Uncovering change in emotion regulation in clinical practice with the threshold autoregressive model

Author: Jens de Groot

Research group: LaBlab

Supervisors: Jocelien Olivier and Laura Bringmann

Co-supervisor: Sebastian Castro-Alvarez

Date: June 2022

Abstract

The threshold autoregressive (TAR) model is a model that describes state-dependent changes in regulatory behaviour of a system. Recently, the TAR model has been used by psychologists to study its potential in measuring state-dependent affect regulation over multiple timepoints across multiple individuals. This research is a follow up in which we study how many timepoints are sufficient for the model to make sound inferences about possible state-dependent regulation of affect for a single person. We make use of simulation trials to assess the capability of the TAR model to converge and produce good estimates for each model parameter while we vary the number of timepoints and some regulation parameters of the TAR model. We further apply a real dataset to the model in order to illustrate how to estimate the model and interpret the results. The results are discussed in light of clinical practice.

Contents

An important aspect of human behaviour that is studied by psychologists is the dynamics of human affect over time. There are at least two models that have been used by researchers to analyze affect dynamics. One model is the autoregressive (AR) model (Brose, Schmiedek, Koval, & Kuppens, 2015; Koval et al., 2015; Koval, Butler, Hollenstein, Lanteigne, & Kuppens, 2015; Koval, Kuppens, Allen, & Sheeber, 2012; Koval, Pe, Meers, & Kuppens, 2013; Koval & Kuppens, 2012; Kuppens, Allen, & Sheeber, 2010; Suls, Green, & Hillis, 1998). For example, Suls et al. (1998) used the AR model to measure *emotional inertia* in healthy individuals. Here, *emotional inertia* is the resistance of an individual to change their emotional state over time. Therefore, emotional inertia represents the autoregressive effect in the model. For example, a high autoregressive effect (i.e., high emotional inertia) indicates that an individual has more consecutive scores on an affect scale that lay close to each other. In other words, the individual lingers for longer into an emotional state. The opposite goes for a low autoregressive effect. Thus, the autoregressive effect or *inertia* can be seen as a measure of affect regulation that can differ between persons (Suls et al., 1998; Kuppens et al., 2010, 2012). Typically, a higher autoregressive effect indicates a weaker regulation of affect. (de Haan-Rietdijk, Gottman, Bergeman, & Hamaker, 2016).

The AR model has been proven successful in analysing and understanding of the dynamics of human affect. However, the AR model assumes that affect regulation is constant over time. This is a problem, as research shows that affect regulation may change with affect intensity (de Haan-Rietdijk, Gottman et al., 2016). Therefore, the threshold-autoregressive (TAR) model has been used recently (de Haan-Rietdijk, Gottman et al., 2016). Namely, the TAR model allows a state-dependent regulation of affect. This means that change in affect regulation is possible. In this way, the interpretation of time series data can become more valuable, as clinicians could track down changes in affect regulation.

Consequently, in a simulation study, de Haan-Rietdijk, Gottman et al. (2016) assessed how well a true TAR process – that was generated with simulated data – could be detected with the Bayesian estimation method. This ability with which an estimation technique detects a process that is truly there is defined as the *detection rate* or *power.* It could be observed that the power increased with an increasing number of timepoints and subjects. In the following sections, de Haan-Rietdijk, Gottman et al. (2016) used real datasets with a sufficient number of timepoints and subjects. For example, they found that individuals had a lower inertia for more intense negative mood states (de Haan-Rietdijk, Gottman et al., 2016).

However, what is missing is that the TAR model for a single person has not been investigated into detail. For instance, it is yet precarious how many timepoints are sufficient for the TAR model to converge for a single person. This is very important to know for several reasons. Firstly, clinicians often evaluate only one person at a time. Secondly, time series measurements on affect regulation are time consuming and it may take days, weeks or even months to complete those measurements (Kossakowski, Groot, Haslbeck, Borsboom, & Wichers 2017; Koval et al., 2012; Suls et al., 1998). It is therefore useful to know what the smallest number of timepoints is at which the TAR model functions well enough to make sound inferences about one's affect dynamics.

This study is built up as follows. Firstly, we provide background literature of research on both the AR and the TAR model. Next, we describe the basic AR and TAR model. After that, we determine the proper sampling conditions for our simulation study. Thereafter, we evaluate model convergence criteria and the accuracy of TAR model estimates with a simulation study, while we vary the number of measurements (timepoints) for a single person. We do this in order to get an answer to the question of how many timepoints we need to let the TAR model converge and function well for a single person. Lastly, an empirical dataset of a single person on affect regulation is used to show how the model

would work in a real-life example. To achieve all this, we use Bayesian inference methods throughout the entire study. Lastly, we discuss the results in light of the existing literature.

1 - Background

In the AR model, each single observation can be regressed on one or more previous observations. The autoregressive coefficients determine to what extent an observation is predicted by its previous observations. Let us now focus on a system in which an observation only depends on the previous observation. In this case, we have one autoregressive coefficient that links the two observations together. To begin with, the system has an equilibrium, which is the state that the system tends to be near to. The autoregressive coefficient can take values between -1 and 1. For example, a system with a high autoregressive coefficient takes relatively long to recover back to equilibrium. The opposite goes for a system with a low autoregressive coefficient. The autoregressive coefficient can therefore be seen as a measure of resistance of a system to recover after an external shock that perturbs the system away from its equilibrium. In the field of affect dynamics, this is also called *emotional inertia*. For example, Suls and colleagues (1998) found that a bad mood at a current timepoint was significantly predicted by a bad mood at the previous timepoint with inertia $\varphi = 0.25$ (Suls et al., 1998). Bearing this in mind, Kuppens and others (2010) found statistically significant *inertias* for eight emotions ranging from 0.20 to 0.33 while sampling emotional states from daily life of students (Kuppens et al., 2010). In another study, inertias for both positive and negative emotions ranged from 0.15 to 0.40 in a sample of some 300 students (Koval, Sütterlin, & Kuppens, 2016). This indicates a low to moderate predictability of emotional states. In other words, the emotional states of some of these individuals tend to recover rather quickly to the equilibrium.

However, a limitation of the AR model is that it assumes that affect regulation is constant over time. This means that inertia is presumed stationary and does not vary over time. To overcome this limitation, the AR model can be extended to the so-called TAR model. In this model, affect regulation is allowed to vary over time. This is realized by adding a so-called *threshold* parameter to the TAR model, hence this is called the *threshold*-AR or TAR model. These thresholds actually divide the AR model into multiple individual AR processes or regimes that each have their own autoregressive coefficients (Tsay, 1989). In the simplest case we have a TAR model with only one threshold, thus two *regimes* with their own emotional inertias. As a consequence of this, emotional inertia can change as the intensity of affect changes. To concretize this idea, we shortly discuss some results of the study of de Haan-Rietdijk et al. (2016). In this study, they focussed on the multilevel extension of the TAR model. They used a dataset that describes the affect dynamics of 129 newlywed couples during a 15 minutes lasting discussion (Gottman, Swanson, & Murray, 1999; de Haan-Rietdijk, Gottman et al., 2016) They used the model to estimate the two inertias for each person. The result was that the data followed a TAR process, with an inertia of 0.3 and another inertia of around 0.6 for both men and women. The first inertia was obtained for an affect intensity was below a certain threshold, whereas the second inertia was obtained when their affect intensity was above this threshold (de Haan-Rietdijk, Gottman, et al., 2016).

2 – Methods

In this section, we discuss the mathematical properties of both the AR and TAR model for a single case and Bayesian inference. The chapter is built up as follows. We first discuss the generalized version of the AR model. Next, we discuss a special case of the AR model with one regression coefficient and graphically illustrate how this version of the model is used in the field of affect dynamics. We further

present an alternative parameterization of the AR model that is used in an article of de Haan-Rietdijk, Gottman et al. (2016). The sections for the TAR model are built up the same; we explain a generalized version, then focus on a special case with two regimes and finally discuss its alternative equivalent (de Haan-Rietdijk, Gottman, et al., 2016). Finally, we discuss Bayesian inference, which we used to estimate the TAR model for a single person.

2.1 – The AR model

2.1.1 – The AR model in general

The autoregressive (AR) model is described by one or multiple regression coefficients that link one or multiple previous observations to the current observation at a certain timepoint (Figueiredo, Figueiras, Park, Farrar, & Worden, 2011). In this way, a variable on a timepoint that depends on itself at lagged timepoints can be modelled. In general, let p be the number of previous timepoints that can be regressed on the current timepoint. In that case, we have an autoregressive process of order p, denoted as $\mathrm{AR}(p)$. The $\mathrm{AR}(p)$ model can be represented as follows:

$$Y_t = c + \sum_{i=1}^{p} \varphi_i Y_{t-i} + \varepsilon_t \ (1),$$

where c is a constant, which can be interpreted as the intercept. Y_{t-i} are the previous observations. The parameters φ_i are the autoregressive coefficients that determine the extent with which the current observation is linked to the previous observations. Lastly, ε_t represents random shocks that happen in daily life and is also called 'white noise'. It is assumed that the random shocks are normally distributed around a mean of 0, with a variance specific for these random shocks. Thus,

$$\varepsilon_t \sim Normal(0, \sigma_\varepsilon^2) \quad (2)$$

Furthermore, the parameters can be estimated using several methods (e.g., Box, Jenkins, Reinsel, & Ljung, 2015). One method that is used and will be elaborated on later is Bayesian Inference.

2.1.2 – The AR(1) model and affect regulation

In the previous section, we wrote about the generalized AR model. However, one special case of the AR model that lends itself very well to study affect dynamics in psychology is the AR(1) model (Suls et al., 1998; Kuppens et al., 2010; Koval & Kuppens, 2012). The $\mathrm{AR}(1)$ model for a single person can be defined as follows:

$$Y_t = c + \varphi Y_{t-1} + \varepsilon_t \ (3)$$

As for the general version of the model, c is the intercept, φ is the autocorrelation parameter and ε_t is the random shock, or innovation parameter. In the article of de Haan-Rietdijk, Gottman et al. (2016), the $\mathrm{AR}(1)$ model is used to simulate data, they make use of an alternative parameterization of the model, because the intercept is not meaningful, but having a mean or equilibrium is interesting to interpret. In this article, we work with this version of the model. In this way, the $\mathrm{AR}(1)$ model becomes:

$$Y_t = \mu + \varphi(Y_{t-1} - \mu) + \varepsilon_t \ (4)$$

Without this reparameterization, the intercept c would be equal to

$$c = \mu(1 - \varphi) \quad (5)$$

As already mentioned, this intercept value has no intuitive interpretation. Therefore, de Haan-Rietdijk et al. (2016) and we prefer to use Equation 3.

Now, the dynamics of the AR(1) can be visualized in Figure 1. Both panels represent the NA score for two hypothetical individuals A and B (de Haan-Rietdijk, Gottman, et al., 2016). The horizontal lines in the plot represent the *equilibrium* a person tends toward. For example, for both individuals the equilibrium is equal to 15. However, their inertias may differ. For instance, for individual A (upper panel) with an inertia of $\varphi = 0.2$, the affect score at a certain timepoint has little influence on the affect score at the next timepoint. Thus, for this individual the affect quickly recovers towards their equilibrium. However, individual B has an inertia of 0.8, so that there is a slower recovery towards the equilibrium, with consecutive scores below or above the equilibrium. Corresponding state-space plots are seen on the right of the figure, in which the underlying autoregression is depicted (de Haan-Rietdijk, Gottman, et al., 2016). Also, note that the mean over the timeseries is equal to the equilibrium for both person A and B.

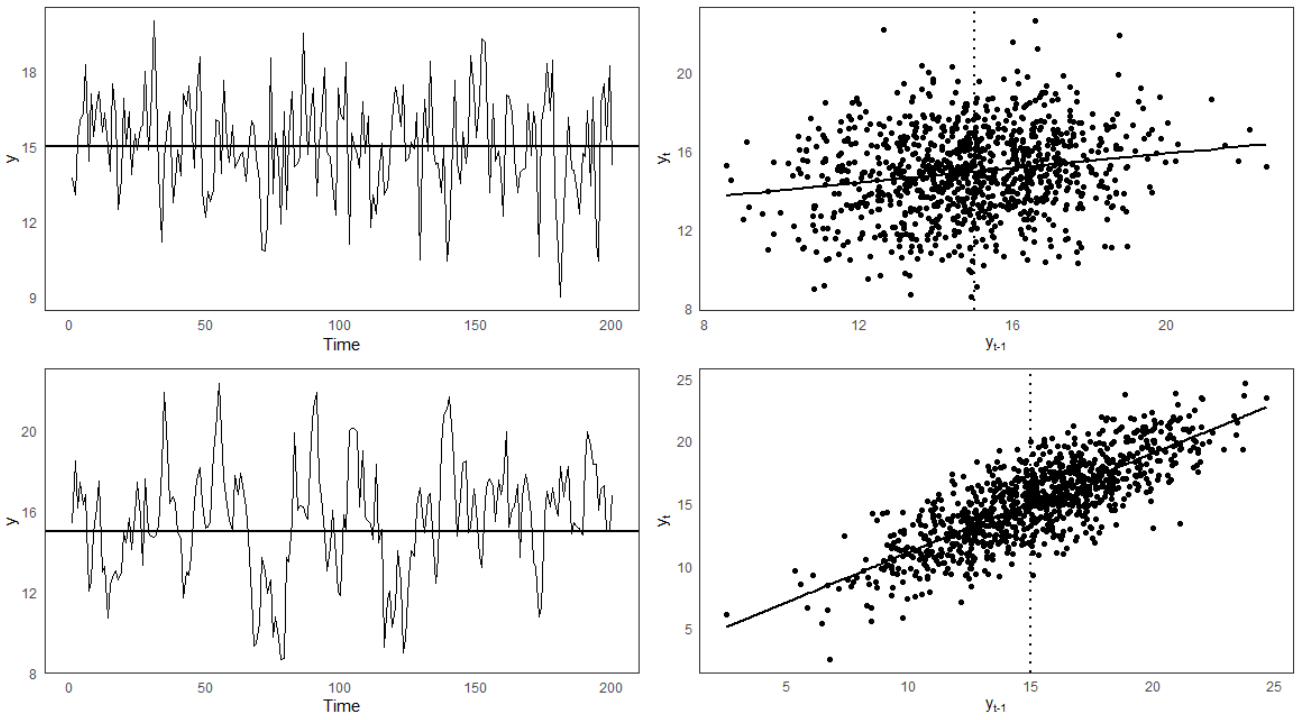


Figure 1. Hypothetical NA scores for persons A and B with state-space plots representing the underlying autoregression. The vertical dashed line in these plots represents the equilibrium. The random shock or innovation parameter comes from a standard normal distribution with variance 4; $\varepsilon_t \sim Normal(0,4)$. On the right plots, simulated data for 1000 timepoints is included. On the left plot, only the first 200 timepoints are shown.

Thus, Figure 1 represents two cases that have each a different inertia. Notice that the regression lines in each of the state-space plots of the figure have the same slope, irrespective of which side of the equilibrium (dotted line) the points are. Actually, the state-space plots imply that inertia is constant over time and does not vary within persons. However, recent research shows that inertia may actually covary with intensity of affect level (de Haan-Rietdijk, Kuppens, & Hamaker, 2016). In this study, two datasets were used that contained measurements of multiple persons on affect intensity and inertia. It was found that there was a substantial correlation between affect intensity and the inertia for that affect (de Haan-Rietdijk, Kuppens et al., 2016). Thus, despite the fact that this study included multiple persons, it lends support for the hypothesis that inertia may also vary with affect intensity within a system or person. The threshold autoregressive (TAR) model offers a solution to this.

2.2 – The TAR model

2.2.1 – The TAR model in general

The general TAR model is an extension of the AR model in which it is assumed that the strength of the autoregressive effect depends on the observed values of the previous occasions. As a consequence, there are multiple processes that describe the time series of interest and each of these processes is known as a regime. Let us now introduce the TAR(k; p, d) model (Tong & Lim, 1980; Tsay, 1989). Here, k is the number of regimes separated by $k-1$ thresholds. Again, p is the order of the autoregression. Although not applicable to this article, it should be noted that p may differ between regimes (Tsay,

1989). Lastly, d is the lag of the threshold and determines which timepoint before the current timepoint is used as the threshold variable. Thus, the model is defined mathematically as follows:

$$Y_t = \varphi_0^{(j)} + \sum_{i=1}^{p} \varphi_i^{(j)} Y_{t-i} + a_t^{(j)} \quad \text{with } r_{j-1} \leq Y_{t-d} < r_j \text{ (6)},$$

where j is the index for the regime the model is in, with $j = 1, \ldots, k$. In each regime, there exists a particular AR model that is effective in only one regime. Furthermore, r is a threshold value that separates two regimes from each other. Which regime is active depends on what the value of the threshold variable is. Also, Y_{t-d} is always enclosed by the two threshold values of the $(j-1)^{th}$ and j^{th} regime, thus $r_{j-1} \leq Y_{t-d} < r_j$.

To make this more concrete, the TAR(2; 1, 1) model has two regimes, with $r_0 = -\infty, r_1$ and $r_2 = \infty$. In this case we have the two states in which $-\infty < Y_{t-d} < r_j$ and $r_j \leq Y_{t-d} < \infty$. More practically, for the lower regime we can write that $Y_{t-1} < r_1$ and for the upper regime $Y_{t-1} \geq r_1$. For example, in the study of de Haan-Rietdijk, Gottman et al. (2016), they make use of the $\mathrm{TAR}(2; 1,1)$ model, with which they tried to model the same concept as mentioned before; affect regulation. To clarify, they use a two-regime model of lag-order 1 separated by one threshold parameter. A lag-order 1 indicates that only the previous timepoint is used for the threshold variable (i.e. $d = 1$). This results in the following model for a single person:

$$Y_t = \begin{cases} \tau + \varphi_1(Y_{t-1} - \tau) + \varepsilon_t & \text{if} \quad Y_{t-1} < \tau \\ \tau + \varphi_2(Y_{t-1} - \tau) + \varepsilon_t & \text{if} \quad Y_{t-1} \geq \tau \end{cases} \quad (7),$$

where τ is the threshold parameter and intercept at the same time. However, in a TAR process, the parameter τ need not be equal to the equilibrium or mean. For example, if the autocorrelation above the threshold is higher, then scores above the threshold may be more persistent and in greater number than scores below the threshold. In this case, the mean would be higher than the threshold (de Haan-Rietdijk, Gottman et al., 2016). Again, φ_1 and φ_2 are the inertias that are dependent on what the value of Y_{t-1} is relative to the threshold parameter τ.

To illustrate the TAR model with two regimes, of lag-order 1 and with a threshold parameter that has lag 1, we present an example based on de Haan-Rietdijk, Gottman et al (2016). Consider that the affect dynamics of a person follows a TAR process and their time series are presented on the left panel of Figure 2 Here, the two different inertias are $\varphi_1 = 0.1$ and $\varphi_2 = 0.7$. The two regimes are separated by the threshold value of $\tau = 15$. The two states of affect are each a regime in which another inertia value is in effect. Thus, when the NA value of this person is below 15 and thus less intense, their inertia is relatively low (0.1). However, if this person has an affect of above 15, then their inertia is 0.7. This means that a more intense NA has a larger carry-over effect than an affect that is less intense and that inertia is thus state-dependent, as can be seen in Figure 2 (de Haan-Rietdijk, Gottman, et al., 2016).

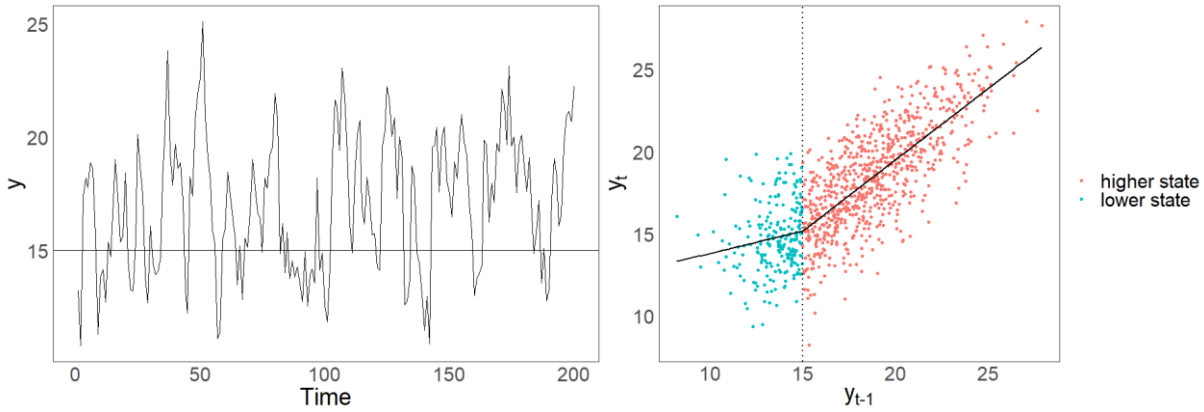


Figure 2. An instance of the $\mathrm{TAR}(2;1,1)$ model. Person C has more consecutive scores above the equilibrium of 15 due to the larger inertia of 0.7. Scores below the equilibrium are scanter because of quicker recovery to the equilibrium caused by the low inertia of 0.1 in the state of lower NA. The state-space plot on the right is partitioned into two regimes by the vertical line representing the equilibrium. Of note is the flexure in the state-space plot that is caused by two differing inertias in each *regime*.

2.3 – Bayesian inference

2.3.1 – In theory

In this article, we implemented the TAR(2; 1, 1) model within the Bayesian framework. Hence, in this section we present a brief overview of Bayesian inference. Bayesian inference is a statistical method that combines prior beliefs with the observations about a certain phenomenon to do statistical inferences about the data. It is based on Bayes' rule. Concretely, the idea is that the knowledge or belief (prior) that we have about a phenomenon can be updated after observing the evidence (data). Consistently, the updated knowledge is the posterior, which in turn can become the prior in yet another following experiment. Suppose that we have some data and we would like to estimate parameter θ. Then Bayes' rule is defined as follows:

$$\Pr(\theta|data) = \frac{\Pr(data|\theta) \times \Pr(\theta)}{\Pr(data)} \quad (8)$$

Here, $\Pr(\theta|data)$ is the posterior distribution of parameter values. That is, it is the distribution of parameter values that is observed after the data are seen. This distribution is determined by the product of the prior distribution $\Pr(\theta)$ and the likelihood function $\Pr(data|\theta)$. The likelihood function is the likelihood of observing the data given the parameter value θ. This product of the prior distribution and the likelihood function is then divided by the probability of observing the data irrespective of parameter values (marginal probability), which is $\Pr(data)$. Thus, we could say that the posterior distribution is proportional to the product of the prior distribution and the likelihood function:

$$\Pr(\theta|data) \propto \Pr(data|\theta) \times \Pr(\theta) \quad (9)$$

It should be noted that a particular group of priors may not even influence the shape of the posterior distribution. Such priors are called 'flat priors'. Flat priors are thus prior distributions that have a probability density function for the parameter value that is hardly influenced by this value itself. In our case we used flat priors, in order to get similar estimates as we would get if we have used frequentist approaches. In this study, flat priors are also used, so that the posterior distribution is more determined by the data, which makes the process of finding the posterior more data-driven.

2.3.2 – In practice

The most popular Bayesian estimation methods are the Markov Chain Monte Carlo (MCMC) algorithms. In short, these algorithms describe an iterative process that samples possible estimated values from an approximation of the posterior distribution without actually computing the posterior distribution. Additionally, it is possible to run a model multiple times with these algorithms. In each run or iteration, parameters of interest are estimated, given the input data. After many runs, there is a substantial sample of parameter estimates for each parameter of interest. This sample is also an approximation of the posterior distribution of parameter values (Kruschke, 2014). An MCMC algorithm should be run with multiple chains. Here, a chain is a sequence of a finite number of model runs. It is recommended to run two or more parallel chains at a time, in order to see if the model converges to a stable posterior distribution. It usually takes a couple of iterations for the chains to converge to the same region in the parameter space. This period, in which the chain converges to the highest density region of the parameter space from its initial value is called the *burn-in* period.

One of the problems of the MCMC algorithms is that there is never certainty that the model converged. Because of this, multiple visual and numerical quality checks of the samples are done. For instance, the entirety of values in the chain should represent the posterior distribution, or at least an approximation thereof. Also, the posterior distribution of chain values should not be influenced too much by the initial values for the parameters. Ideally, chains should be generated in a time-efficient manner, with as few iterations as possible (Kruschke, 2014). Furthermore, a chain of good quality is a chain that fluctuates well and is thus not stuck at one single location in the parameter space, which can be seen in a trace plot. Chains that fluctuate in a similar way within the posterior distribution are said to *mix* or *overlap* well with each other (Kruschke, 2014; Toft, Innocent, Gettinby, & Reid, 2007).

There exist also numerical quality checks of the chain. Two important instances of these are the *autocorrelation* and the Gelman-Rubin diagnostic (Gelman & Rubin, 1992). The autocorrelation is the correlation of the chain values with other chain values that are k steps ahead of the former. When there is a high autocorrelation in a chain at a certain lag k, it means that a chain moves slowly through the parameter space. This means that there is also less variation in chain values, which in turn leads to a lower variance in the posterior distribution. Therefore, the autocorrelation is to some extent predictive of the certainty with which the parameters are estimated (Link & Eaton, 2012). The advantage of a low autocorrelation is that sampling values are more independent from each other. In this way, draws from the posterior distribution cannot be predicted so well by previous draws. Thus, a low autocorrelation can be used such that we can obtain more useful information from the posterior distribution. Also, by reducing the autocorrelation between draws, we can increase the effective sample size (Lanfear, Hua, & Warren, 2016). One way to make the autocorrelation lower is by making use of thinning. Thinning is the process in which every i^{th} step in the chain is stored. The costs and benefits of thinning are shortly discussed. For example, one may use thinning in order to lower the autocorrelation, which benefits sampling independence at the cost of estimation accuracy and time

(Geyer, 1992; Link & Eaton, 2012; Maceachern & Berliner, 1994). On the other hand, not using thinning of chains benefits the accuracy with which parameters are estimated and shortens time consumption at the cost of sampling independence (Link & Eaton, 2012).

The other numerical quality check is known as the Gelman-Rubin statistic or diagnostic. This is a statistic that tells how much variance there is between chains relative to the amount of variance within chains (Gelman & Rubin, 1992). This ratio between between-variance and within-variance decreases and comes close to 1.0, once chains have all settled in the same parameter space with the same extent of fluctuation around the mode value of the parameter. Thus, a Gelman-Rubin statistic value close to 1.0 indicates that the model actually converges. In the end, the statistic converges downwards to 1.0 with more iterations (Brooks & Gelman, 1998; Gelman & Rubin, 1992; Kruschke, 2014). As a rule of thumb, the Gelman-Rubin diagnostic should preferably be below 1.1, because higher values indicate non-convergence (Kruschke, 2014).

3 – Simulations

We performed simulation studies to assess at what number of timepoints the TAR model becomes a feasible model to describe affect dynamics. To achieve this, we first determined the proper settings of some relevant MCMC parameters in a preliminary simulation. Thereafter, we used these settings to perform a simulation study in which we measured model performance under different conditions. Conditions could vary by number of timepoints and true value for the inertia parameters. We assessed the ability of the MCMC algorithm to estimate the parameter values from the simulated data and to detect the underlying TAR process of the simulated data.

3.1 – Choosing the MCMC design

There are MCMC-specific values that may need to be specified in order to obtain proper results. These are the total number of iterations, thinning and fraction of burn-in iterations. We prefer to choose an optimal thinning, such that high autocorrelation can be removed without losing an excessive amount of time. In our case, we use the TAR model in the Bayesian framework for clinical practice. Thus, what we want is that the accuracy with which the parameters are estimated only depends on the given data (and eventual prior beliefs), not on high autocorrelation (see Supplementary Materials and Results). Also, with a decent thinning level, we can make the autocorrelation lower for all model parameters, such that differences in estimation accuracy between model parameters are only explained by the data. Based on a short simulation design and the results thereof that are explained in detail in the Supplementary Materials and Results, we decided to use a thinning factor of 30, total number of iterations 20.000 and a burn-in fraction of 0.5 (Figure S1).

3.2 – Data generation

In the simulation study, we focused on studying the model convergence and accuracy with which the parameters were estimated. Several simulation conditions were realised. Conditions could differ by (1) the true value for the first inertia φ_1, (2) the true value for the second inertia φ_2, and (3) the number of timepoints. The values of φ_1 and φ_2 were varied such that

$$\{\varphi_1, \varphi_2\} = \{0.2, 0.4\}, \{0.2, 0.6\}, \{0.4, 0.6\}, \{0.4, 0.2\}, \{0.6, 0.2\}, \{0.6, 0.4\} \quad (10),$$

which is based on observations in spousal relationships in an earlier study (de Haan-Rietdijk, Gottman et al., 2016). The number of timepoints was varied as follows, based on the literature: 90, 150, 200, 300, 500 (Borkenau & Ostendorf, 1998; Ferrer, Steele, & Hsieh, 2012; Gottman et al., 1999; Hamaker, Grasman, & Kamphuis, 2016; Kuppens et al., 2010; Schuurman, Houtveen, & Hamaker, 2015; Tong & Lim, 1980). This resulted in a total of 30 conditions. Each condition was replicated 100 times, resulting in 3000 analyses. The seed numbers were simply denoted with the numbers 1 to 100. The results of the simulation are shown below and are discussed into more detail. The total duration of the simulation session lasted 4 days, 17 hours and 46 minutes on an Intel core i-9100T 3.10GHz with 8GB of RAM.

3.3 – Results

To start with, the plots that are shown are built up as follows. Each plot grid consists of six plots. In each plot, there is one combination of values for the two inertias for φ_1 (*phi1*) and φ_2 (*phi2*) that is effective. Furthermore, each point line in a plot represents values of a variable of inference that are plotted against the number of timepoints, for one model parameter. For instance, the upper left plot in Figure 5 presents the point estimates of the two inertias phi1 and phi2 against the number of timepoints when the true values of phi1 and phi2 were set at 0.4 and 0.2, respectively. Note that in each plot, the lines may be shifted towards a certain direction from each other. This is done to make sure that the lines do not overlap too much, such that the variation around each point is well visible. In the following sections, we analyse several aspects of the results. The model parameters that needed to be estimated include *phi1, phi2* (inertias), *phi_diff* (difference between inertias), *sigma2* (variance of the trait level of affect) and *tau* (τ, threshold parameter). However, in most cases we omitted the results for the parameters *phi_diff*, *sigma2* and *tau* to save space, but the visualization of these results can be regenerated using the script provided in the Supplementary Materials.

3.3.1 – Model convergence

Firstly, we discuss the results with regard to model convergence. To summarize, there were no indications that the TAR model had problems with converging. In all the simulations for any parameter, the median autocorrelation was 0.014 (IQR: 0.045; 0.003 – 0.048) (data not shown), which is acceptable for model convergence (Kruschke, 2014, p. 184). Furthermore, the maximum value of the Gelman-Rubin diagnostic across all of the simulation results was 1.039 (data not shown). This means that all of the Gelman-Rubin diagnostic values were below the acceptable threshold of 1.05.

3.3.2 – Parameter estimation

Secondly, we have the results that involve parameter estimation. These include the detection rate, point estimates, width of the credibility interval (CI), bias, absolute bias and relative bias. First, the detection rate is the fraction the hundred replications that contains the true value of the parameter in the 95% CI. The detection rate could thus range from 0 to 100%. The detection rate of the model parameters (i.e., phi1, phi2, sigma2, tau) was as expected. The smallest value observed was 0.9 and the largest value was 1. The mean detection rate was 0.96, which is around the expected value of 0.95. This was the detection rate for the four model parameters *phi1, phi2, sigma2* and *tau* (Figure 3).

For the difference between the two inertias, we actually computed the proportion of the replications in which the 95% CI of phi_diff did not include zero. In other words, it is the rate with which the underlying TAR process is detected. Thus, it should be noted that the detection rate has a different

meaning for this parameter. It turned out that the mean and median detection rate were 0.23 and 0.19 across all of the conditions, respectively. Note that, however, the detection rate increases with an increasing number of timepoints (see Figure 4). This suggests an improvement of the model to detect a TAR process as the number of timepoints increases.

The point estimates per simulation for the two inertias are represented as means of the posterior distribution. In Figure 5, the violin plots show the distribution of all of the posterior means from the hundred replicates. For example, one pattern that we can see is that the distribution of point estimates becomes narrower as the number of timepoints and inertia value increases. Also, from Figure 6 we can infer that the width of the CI is indeed systematically smaller when the value for either inertia is greater. Moreover, the values of the point estimates of these inertia parameters depend relatively more on the number of timepoints that is used than the other model parameters (see also information about the relative bias in Figure 7). Especially the width of the CI and the relative bias are of importance, as they are a measure of accuracy with which the model parameters are estimated. It is also very important to maintain a high accuracy for parameter estimation, because then conclusions can be drawn with more reliability. We therefore decided to do an analysis of variance of these parameter values between the timepoint number groups.

Furthermore, for *sigma2*, the mean point estimate was 4.01, which clearly approximates the true value of 4 (95% CI: [3.21, 4.96]), with a median relative bias very close to zero, measured across all the simulation conditions for all the model parameters. The same is true for parameter *tau*, (mean: 15.02, 95% CI: [14.22, 15.88]) with a relative bias very close to zero. Also, the estimates of these parameters became more certain as the number of timepoints increases (data not shown).

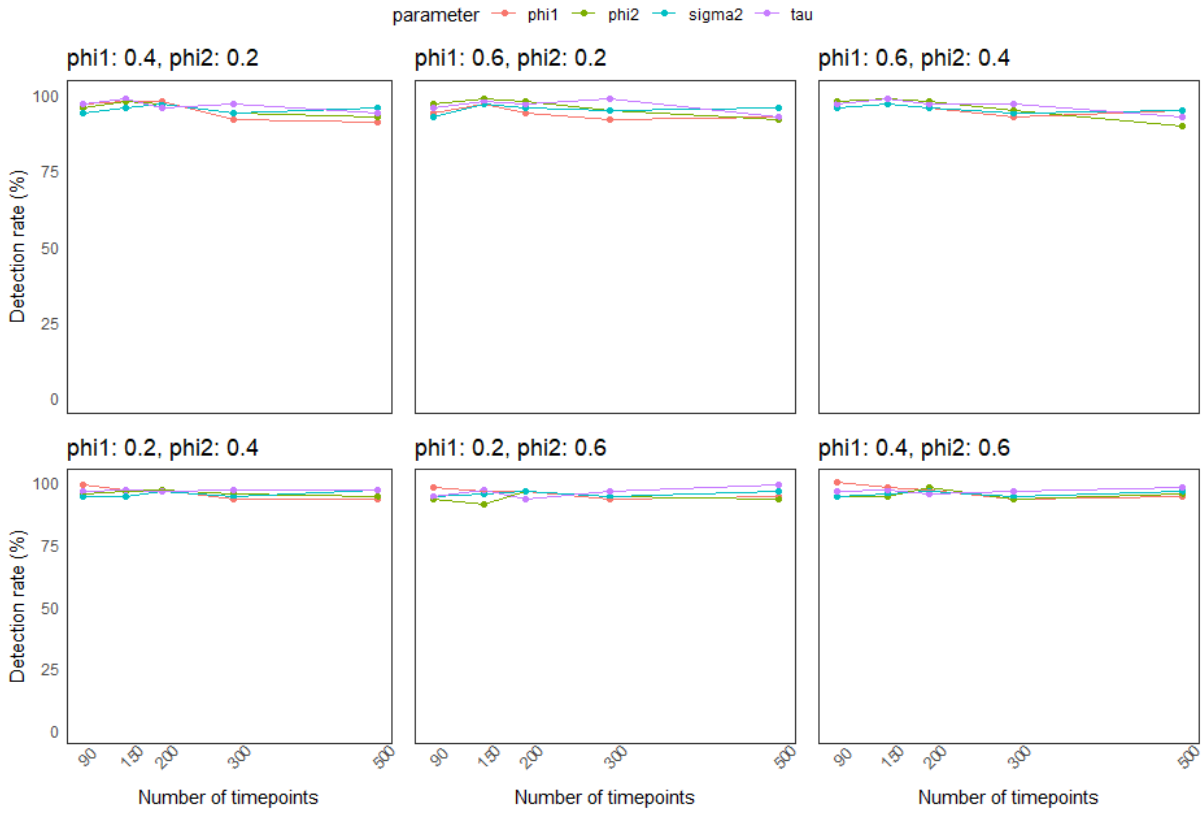


Figure 3. The detection rate of the true value of the model parameter in the simulations for the parameters *phi1, phi2* (inertias)*, sigma2* (variance of the trait level affect) and *tau* (threshold parameter). The detection rate is the percentage of seeds of which the CI contained the true value of the parameter. Each set of 100 seeds represented one timepoint, one model parameter and one set of true values for *phi1* and *phi2*.

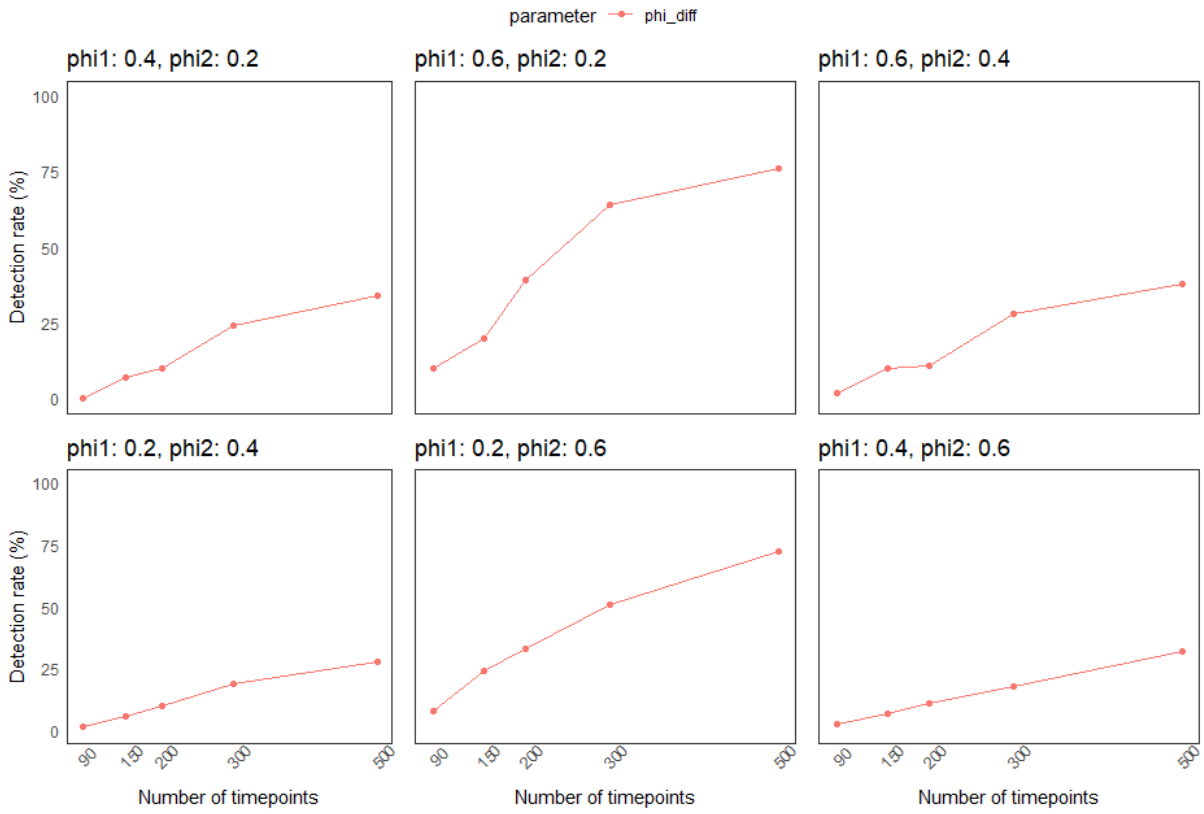


Figure 4. The detection rate of the of the TAR process. The detection rate is the percentage of seeds of which the CI of the difference parameter did not contain the value 0. Each set of 100 seeds represents one timepoint and one set of true values for *phi1* and *phi2*.

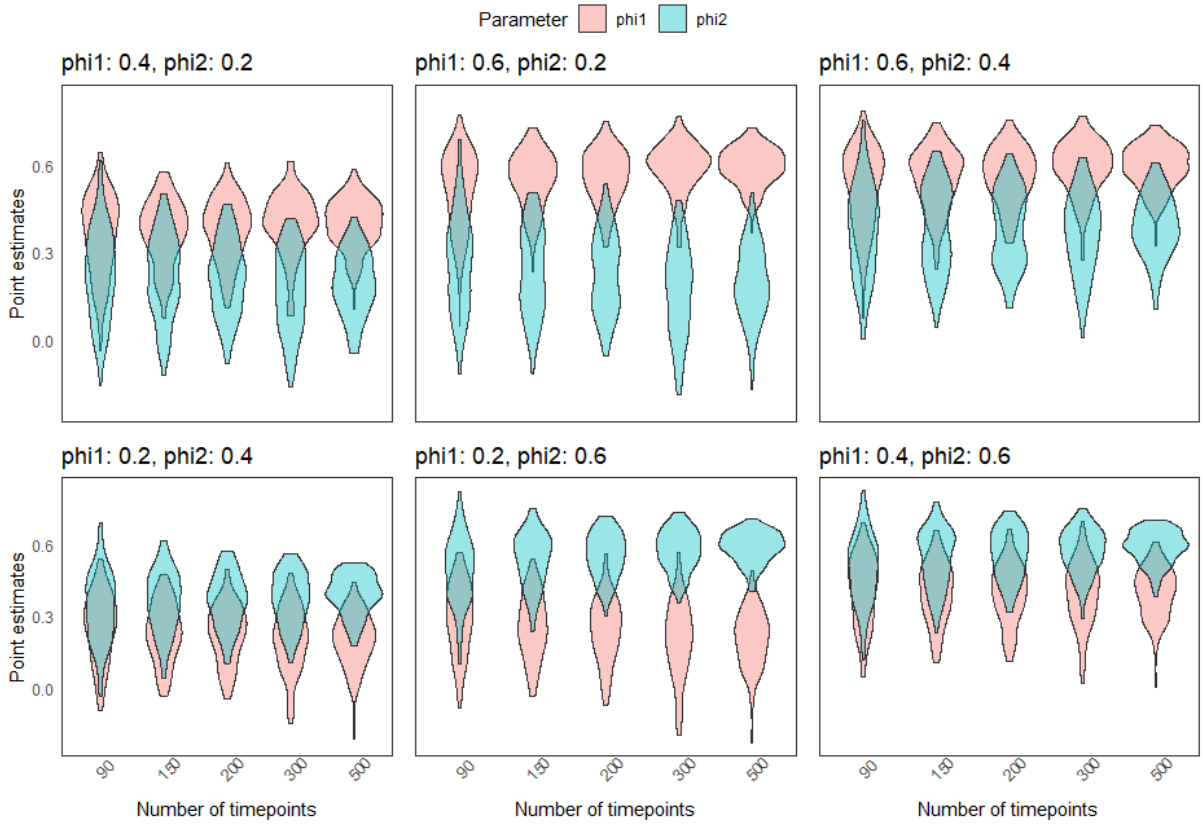


Figure 5. The point estimates of the two inertias (*phi1* and *phi2*) plotted against the number of timepoints for the six possible combinations of true values for the two inertias. A violin shape in each plot for each group of timepoint number represents the density of the dependent variable over the 95% confidence interval. A greater width indicates a higher density.

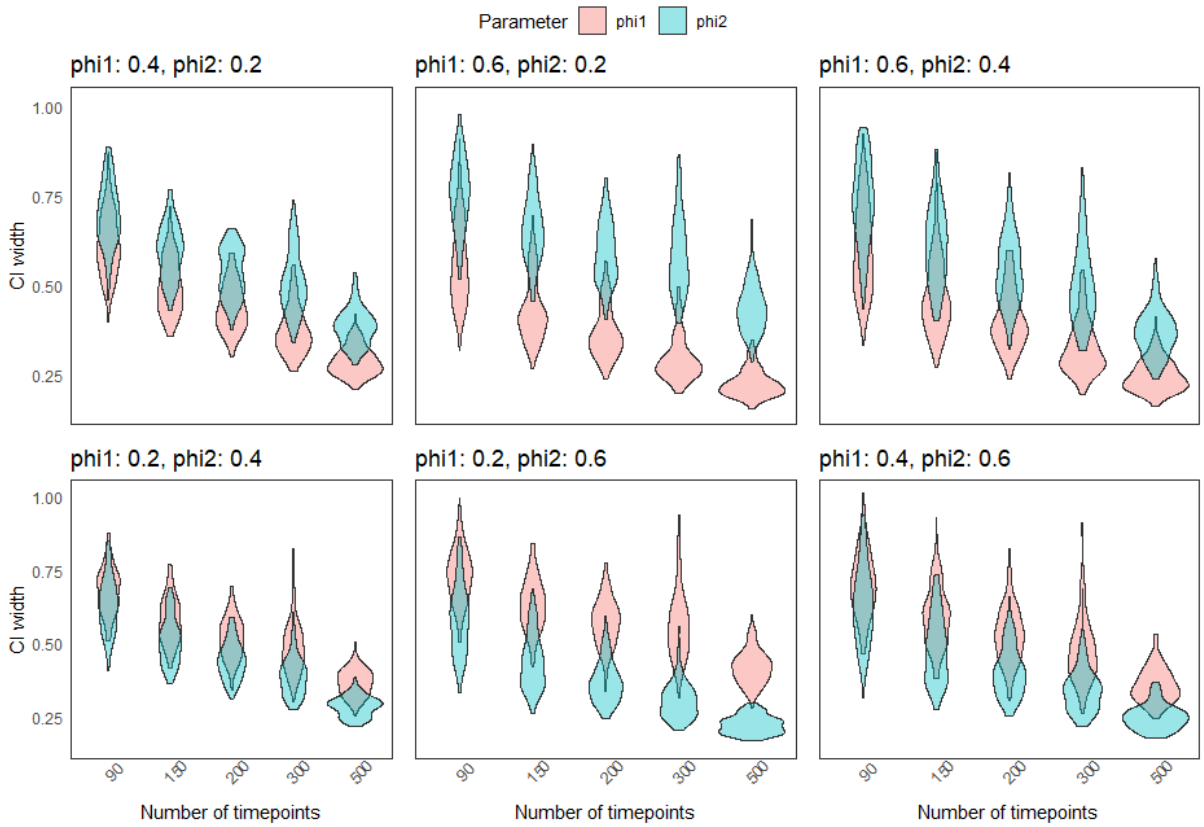


Figure 6. The width of the CI plotted against the number of timepoints for *phi1* and *phi2* (inertias). A violin shape in each plot for each group of timepoint number represents the density of the dependent variable over the 95% confidence interval. A greater width indicates a higher density.

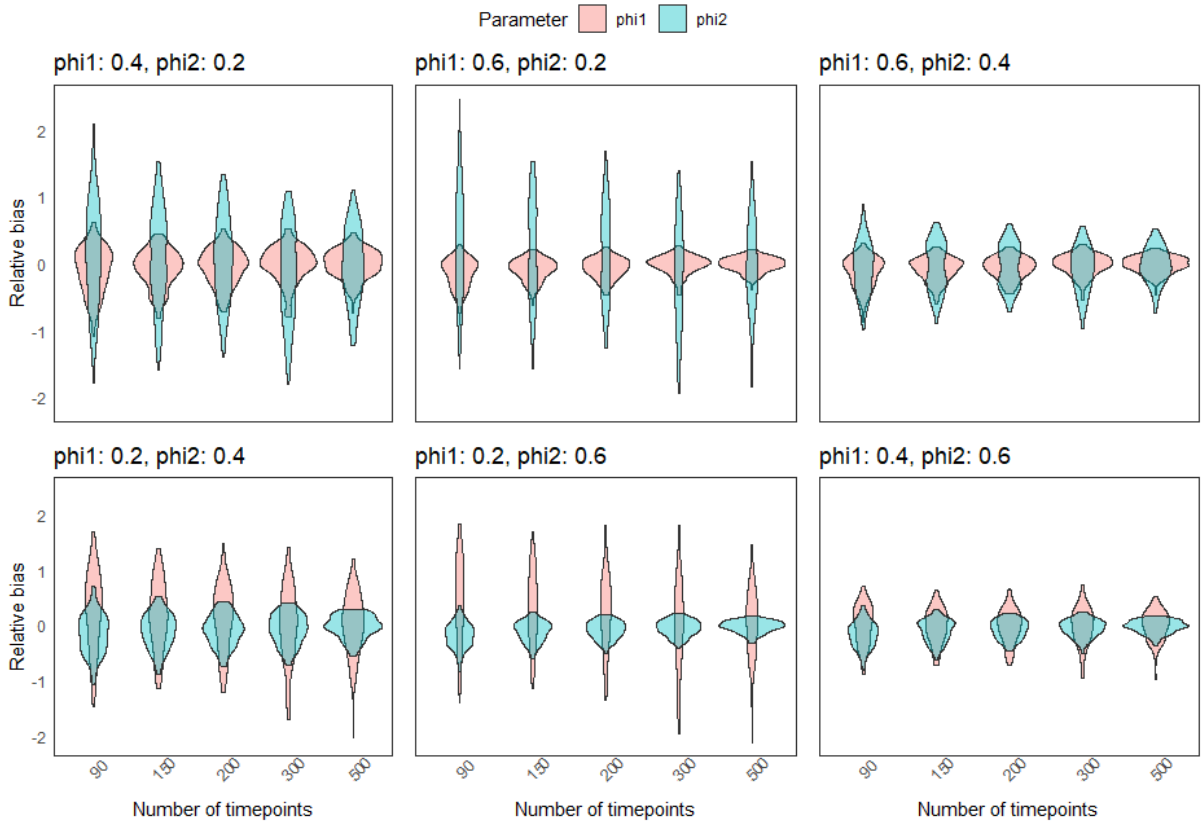


Figure 7. The relative bias plotted against the number of timepoints for the parameters *phi1* and *phi2* (inertias). The absolute bias was calculated by dividing the aforementioned bias by the true value. A violin shape in each plot for each group of timepoint number represents the density of the dependent variable over the 95% confidence interval. A greater width indicates a higher density.

Analysis of variance (ANOVA) between groups of timepoint number of model performance criteria

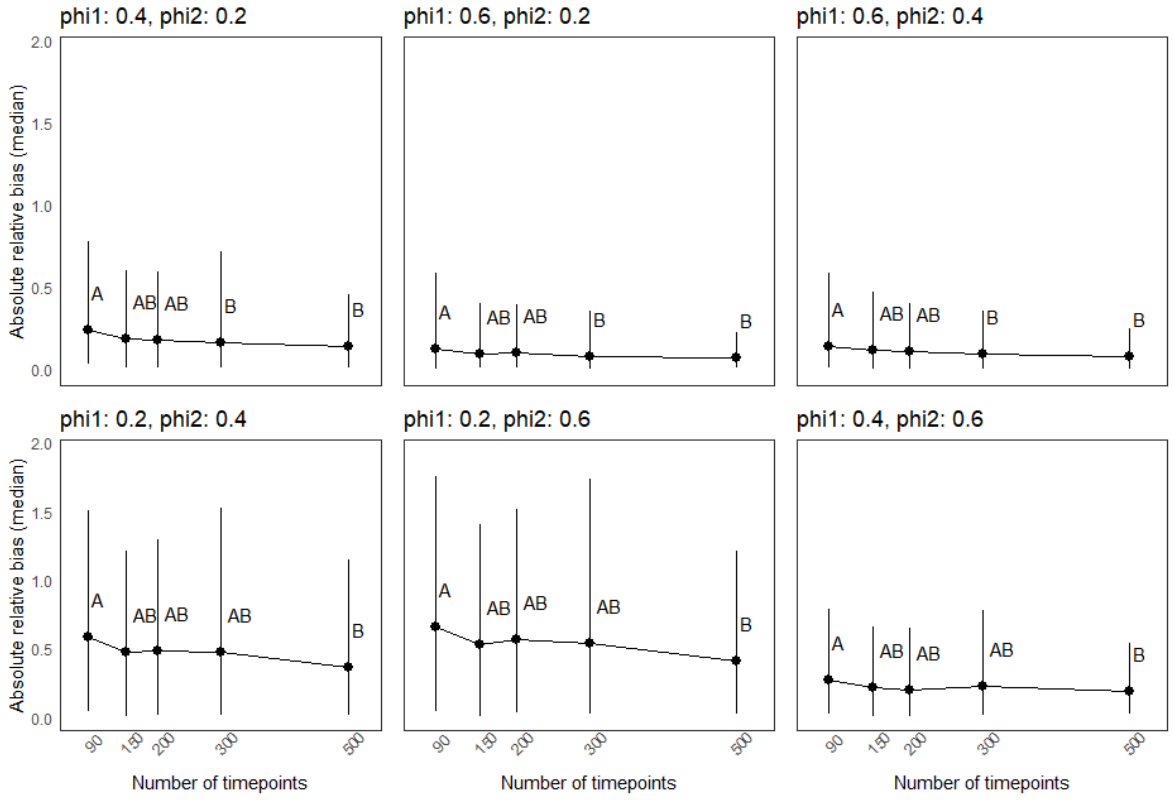
For the width of the CI and the relative bias, we conducted an analysis of variance (ANOVA) between the numbers of timepoints for all of the model parameters except *phi_diff* of all of the six possible combinations for the two inertias. We did a slight, but important modification on the variable that was the relative bias. From the relative bias, we created the so-called *absolute* relative bias, such that groups could not differ due to over- or underestimation. Now, the absolute relative bias is merely a measure of the accuracy with which the parameters are estimated. The absolute relative bias was created by simply taking the absolute value of the relative bias.

The ANOVA was performed after we did tests for normality of the data and homogeneity of variances between numbers of timepoints. We used the Shapiro-Wilk test to test for normality of the data in the timepoint groups for all of the six combinations of inertias for both inertias and both the variables width of the CI and absolute relative bias, leaving us with 48 sets of 5 comparable timepoint number groups. To summarize, in all of the 48 groups, there was at least one timepoint number that contained a non-normal distribution of variable values. This makes a regular ANOVA that assumes normality of all groups unfeasible. Next, we tested the homogeneity of variance in the 48 sets of 5 groups, using the Fligner-Killeen test of homoscedasticity. This test has the advantage that it does not require normality of the data (Fligner & Killeen, 1976). In 46 out of the 48 sets, there was heteroscedasticity between the timepoint number groups. Therefore, we decided to do a Kruskal-

Wallis rank sum test to test for significant differences between the medians of the timepoint number groups. In this case, the Kruskal-Wallis test is a test that does not assume normality of the data and is therefore close to a suitable solution for the ANOVA. The Conover's non-parametric all-pairs comparison test was used as a post-hoc analysis to investigate which timepoint number groups differed significantly from one another in the 48 sets. The level of significance was held at $\alpha = 0.05$.

The results for the ANOVA for the two inertias are summarized in the figures 8A through 9B. Note that here as well, the results for the other parameters have been omitted, because they were similar. The visualization of these results can be regenerated with a script provided in the Supplementary Materials. Figure 8 represents the results for the absolute relative bias and Figure 9 represents the results for the width of the CI. One of the results is that only the group of simulations with 500 timepoints always differs significantly from the group of 90 timepoints with regard to the values of the absolute relative bias and width of the CI. However, for only the width of the CI, all of the time point number groups differ significantly from one another. This suggests that there is little variation between timepoints in the accuracy with which the MCMC algorithm estimates the two inertias, but more variation in the certainty of the parameter estimates. In this case, the certainty seems to be of more concern than the accuracy when a clinician has to choose the appropriate number of timepoints. We will elaborate on this more in the discussion section.

A



B

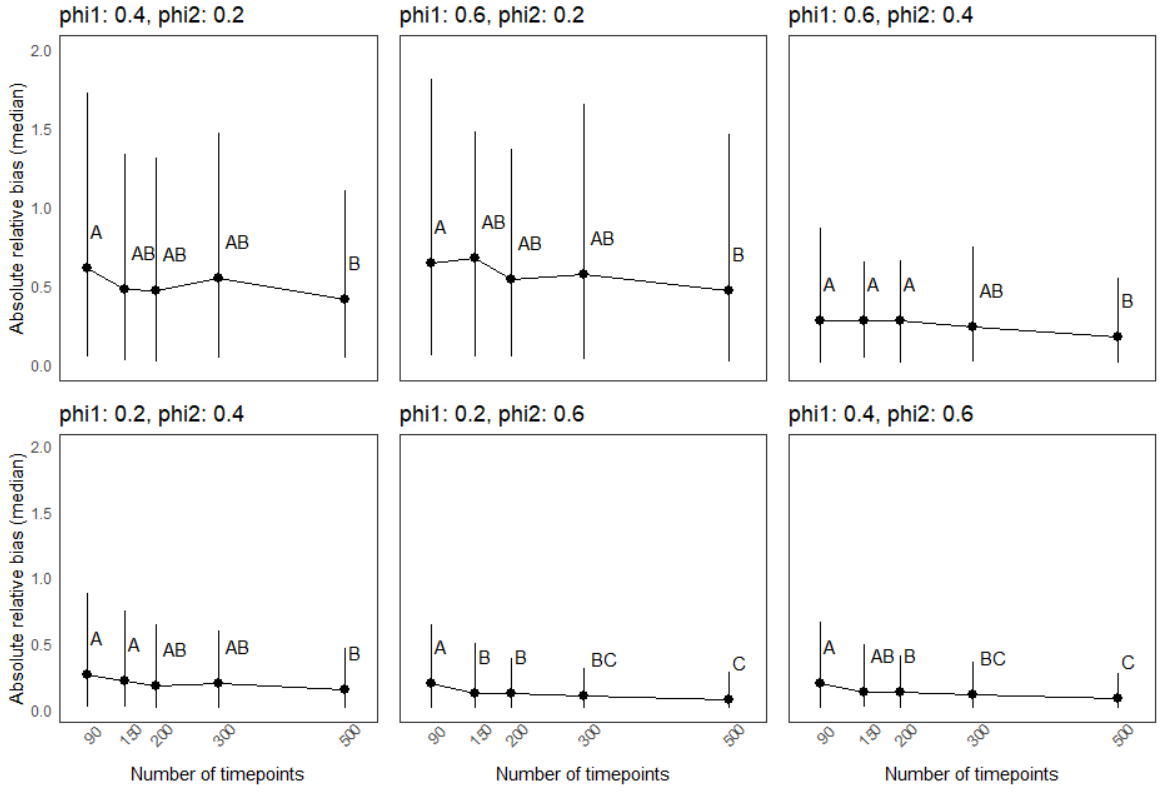
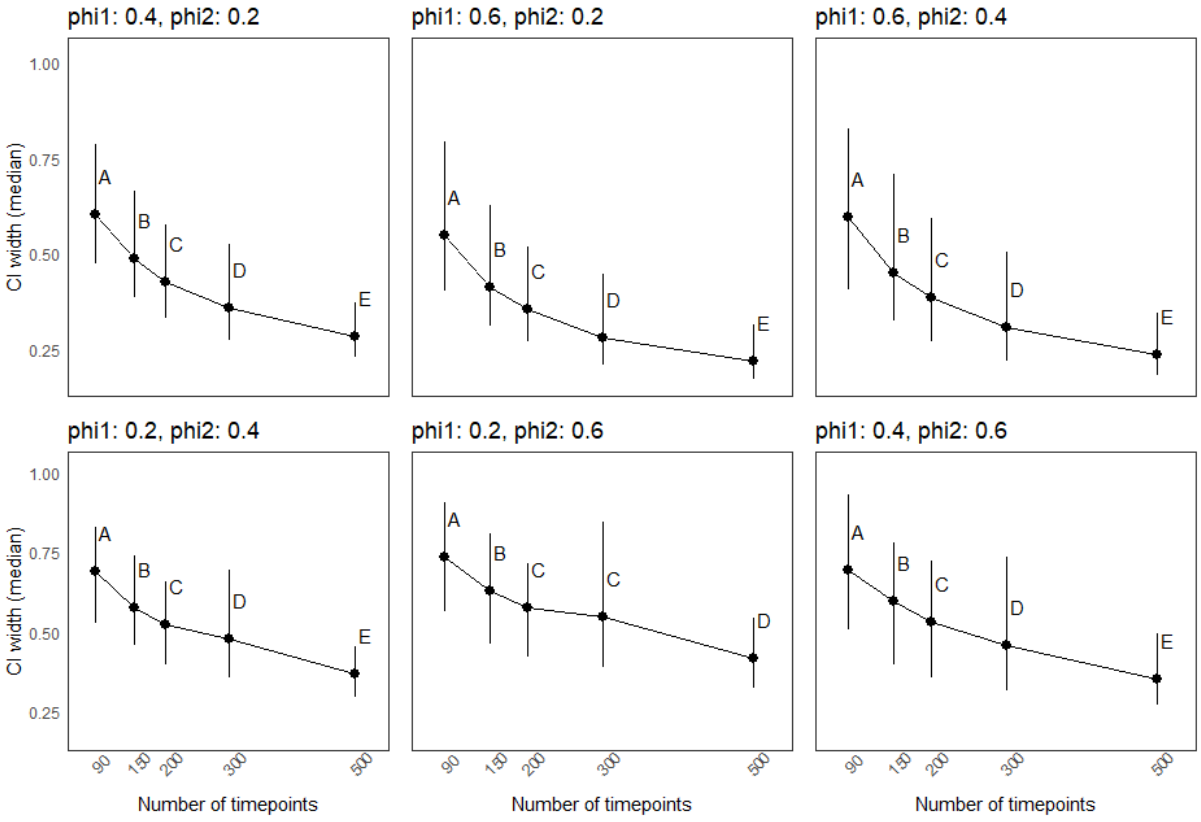


Figure 8. ANOVA of the results of the absolute relative bias compared between the five timepoint number groups for all of the six combinations of true inertia values, for (**A**) *phi1* and (**B**) *phi2*. The significance labels are placed

to the right and upwards from the median point of each timepoint number group value. Different letters between two groups indicates a significant difference between those groups. A group that has a letter in common with two or more other groups does not significantly differ from either group. A point in each line indicates the median of a time point number group, while length of the vertical lines are determined by the 2.5 percentile and the 97.5 percentile.

A


phi1: 0.4, phi2: 0.2
phi1: 0.6, phi2: 0.2
phi1: 0.6, phi2: 0.4
phi1: 0.2, phi2: 0.4
phi1: 0.2, phi2: 0.6
phi1: 0.4, phi2: 0.6
CI width (median)
1.00
0.75
0.50
0.25
A
B
C
D
E
90
150
200
300
500
Number of timepoints

B

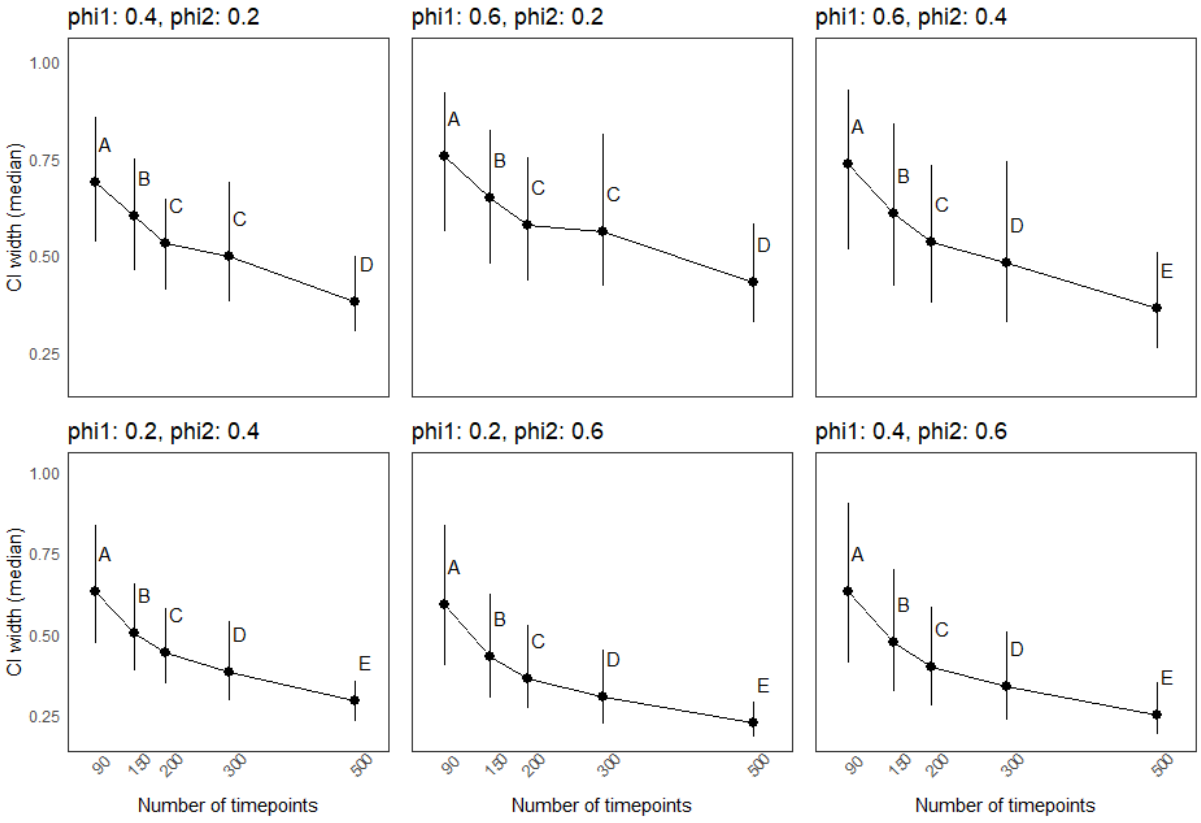

phi1: 0.4, phi2: 0.2
phi1: 0.6, phi2: 0.2
phi1: 0.6, phi2: 0.4
phi1: 0.2, phi2: 0.4
phi1: 0.2, phi2: 0.6
phi1: 0.4, phi2: 0.6
CI width (median)
1.00
0.75
0.50
0.25
A
B
C
D
E
90
150
200
300
500
Number of timepoints

Figure 9. ANOVA of the results of the width of the CI compared between the five timepoint number groups for all of the six combinations of true inertia values, for (**A**) *phi1* and (**B**) *phi2*. The significance labels are placed to the right and upwards from the median point of each timepoint number group value. Different letters between two groups indicates a significant difference between those groups. A group that has a letter in common with two or more other groups does not significantly differ from either group. A point in each line indicates the median of a time point number group, while length of the vertical lines are determined by the 2.5 percentile and the 97.5 percentile.

4 – Empirical applications for the TAR model

4.1 – Affect and sleep dynamics in a patient diagnosed with major depressive disorder

4.1.1 – The dataset

The dataset consists of data that were collected during a case study conducted by the participant of the study themselves. The participant has a history of major depression. For more details on their psychiatric course, readers are referred to (Groot, 2010). The data were collected on 239 consecutive days between August 13, 2012 and April 11, 2013. The maximal number of measurements per day could be 10. The total number of observations was 1474, resulting in an average completion of 62% per day (Kossakowski et al., 2017; Wichers et al., 2016). The dataset contains a total of 85 variables, of which 50 are momentary items. These items were collected using a semi-random version of the experience sampling method (ESM; which was also used in the studies mentioned in the beginning of this paper). Items that were used represent facets of affect a person can have. Examples of affect items include irritation, contentment, loneliness, anxiety, enthusiasm, cheerfulness, and more (Wichers et al., 2016). Such items were often measured on a Likert scale (Joshi, Kale, Chandel, & Pal, 2015), ranging from 1 to 7 or -3 to 3. For a more extensive overview of ESM and the dataset itself, see (Larson & Csikszentmihalyi, 2014) and (Kossakowski et al., 2017), respectively.

4.1.2. – Research question and hypotheses

For this empirical dataset, we are interested in whether inertia in positive affect (PA), negative affect (NA) and sleep are state-dependent. For PA, we hypothesize that the inertia for the lower state is higher than the inertia for the upper state. This is because of the negative correlation between PA mean level and inertia (de Haan-Rietdijk et al., 2016). On the other hand, for NA, we presume that the inertia for the lower state is lower than the inertia for the higher state, because of the reportedly positive correlation between NA level and inertia (de Haan-Rietdijk, Kuppens, et al., 2016).

In summary, this yields the following TAR model:

$$Y_t = \begin{cases} \tau + \varphi_1(PA_{t-1} - \tau) + \varepsilon_t & \text{if} \quad PA_{t-1} < \tau \\ \tau + \varphi_2(PA_{t-1} - \tau) + \varepsilon_t & \text{if} \quad PA_{t-1} \geq \tau \end{cases}$$

According to our hypothesis, we assume that $\varphi_1 > \varphi_2$. For NA we have the same model, but then substitute PA with NA. With the model for NA, we instead assume that $\varphi_1 < \varphi_2$.

4.1.3. Pre-processing of the dataset

Firstly, variables that needed recoding were recoded. This was done for some mood variables and most of the variables related to sleep. For instance, the variables that represent measurements of anxiety (*mood_anxious*) and feelings of guilt (*mood_guilty*) were recoded such that they could range from 1 to 7 instead of -3 to 3.

Next, we aggregated variables of interest into subscales. For instance, variables that are related to affect were aggregated into the two affect dimensions PA and NA. More precisely, the PA dimension entailed the variables that represented measurements on feeling: relaxed (*mood_relaxed*), satisfied *(mood_satisfi),* enthusiastic (*mood_enthus*), cheerful (*mood_cheerf*), strong (*mood_strong*), concentrated (*pat_concent*). The NA dimension encompassed measurements of feeling: down (*mood_down*), irritated (*mood_irritat*), lonely (*mood_lonely*), anxious (*mood_anxious*), guilty (*mood_guilty*), agitated (*mood_agitate*), worried (*pat_worry*) and ashamed (*se_ashamed*). This grouping of items is based on an extensive body of literature in which researchers determined a factor structure for the Positive and Negative Affect Schedule in children (Wróbel, Finogenow, Szymańska, & Laurent, 2019; Yamasaki, Katsuma, & Sakai, 2006), adolescents (Melvin & Molloy, 2000; Villodas, Villodas, & Roesch, 2011; Watson, Clark, & Tellegen, 1988), (young) adults (Crawford & Henry, 2004; Pandey & Srivastava, 2008; Pires, Filgueiras, Ribas & Santana, 2013; Terraciano, McCrae, & Costa Jr, 2003; Tuccitto, Giacobbi Jr, & Leite, 2010), elderly (Buz, Pérez-Arechaederra, Fernández-Pulido, & Urchaga, 2015; Kercher, 1992) and people with mental disorders (Díaz-García et al., 2020; Lim, Yu, Kim, & Kim, 2010).

Lastly, since the TAR model is a discrete time model, it must be that there is an equal amount of time between two measurements. Therefore, we first inserted rows with missing values for which there could be actual measurements. Thereafter we inserted rows with missing values at timepoints where measurements could have been possible (e.g., at night-time). All this led to a total of 16 possible observations per day, including these 'missing' observations [see also (Asparouhov, Hamaker, & Muthén, 2018) for a more detailed description of this procedure].

4.1.4. Results

The analysis took 4 hours, 18 minutes and 9 seconds on a device with 8 GB RAM and 2.80 GHz processing speed.

Table 1. Point estimates (posterior means) and 95% credible intervals for the selected parameters of the TAR model for a single person with clinical data.

	Mean	**95% CI**
Positive affect (PA)		
φ_1	0.56	[0.46, 0.64]
φ_2	0.41	[0.26, 0.53]
$\varphi_1 - \varphi_2$	0.15	[-0.05, 0.35]
τ	25.2	[24.5, 26.0]
σ_ε^2	13.6	[12.6, 14.6]

Negative affect (NA)

φ_1	0.38	[0.24, 0.51]
φ_2	0.58	[0.51, 0.65]
$\varphi_1 - \varphi_2$	-0.20	[-0.38, -0.02]
τ	16.4	[15.8, 17.0]
σ_ε^2	12.6	[11.7, 13.6]

Based on the credible intervals, we conclude that only negative affect is state-dependent. For positive affect and sleep, there was no indication of state-dependent regulatory mechanisms.

Table 1 summarizes the results of the model parameters estimated from the empirical dataset. Of note is that only NA showed a credibility interval for the difference between inertias that did not include zero. This means that NA might follow a state-dependent regulation in this dataset. In this case, $\varphi_1 - \varphi_2 < 0$. Although, we did not find significantly state-dependent regulatory mechanisms for PA and sleep, we can still say something about the mean value of their inertias. For instance, the mean level of φ_1 is smaller than the mean level of φ_2 for both PA and sleep. With regard to the thresholds for PA and NA, we see that the threshold for PA is significantly higher than the threshold for NA. This might suggest that the subject overall experienced higher positive feelings than negative feelings. The variance of both PA and NA scores were similar, indicating a comparable variability for both emotion dimensions.

Figure 10 shows the state-space plot for NA, of which the data follows a TAR process. The dotted line indicates the threshold value. The line on the left from the threshold is less steep than the line on the right from the threshold. This indicates that the person has a higher inertia for more intense negative mood states.

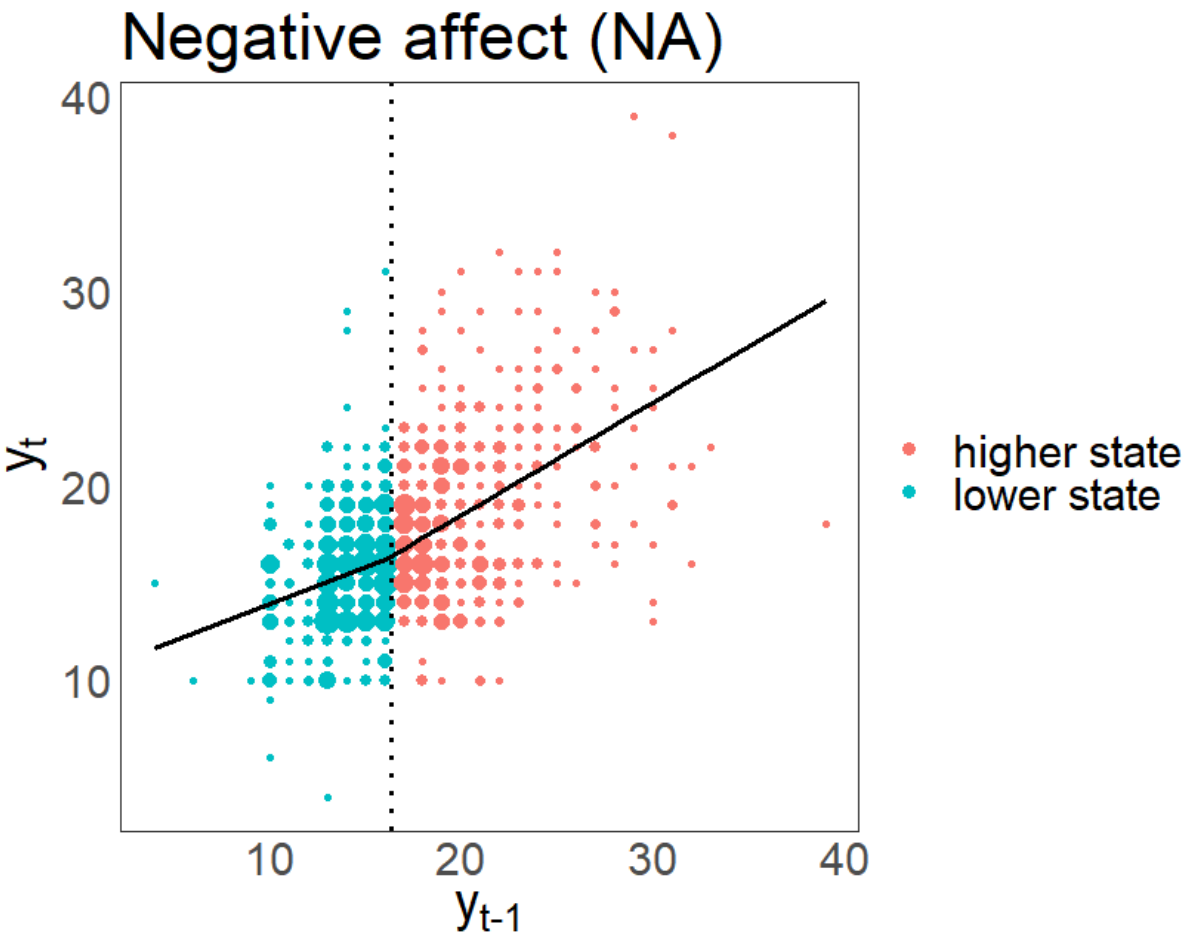


Figure 10. State-space plot of negative affect (NA). The dotted line indicates the threshold value. The steepness of the lines indicates the autoregressive effect that each timepoint has on the next. The inertia is around 0.38 in the area that is left from the threshold value, whereas the inertia is around 0.58 in the area that is right from the threshold value. The size of the points indicates the observation count for a data point.

Discussion

In this article, we tried to answer the question of how many timepoints are sufficient to make the TAR model a reliable model. We used Bayesian estimation methods to achieve this. We first tuned the MCMC parameters, such as the thinning and total number of iterations to appropriate values. Thereafter, we simulated data that followed a TAR process under 30 conditions with 100 replications per condition. We calculated several variables of interest that represented aspects of model performance. Next, we performed an ANOVA on the absolute relative bias and width of the CI for the model parameters. Lastly, we used an empirical dataset to show how to fit and interpret the TAR model. All of these results are discussed below in respective order.

The number of iterations we chose was 20.000 with thinning factor 30. We thought that this number of iterations was reasonable, as it was the minimal number at which the autocorrelation stabilized and the fluctuations in the Gelman-Rubin diagnostic mostly stabilized. Taking a minimal value was important here, so that time consumption for the simulation study could be minimized. With regard to the thinning, we initially considered a thinning factor of 40. This was because the lines in plot with autocorrelation against thinning factor mixed well at this point (Figure S1D). However, we later chose a thinning factor of 30, because that would be less time-consuming. Besides, we gained little

improvement in the autocorrelation with a thinning factor of 40 relative to the thinning factor of 30. Although we did use the autocorrelation of 0.05 as a criterion to assess model convergence in the MCMC test phase, we omitted this in the simulation study. Namely, in the MCMC test phase the autocorrelation was a useful measure as to indicate which thinning factor would be needed approximately. However, maintaining an autocorrelation of lower than 0.05 in the simulation session would be fairly strict, which is why we dropped this as a criterion for model convergence. The fraction of burn-in iterations in this study was half of the total iterations. This fraction of burn-in iterations was also chosen in another study concerning the TAR-model (de Haan-Rietdijk, Gottman et al., 2016). We think that this was an appropriate choice, as having a fraction that is too small may lead to a significant bias in estimates (Cowles, Roberts, & Rosenthal, 1999).

After that, we ran the simulation study. An important point here is the estimation of the two inertias. The TAR model overall had no problems with convergence for the simulated data, taking into account the Gelman-Rubin diagnostic that did not exceed 1.05. On top of that, there are some remarks with regard to the parameter estimation. Our power analysis revealed that the estimation method could detect the TAR process in only 23% of all cases. Even with 500 timepoints, the power could not come above 0.80, which is a recommendable level according to de Haan-Rietdijk, Gottman et al. (2016). This means that there is a high type II error rate. This implies that in many cases, the estimation algorithm fails to detect a true effect. Our research is in line with that of other research (de Haan-Rietdijk, Gottman et al., 2016). For example, with a multilevel TAR model, it was found that the power rate decreased consistently, such that more timepoints were needed for an adequate power when the number of subjects decreased. This observed relation between number of subjects and timepoints may explain why the type II error rates are high when few timepoint measurements from just a single person are used (de Haan-Rietdijk, Gottman et al., 2016).

From a clinical perspective, this means that a researcher may observe a difference between the two inertias for a patient from some data, whereas it is not statistically significant. For example, the credibility interval of the difference between the two inertias may still include 0. However, knowing that there is a relatively high type II error rate may prompt the researcher to expect that there may still be a remarkable difference. Hence, if the estimation algorithm is supplied with fewer than 500 timepoints, there is still sufficient reason to do hypothesis generating research. For hypothesis testing research, however, the researcher should rely on methods that collect data more intensively. For example, in the dataset of Peter Groot (2010), measurements were taken 10 times a day. This may be costly for a patient who is under clinical assessment. However, this may be accounted for by only focusing on measurement variables that directly relate to a person's vulnerability.

Furthermore, there is the ANOVA that was performed on the absolute relative bias and the width of the CI of the model parameters. We could observe that the absolute relative bias showed relatively little variation across the number of timepoints. On the contrary, there was more variation in the width of the CI across timepoints. From this it follows that the number of timepoints chosen by a clinician is determinative of the certainty with which the parameters of interest are estimated. Thus, if a clinician's goal is to focus on hypothesis generating research and still maintain a high certainty with which parameters are estimated, then 300 timepoints or more would be required. We could say that the choice of the number of timepoints should depend on the set up of the research (hypothesis generating or -testing) and the certainty with which the clinician wants to estimate their parameters.

Also, a clinician may want to reconsider the choice of the threshold variable, such that it fits to the question that is asked. A limitation of the TAR model that is used in this paper is that it assumes that the threshold variable is the same as the regression variable. This can be overcome by the implementation of another variable as the threshold variable. For example, there exists an intimate

relationship between measures of sleep quality and subsequent mood states (Konjarski, Murray, Lee, & Jackson, 2018). Hence, the TAR model that we used can be modified by substituting the emotional threshold variable by a variable that represents sleep quality. In addition, the current TAR model contains only one variable that is regressed on itself. A multivariate extension of the TAR model is the vector-TAR model (VTAR), where multiple variables are regressed on the dependent variable (Takano, Stefanovic, Rosenkranz, & Ehring, 2021). In this way, there might exist separate regimes in which some variables have a stronger or weaker effect on the variable of interest. For example, it is interesting to speculate that poor quality of sleep might be more predictive of weak emotional regulation if an individual experiences higher intensities of negative affect than if an individual experiences low intensities of negative affect. The latter requires a reciprocal relationship between affect and sleep, which seems to exist (Konjarski et al., 2018).

In general, time is an important aspect with regard to the TAR model. The TAR model in this paper assumes that there is equal spacing between timepoints. Hence, it is a discrete timepoint TAR model. Therefore, we had to include missing values in the empirical dataset of Groot (2010) at the timepoints where there could actually have been data. This is a limitation, because inserting missing values requires an extra step in the pre-procession of the dataset. One way to circumvent this is to account for variation in the length of the periods between measurements by using continuous-time dynamic models. In this way, the clinician may be able to handle uneven time intervals. More details on how to use continuous-time dynamic models can be found here (de Haan-Rietdijk, Voelkle, Keijsers, & Hamaker, 2017).

In relation to this, there exists another way in which time can play a role in the TAR model. For example, the inertia parameter in the TAR model used here only depends on affect intensity. However, it might be interesting to make the different inertia parameters time-variant. For example, there is evidence that an emotional state may influence the next emotional state more as time progresses (Bringmann, Ferrer, Hamaker, Borsboom, & Tuerlinkcx, 2018). This implies that inertia can change over time. This can be modelled using a time-varying autoregressive (TVAR) model (Bringmann et al., 2017). Interestingly, this TVAR model can be combined with the TAR model. With this, we may even model a situation in which the change in affect regulation depends on affect intensity.

To conclude, the TAR model for a single person can be useful for both hypothesis generating and hypothesis testing research. In general, the choice of the number of timepoints should depend on the goal of the research and the certainty with which the clinician wants to estimate their parameters. Together with that, there exists a broad range of extensions on the TAR model of which the use can be investigated in future research.

References

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. Structural Equation Modeling: A Multidisciplinary Journal, 25(3), 359-388. https://doi.org/10.1080/10705511.2017.1406803

Borkenau, P., & Ostendorf, F. (1998). The Big Five as states: How useful is the five-factor model to describe intraindividual variations over time?. Journal of Research in Personality, *32*(2), 202-221. https://doi.org/10.1006/jrpe.1997.2206

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley & Sons.

Bringmann, L. F., Ferrer, E., Hamaker, E. L., Borsboom, D., & Tuerlinckx, F. (2018). Modeling nonstationary emotion dynamics in dyads using a time-varying vector-autoregressive model. Multivariate behavioral research, 53(3), 293-314. https://doi.org/10.1080/00273171.2018.1439722

Bringmann, L. F., Hamaker, E. L., Vigo, D. E., Aubert, A., Borsboom, D., & Tuerlinckx, F. (2017). Changing dynamics: Time-varying autoregressive models using generalized additive modeling. Psychological methods, 22(3), 409. https://doi.org/10.1037/met0000085

Bringmann, L. F., Lemmens, L. H. J. M., Huibers, M. J. H., Borsboom, D., & Tuerlinckx, F. J. P. M. (2015). Revealing the dynamic network structure of the Beck Depression Inventory-II. Psychological medicine, *45*(4), 747-757. https://doi.org/10.1017/S0033291714001809

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. https://doi.org/10.1371/journal.pone.0060188

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. Journal of computational and graphical statistics, *7*(4), 434-455. https://doi.org/10.2307/1390675

Brose, A., Schmiedek, F., Koval, P., & Kuppens, P. (2015). Emotional inertia contributes to depressive symptoms beyond perseverative thinking. Cognition and Emotion, *29*(3), 527-538. https://doi.org/10.1080/02699931.2014.916252

Buz, J., Pérez-Arechaederra, D., Fernández-Pulido, R., & Urchaga, D. (2015). Factorial structure and measurement invariance of the PANAS in Spanish older adults. The Spanish Journal of Psychology, 18. https://doi.org/10.1017/sjp.2015.6

Caner, M., & Hansen, B. E. (2001). Threshold autoregression with a unit root. Econometrica, 69(6), 1555-1596. https://doi.org/10.1111/1468-0262.00257

Cowles, M. K., Roberts, G. O., & Rosenthal, J. S. (1999). Possible biases induced by MCMC convergence diagnostics. Journal of Statistical Computation and Simulation, *64*(1), 87-104. https://doi.org/10.1080/00949659908811968

Crawford, J. R., & Henry, J. D. (2004). The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. British journal of clinical psychology, *43*(3), 245-265. https://doi.org/10.1348/0144665031752934

de Haan-Rietdijk, D., Gottman, J. M., Bergeman, C. S., & Hamaker, E. L. (2016). Get over it! A multilevel threshold autoregressive model for state-dependent affect regulation. Psychometrika, 81(1), 217-241. https://doi.org/10.1007/s11336-014-9417-x

de Haan-Rietdijk, S., Kuppens, P., & Hamaker, E. L. (2016). What's in a day? A guide to decomposing the variance in intensive longitudinal data. Frontiers in Psychology, 7, 891. https://doi.org/10.3389/fpsyg.2016.00891

de Haan-Rietdijk, S., Voelkle, M. C., Keijsers, L., & Hamaker, E. L. (2017). Discrete-vs. continuous-time modeling of unequally spaced experience sampling method data. Frontiers in psychology, 8, 1849. https://doi.org/10.3389/fpsyg.2017.01849

Díaz-García, A., González-Robles, A., Mor, S., Mira, A., Quero, S., García-Palacios, A., ... & Botella, C. (2020). Positive and Negative Affect Schedule (PANAS): psychometric properties of the online Spanish

version in a clinical sample with emotional disorders. BMC psychiatry, *20*(1), 1-13. https://doi.org/10.1186/s12888-020-2472-1

Ferrer, E., Steele, J. S., & Hsieh, F. (2012). Analyzing the dynamics of affective dyadic interactions using patterns of intra-and interindividual variability. Multivariate Behavioral Research, *47*(1), 136-171. https://doi.org/10.1080/00273171.2012.640605

Figueiredo, E., Figueiras, J., Park, G., Farrar, C. R., & Worden, K. (2011). Influence of the autoregressive model order on damage detection. Computer-Aided Civil and Infrastructure Engineering, *26*(3), 225-238. https://doi.org/10.1111/j.1467-8667.2010.00685.x

Fligner, M. A., & Killeen, T. J. (1976). Distribution-free two-sample tests for scale. Journal of the American Statistical Association, *71*(353), 210-213. https://doi.org/10.2307/2285771

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. Statistical science, *7*(4), 457-472. DOI: 10.1214/ss/1177011136

Geyer, C. J. (1992). Practical markov chain monte carlo. Statistical science, 473-483. DOI: 10.1214/ss/1177011137

Gottman, J., Swanson, C., & Murray, J. (1999). The mathematics of marital conflict: Dynamic mathematical nonlinear modeling of newlywed marital interaction. Journal of Family Psychology, *13*(1), 3. https://doi.org/10.1037/0893-3200.13.1.3

Groot, P. C. (2010). Patients can diagnose too: how continuous self-assessment aids diagnosis of, and recovery from, depression. Journal of Mental Health, *19*(4), 352-362. https://doi.org/10.3109/09638237.2010.494188

Hamaker, E. L., Grasman, R. P., & Kamphuis, J. H. (2016). Modeling BAS dysregulation in bipolar disorder: Illustrating the potential of time series analysis. Assessment, *23*(4), 436-446. https://doi.org/10.1177/1073191116632339

Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. Psychological bulletin, *141*(4), 901. DOI: 10.1037/a0038822

Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert scale: Explored and explained. British journal of applied science & technology, *7*(4), 396. DOI: 10.9734/BJAST/2015/14975

Kercher, K. (1992). Assessing subjective well-being in the old-old: The PANAS as a measure of orthogonal dimensions of positive and negative affect. Research on Aging, *14*(2), 131-168. https://doi.org/10.1177/0164027592142001

Konjarski, M., Murray, G., Lee, V. V., & Jackson, M. L. (2018). Reciprocal relationships between daily sleep and mood: A systematic review of naturalistic prospective studies. Sleep medicine reviews, 42, 47-58. https://doi.org/10.1016/j.smrv.2018.05.005

Kossakowski, J., Groot, P., Haslbeck, J., Borsboom, D., & Wichers, M. (2017). Data from 'critical slowing down as a personalized early warning signal for depression'. Journal of Open Psychology Data, *5*(1). http://doi.org/10.5334/jopd.29

Koval, P., & Kuppens, P. (2012). Changing emotion dynamics: individual differences in the effect of anticipatory social stress on emotional inertia. Emotion, *12*(2), 256. https://doi.org/10.1037/a0024756

Koval, P., Brose, A., Pe, M. L., Houben, M., Erbas, Y., Champagne, D., & Kuppens, P. (2015). Emotional inertia and external events: The roles of exposure, reactivity, and recovery. Emotion, *15*(5), 625. https://doi.org/10.1037/emo0000059

Koval, P., Butler, E. A., Hollenstein, T., Lanteigne, D., & Kuppens, P. (2015). Emotion regulation and the temporal dynamics of emotions: Effects of cognitive reappraisal and expressive suppression on emotional inertia. Cognition and Emotion, *29*(5), 831-851. https://doi.org/10.1080/02699931.2014.948388

Koval, P., Kuppens, P., Allen, N. B., & Sheeber, L. (2012). Getting stuck in depression: The roles of rumination and emotional inertia. Cognition & emotion, *26*(8), 1412-1427. https://doi.org/10.1080/02699931.2012.667392

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. https://doi.org/10.1037/a0033579

Koval, P., Sütterlin, S., & Kuppens, P. (2016). Emotional inertia is associated with lower well-being when controlling for differences in emotional context. Frontiers in Psychology, 6, 1997. https://doi.org/10.3389/fpsyg.2015.01997

Kruschke, J. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan.

Kuppens, P., Allen, N. B., & Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. Psychological science, *21*(7), 984-991. https://doi.org/10.1177/0956797610372634

Lanfear, R., Hua, X., & Warren, D. L. (2016). Estimating the effective sample size of tree topologies from Bayesian phylogenetic analyses. Genome biology and evolution, 8(8), 2319-2332. https://doi.org/10.1093/gbe/evw171

Larson, R., & Csikszentmihalyi, M. (2014). The experience sampling method. In Flow and the foundations of positive psychology (pp. 21-34). Springer, Dordrecht. https://doi.org/10.1007/978-94-017-9088-8_2

Lim, Y. J., Yu, B. H., Kim, D. K., & Kim, J. H. (2010). The positive and negative affect schedule: Psychometric properties of the Korean version. Psychiatry investigation, *7*(3), 163. DOI: 10.4306/pi.2010.7.3.163

Link, W. A., & Eaton, M. J. (2012). On thinning of chains in MCMC. Methods in ecology and evolution, 3(1), 112-115. https://doi.org/10.1111/j.2041-210X.2011.00131.x

MacEachern, S. N., & Berliner, L. M. (1994). Subsampling the Gibbs sampler. The American Statistician, 48(3), 188-190. https://doi.org/10.2307/2684714

McNally, R. J., Mair, P., Mugno, B. L., & Riemann, B. C. (2017). Co-morbid obsessive–compulsive disorder and depression: A Bayesian network approach. Psychological medicine, *47*(7), 1204-1214. https://doi.org/10.1017/S0033291716003287

Melvin, G. A., & Molloy, G. N. (2000). Some psychometric properties of the Positive and Negative Affect Schedule among Australian youth. Psychological reports, *86*(3_suppl), 1209-1212. https://doi.org/10.2466/pr0.2000.86.3c.1209

Moberly, N. J., & Watkins, E. R. (2008). Ruminative self-focus and negative affect: an experience sampling study. Journal of abnormal psychology, *117*(2), 314. https://doi.org/10.1037/0021-843X.117.2.314

Narayan, P. K. (2006). The behaviour of US stock prices: Evidence from a threshold autoregressive model. Mathematics and computers in simulation, 71(2), 103-108. https://doi.org/10.1016/j.matcom.2005.11.016

Pandey, R., & Srivastava, N. (2008). Psychometric evaluation of a hindi version of positive-negative affect schedule. Industrial Psychiatry Journal, *17*(1), 49.

Pires, P., Filgueiras, A., Ribas, R., & Santana, C. (2013). Positive and negative affect schedule: psychometric properties for the Brazilian Portuguese version. The Spanish Journal of Psychology, 16. https://doi.org/10.1017/sjp.2013.60

Rzeszutek, M., & Gruszczyńska, E. (2021). Inertia, innovation, and cross-lagged effects in negative affect and rumination: daily diary study among people living with HIV. Anxiety, Stress, & Coping, *34*(4), 411-422. https://doi.org/10.1080/10615806.2021.1887481

Schuurman, N. K., Houtveen, J. H., & Hamaker, E. L. (2015). Incorporating measurement error in n= 1 psychological autoregressive modeling. Frontiers in psychology, 6, 1038. https://doi.org/10.3389/fpsyg.2015.01038\

Strikholm, B., & Teräsvirta, T. (2006). A sequential procedure for determining the number of regimes in a threshold autoregressive model. The Econometrics Journal, 9(3), 472-491.

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. https://doi.org/10.1177/0146167298242002

Takano, K., Stefanovic, M., Rosenkranz, T., & Ehring, T. (2021). Clustering individuals on limited features of a vector autoregressive model. Multivariate Behavioral Research, 56(5), 768-786. https://doi.org/10.1080/00273171.2020.1767532

Terraciano, A., McCrae, R. R., & Costa Jr, P. T. (2003). Factorial and construct validity of the Italian Positive and Negative Affect Schedule (PANAS). European journal of psychological assessment, *19*(2), 131. https://doi.org/10.1027/1015-5759.19.2.131

Toft, N., Innocent, G. T., Gettinby, G., & Reid, S. W. (2007). Assessing the convergence of Markov Chain Monte Carlo methods: an example from evaluation of diagnostic tests in absence of a gold standard. Preventive veterinary medicine, *79*(2-4), 244-256. https://doi.org/10.1016/j.prevetmed.2007.01.003

Tong, H., & Lim, K. S. (1980). Threshold Autoregression, Limit Cycles and Cyclical Data. Journal of the Royal Statistical Society. Series B (Methodological), 42(3), 245-292. https://doi.org/10.1111/j.2517-6161.1980.tb01126.x

Tsay, R. S. (1989). Testing and modeling threshold autoregressive processes. Journal of the American statistical association, *84*(405), 231-240. https://doi.org/10.2307/2289868

Tuccitto, D. E., Giacobbi Jr, P. R., & Leite, W. L. (2010). The internal structure of positive and negative affect: A confirmatory factor analysis of the PANAS. Educational and psychological measurement, *70*(1), 125-141. https://doi.org/10.1177/0013164409344522

van Borkulo, C., Boschloo, L., Borsboom, D., Penninx, B. W., Waldorp, L. J., & Schoevers, R. A. (2015). Association of symptom network structure with the course of depression. JAMA psychiatry, *72*(12), 1219-1226. DOI: 10.1001/jamapsychiatry.2015.2079

Villodas, F., Villodas, M. T., & Roesch, S. (2011). Examining the factor structure of the positive and negative affect schedule (PANAS) in a multiethnic sample of adolescents. Measurement and Evaluation in Counseling and Development, *44*(4), 193-203. https://doi.org/10.1177/0748175611414721

Wang, D., Schneider, S., Schwartz, J. E., & Stone, A. A. (2020). Heightened stress in employed individuals is linked to altered variability and inertia in emotions. Frontiers in psychology, 11, 1152. https://doi.org/10.3389/fpsyg.2020.01152

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. Journal of personality and social psychology, *54*(6), 1063. DOI:10.1037/0022-3514.54.6.1063

Wichers, M., Groot, P. C., Psychosystems, E. S. M., & EWS Group. (2016). Critical slowing down as a personalized early warning signal for depression. Psychotherapy and psychosomatics, *85*(2), 114-116. https://doi.org/10.1159/000441458

Wróbel, M., Finogenow, M., Szymańska, P., & Laurent, J. (2019). Measuring positive and negative affect in a school-based sample: a polish version of the PANAS-C. Journal of Psychopathology and Behavioral Assessment, *41*(4), 598-611. https://doi.org/10.1007/s10862-019-09720-7

Yamasaki, K., Katsuma, R., & Sakai, A. (2006). Development of a Japanese version of the positive and negative affect schedule for children. Psychological Reports, *99*(2), 535-546. https://doi.org/10.2466/pr0.99.2.535-546