mental health: new understanding, new hope.* World Health Organization: Geneva.

WHO. (2008). *ICD-10: International Statistical Classification of Diseases and Related Health Problems, 10th Revised edition.* New York, NY: Author.

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., Groot, P. C., & Psychosystems. (2016). Critical slowing down as a personalized early warning signal for depression. *Psychotherapy and Psychosomatics*, *85*(2), 114–116.

Wichers, M., Wigman, J., & Myin-Germeys, I. (in press). Micro-level affect dynamics in psychopathol-

ogy viewed from complex dynamical system theory. *Emotion Review*.

Wigman, J., van Os, J., Borsboom, D., Wardenaar, K., Epskamp, S., Klippel, A., ... Wichers, M. (2015). Exploring the underlying structure of mental disorders: Cross-diagnostic differences and similarities from a network perspective using both a top-down and a bottom-up approach. *Psychological Medicine*, *45*(11), 2375–2387.

Wild, B., Eichler, M., Friederich, H.-C., Hartmann, M., Zipfel, S., & Herzog, W. (2010). A graphical vector autoregressive modelling approach to the analysis of electronic diary data. *BMC medical research methodology*, *10*(1), 1.

Wimsatt, W. C. (1981). Robustness, reliability, and overdetermination. In M. Brewer & B. Collins (Eds.), *Scientific inquiry and the social sciences* (pp. 124–163). San Francisco, CA: Jossey-Bass.

Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Cambridge, MA: Harvard University Press.

Wood, S. N. (2006). *Generalized additive models: An introduction with R*. Boca Raton, FL: Chapman and Hall/CRC.

Wood, S. N. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika*, *100*(1), 221–228.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.

Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of fMRI data. *NeuroImage*, *14*(6), 1370–1386.

Zelenski, J. M., & Larsen, R. (2000). The distribution of basic emotions in everyday life: A state and trait perspective from experience sampling data. *Journal of Research in Personality*, *34*(2), 178–197.

Zhang, D., Snyder, A. Z., Fox, M. D., Sansbury, M. W., Shimony, J. S., & Raichle, M. E. (2008). Intrinsic functional relations between human cerebral cortex and thalamus. *Journal of Neurophysiology*, *100*(4), 1740–1748.