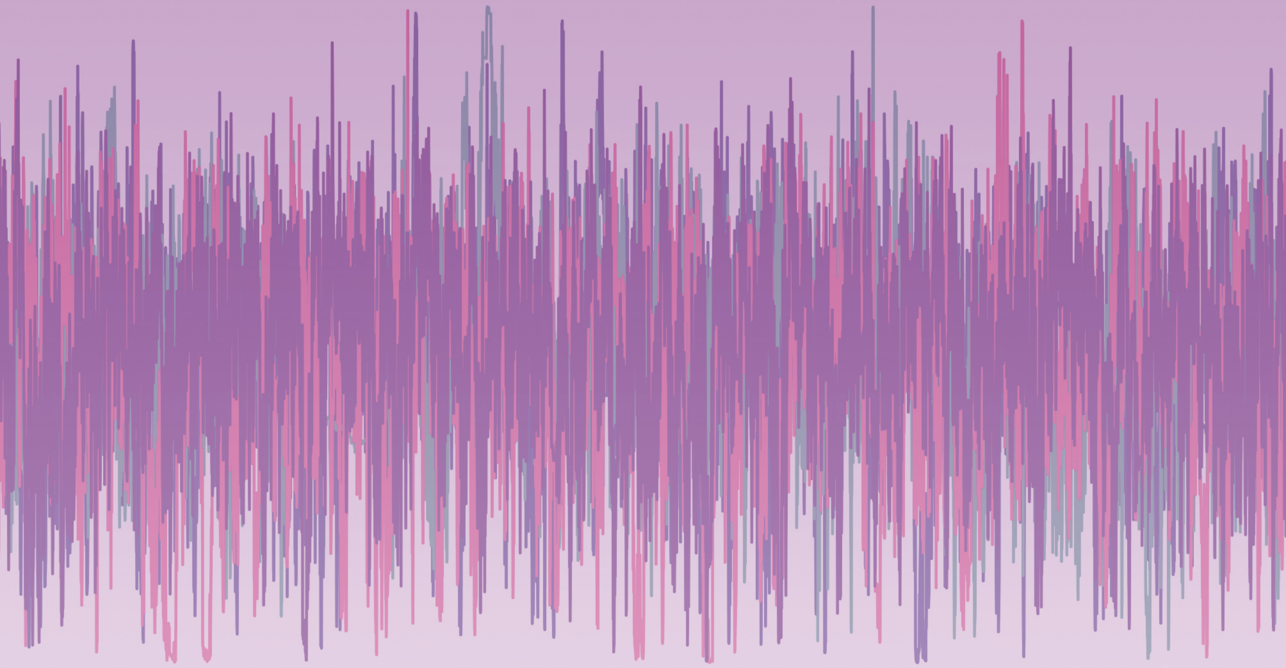
Some Notes on Bayesian Time Series Analysis in Psychology



Tanja Krone

Stellingen
behorende bij het proefschrift
Some Notes on Bayesian Time Series Analysis in Psychology
door **Tanja Krone**.

1. Voor het schatten van een autoregressief model hebben iteratieve schattingsmethoden de voorkeur (dit proefschrift, hoofdstuk 2).
2. De steekproefomvangen, dat wil zeggen de aantallen geobserveerde tijdspunten en individuen, hebben grote invloed op de kwaliteit van de schatting van het autoregressieve lag-1 model (dit proefschrift, hoofdstukken 2 en 3).
3. Het Bayesiaans dynamische lineaire model kan veel uitdagingen gevonden in empirische tijdseries aan (dit proefschrift, hoofdstukken 4, 5, en 6).
4. Een goed statistisch model voor multisubject, longitudinale, psychologische data laat ruimte voor verschillen tussen individuen (dit proefschrift, hoofdstukken 4, 5, en 6).
5. Voor een goed gebruik van statistisch modelleren moet men de psychologische theorie kunnen verbinden met de gebruikte statistische methoden.
6. Een gebrek aan convergentie hoeft niet te betekenen dat het model niet geïdentificeerd is.
7. Zowel ontspanning, zoals het bezoeken van een sauna, als inspanning, zoals het voltooien van intensieve vechtsport training, zijn van levensbelang voor een promovenda.
8. Voor totale ontspanning zou men moeten leven als een kat: eten wordt gebracht, slapen kan overal en aandacht komt op commando.
9. It’s nice to be important, but more important to be nice — John Templeton

Some Notes on

Bayesian Time Series Analysis

in Psychology

Tanja Krone

ISBN: 978-90-367-8990-5
ISBN electronic version: 978-90-367-8988-2
Printed by: Ipskamp
This work is funded by the Netherlands Organization for Scientific Research (NWO), grant number 406-11-018.

rijksuniversiteit
groningen

Some Notes on Bayesian Time Series Analysis in Psychology

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus prof. dr. E. Sterken,
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

donderdag 22 september 2016 om 16.15 uur

door

Tanja Krone

geboren op 11 augustus 1985
te Almere

Promotor
Prof. dr. M.E. Timmerman

Copromotor
Dr. C.J. Albers

Beoordelingscommissie
Prof. dr. D. Borsboom
Prof. dr. R.R. Meijer
Prof. dr. M.C. Wichers

Contents

Chapter 1

Introduction

One important aspect in psychology is the ever changing nature of the subjects of interest, human beings. A human being changes every day, every hour: how one feels, what one is doing, how one reacts to external events and to other humans. All of this is part of the complex nature of human beings and one of the reasons why psychology is such an exciting and challenging field to study. Studying the processes and dynamics over time is vital to gain insight in complex mechanisms underlying human behavior and emotions. An important tool in studying these processes is time series analysis, where repeatedly gathered, time dependent data pertaining to the same variable is studied.

Time series data is collected with such a frequency and over such a time span that it characterizes the process of interest. A time series data set may pertain to, for example, the mood of an individual measured multiple times a day over a month, or the number of symptoms experienced by individuals in different treatment groups measured weekly over a year. While time series data still requires an intensive method of data gathering, it has been strongly facilitated over the last decades. The introduction of mobile devices, such as smart phones, allows for less invasive and cumbersome methods of data gathering, while also simplifying the contact between the researcher and the individual. One method which uses these possibilities to a great extent, is ecological momentary assessment, also known as experience sampling (Larson & Csikszentmihalyi, 1983; Shiffman, Stone, & Hufford, 2008; Bolger & Laurenceau, 2013; Bos, Schoevers, & Aan het Rot, 2015). In ecological momentary assessment, questionnaires are administered multiple times per day, at either predetermined or random intervals. The wealth of information that is captured in these data leads to an increase in the amount of time series analysis in psychological sciences.

The main goals in time series analysis are forecasting and describing. When forecasting, one tries to predict the next point in the time series. In psychological sciences this may be used to anticipate when a treatment is complete, or to predict when the symptoms of a disorder change. While this is a very worthwhile goal, it is hard to achieve.

The goal of describing a time series is to discern the patterns and dynamics that characterize the data. As such, a picture may be formed of how the data changes upon external events, internal changes and the passing of time. To create a model describing time series data, one has to have a hypothesis pertaining to the process which is characterized by the data. This hypothesis is reflected both in the choice of the model class used, e.g., a random coefficients model, a state space model or a Bayesian Dynamic Model, and the elements implemented in the model, e.g., the slope and the autocorrelation, for the analysis of the data.

1.1 Describing time series

Model classes Three model classes which may be used in time series, are the random coefficients model, the state space model and the Bayesian dynamic model. The random coefficients model can describe the time series data of multiple individuals in a single model (Hox, 2010; Snijders & Bosker, 1999). As a time series model, the random coefficients model can handle both the trend and the dynamics in the data. The random coefficients model includes fixed and random effects, which yields a more efficient estimation than individual analyses. However, it limits the possibilities of interpreting the parameters of the individuals. One important advantage of the random coefficients model is that it simplifies the interpretation of hierarchically structured data. The results can be interpreted at each level of the data, i.e., the time point level and the individual level.

The state space model (SSM) is a highly versatile model for intensively measured, functionally related data, such as time series data (Durbin & Koopman, 2012). The SSM models a latent variable, or the latent state vector, underlying the observed score. In a SSM, the system equation models the latent variable, while the observation equation links the latent variable to the observed score. The SSM can handle missing data and allows for non-normally distributed residuals in the observed data through the implementation of a link function, similar to the one used in generalized linear models (Nelder & Wedderburn, 1972; McCullagh & Nelder, 1989). Furthermore, it is possible to estimate any random coefficients model for time series using a SSM.

However, to our knowledge no SSM has been developed yet that incorporates both missing data and non-normally distributed observed residuals simultaneously. This is largely due to the limitations of maximum likelihood estimation, traditionally the estimation method for SSM, which is hard to implement for non-normally distributed residuals. This means that while the SSM has a vast range of possibilities, not all of these are implemented yet, and thus these are not accessible to most researchers.

The Bayesian dynamic model (BDM) is the Bayesian counterpart of the SSM (West & Harrison, 1997). The BDM is estimated using Bayesian Markov Chain Monte Carlo (MCMC) estimation. This allows for more flexibility than the SSM with regard to the distributions used in specifying the model. Therefore, one may estimate a model for a data set containing both missing data and non-normally distributed residuals. Furthermore, the Bayesian framework allows for the inclusion of prior expectations in the model.

Model elements To create a model befitting the time series at hand, the trend and the dynamics of the data must be studied. The trend pertains to the long term movement in the data. A trend can be positive, creating an upwards movement, or negative, creating a downwards movement. While the trend is often modeled as being linear, it may also be modeled as, or in combination with, for example, an exponential or cubic trend. The trend can be used to indicate long term effects. For example, in studies comparing the effect of different treatments, the trend may indicate which of these treatment shows the strongest decay in symptoms over time. It is also possible that there is no trend in the data, indicating that a variable does not show long term changes.

The dynamics pertain to the short term movements in the data. Among the dynamics are the variability within a time series, and, in a multivariate setting, the association with other observed series. As we study data in a longitudinal setting, the dynamics of the series over time are of vital importance. One element that captures an important part of the dynamics of a time series, is the autocorrelation. The autocorrelation is the correlation between the elements of a time series, separated by a given interval (Box & Jenkins, 1976; Yule, 1927; Walker, 1931). This is used in autoregressive models, where the scores on one or more earlier time points are used as independent variables in estimating the score on the current time point. An often used model is the autoregressive lag 1 (AR(1)) model, which models data using the autocorrelation of two consecutive, equidistant scores. The same principle can be applied to the noise of a time series model, in which case it is called a moving average model (Box & Jenkins, 1976).

1.2 Outline of the thesis

1.2.1 Estimation of the autoregressive model

The AR(1) model has been estimated with a vast range of estimators, e.g., r_1 (Yule, 1927), C-statistic (Young, 1941) and maximum likelihood estimation. In Chapters 2 and 3, I study the effect of different estimation methods and data properties on the estimation of the AR(1) model.

In **Chapter 2** I introduce the AR(1) model for univariate data of a single individual. I compare six estimators for the AR(1) model, being four frequentist and two Bayesian MCMC estimators. To compare the estimators, I perform a simulation study where I vary the number of time points and the size of the autocorrelation within the time series data. I use five measures to compare the results of the different estimators over the different conditions with regard to bias, variability and power. Furthermore, I show results of misrepresenting the data by analyzing data with a different model than was used to generate the data.

In **Chapter 3** I introduce the multilevel AR(1) model for univariate data of multiple individuals. Here, I compare two estimation methods, being maximum likelihood estimation and Bayesian MCMC estimation. These two methods showed the best results in our simulation study concerning single case data (Chapter 2). Furthermore, I examine the difference between the random and fixed coefficients approach to multilevel modeling. In the random approach the individuals are assumed to be drawn randomly from a certain population, in the fixed approach no such assumption is made. To compare the four resulting estimators, I perform a simulation study varying the length of the time series, the number of individuals per sample, the mean of the autocorrelation and the standard deviation of the autocorrelation. I use six measures to compare the results of the different estimators over the different conditions with regard to bias, variability and power

1.2.2 Empirical data analysis using the BDM

Based on the simulation studies, I find that iterative estimation through maximum likelihood estimation and Bayesian MCMC has several merits when analyzing AR(1) data compared to other estimators. The Bayesian analysis has not yet been used extensively in psychological sciences, but offers certain advantages with regard to the flexibility of the models. Thus, I continue this thesis by using Bayesian MCMC on psychological time series data and exploring the possibilities this brings.

The analysis of empirical data requires attention to several points, beyond the estimation method and data properties. An important point is that empirical data often has practical issues interfering with the analysis of the data. Examples of these practical issues are missing data and the inclusion of external variables. An aim in studying empirical data is often the integration of the relevant psychological hypotheses and the observed data, with the help of the used statistical model. To this end, the statistical model must be able to quantify the important hypotheses of the psychological framework on basis of which the data is studied. In Chapters 4, 5 and 6 I study the analysis of empirical data with the BDM, where I focus on the handling of the practical issues and the integration of the hypotheses and the statistical model.

In **Chapter 4** I use the Bayesian Dynamic Model (BDM) to examine differential trends between treatment groups. Here, I show that the BDM can handle the combination of non-normally distributed residuals, inclusion of external variables both as active covariates and post-hoc, and missing data in one model. Using the BDM, I study the trend in a data set containing univariate count data of 72 individuals, with 10 to 50 time points. I compare the effects of three panic disorder treatments for individuals with and without agoraphobia, using the number of panic attacks experienced per week as dependent variable. Further, I compare different models to see whether there is an autocorrelation in the error, and whether pre-treatment symptoms influence the number of panic attacks at the beginning of the treatment.

In **Chapter 5** I use the BDM to create a model for multivariate, multi-individual time series pertaining to perceived emotions. I combine the framework of emotion dynamic features as described by Kuppens and Verduyn (2015) with the BDM, creating a vector autoregressive (VAR) BDM. Using the VAR-BDM, I quantify six emotion dynamic features, being within person and innovation variability, granularity, inertia, cross-lag regression and the intensity of the emotions. This is the first time these features are combined into one model for a multi-individual, multivariate data set including missing data. Before the empirical application, I use a short simulation study to show how many data points would be needed for the full model in a multivariate setting. As the requirements for the full model are not met in our empirical data set, we use a simplified model. Using the simplified VAR-BDM, I study the dynamics of three emotions for three individuals, with 47 to 70 time points, in one analysis.

Finally, in **Chapter 6** I use the VAR-BDM from Chapter 5 to compare different models for bivariate affect time series. For this affect data, a theoretical framework based on emotion dynamics effect framework is used. I quantify six affect features: within person and innovation variability, inertia, cross-lag regression, intensity and co-occurrence of affect. While each of these features have been studied extensively before, they have not yet been combined in one model. Each affect dynamic is linked to a parameter in the simplified VAR-BDM. Using this model, I study the bivariate affect for 12 individuals, each with 53 to 70 consecutive measurements. I compare several models, to see whether there is a weekly cycle in the affect experienced, and whether there is an autoregression present in the white noise.

Chapter 2

A comparative simulation study of AR(1) estimators in short time series

Abstract

Various estimators of the autoregressive model exist. We compare their performance in estimating the autocorrelation in short time series. In Study 1, under correct model specification, we compare the frequentist r_1 estimator, C-statistic, ordinary least squares estimator (OLS) and maximum likelihood estimator (MLE), and a Bayesian method, considering flat (B_f) and symmetrized reference (B_{sr}) priors. In a completely crossed experimental design we vary lengths of time series (i.e., T = 10, 25, 40, 50 and 100) and autocorrelation (from -0.90 to 0.90 with steps of 0.10). The results show the lowest bias for the B_{sr}, and the lowest variability for r_1. The power in different conditions is highest for B_{sr} and OLS. For T = 10, the absolute performance of all measurements is poor, as expected. In Study 2, we study robustness of the methods through misspecification by generating the data according to an ARMA(1,1) model, but still analysing the data with an AR(1) model. We use the two methods with the lowest bias for this study, i.e., B_{sr} and MLE. The bias gets larger when the non-modelled moving average parameter becomes larger. Both the variability and power show dependency on the non-modelled parameter. The differences between the two estimation methods are negligible for all measurements.

This chapter is published as: Krone, T., Albers, C. J., & Timmerman, M. E. (2016a). A comparative simulation study of AR(1) estimators in short time series. *Quality & Quantity (in press)*. doi:10.1007/s11135-015-0290-1

2.1 Introduction

Time series analysis has been valuable for achieving insight into the nature of longitudinal processes. Especially the autoregressive moving average (ARMA) model (Box & Jenkins, 1976) has gained enormous popularity in various research areas. The autoregressive part models the serial dependence between consecutive measurements. The moving average part models the serial dependence between consecutive error terms. The ARMA(p, q) model is given by:

$$y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j e_{t-j} + e_t, \quad e_t \sim N(0, \sigma_e^2), \tag{2.1}$$

with y_t the score at time t $(t = 1, 2, .., T)$, μ the population mean, ϕ_i the autocorrelation for lag i $(i = 1, 2, ..., p)$, θ_j the moving average parameter at lag j $(j = 1, 2, ..., q)$ and e_t the residual.

One of the simplest versions of the ARMA(p, q) is the AR(1) model:

$$y_t = \mu + \phi(y_{t-1} - \mu) + e_t, \quad e_t \sim N(0, \sigma_e^2), \tag{2.2}$$

where, for simplicity, the subscript 1 is omitted from ϕ. Several estimation methods have been proposed to estimate the AR(1) model. These estimation methods include closed form estimation methods, such as the r_1 estimator (Yule, 1927; Walker, 1931; Box & Jenkins, 1976), C-statistic (Young, 1941) and Ordinary Least Squares (OLS) estimator, and iterative estimation methods, such as frequentist Maximum Likelihood Estimation (MLE) and Bayesian Markov Chain Monte Carlo (MCMC) estimation. The performance of the closed form estimation methods in terms of efficiency have been examined and compared in some simulation studies (Huitema & McKean, 1991; DeCarlo & Tryon, 1993; Arnau & Bono, 2001; Solanas, Manolov, & Sierra, 2010). Generally, in particular for shorter time series (e.g., length $T \leq 50$), the closed form estimation methods have been shown to have biased autocorrelation estimates and/or high variability. Because the closed form and iterative estimation methods have not been mutually compared so far, it is unclear which estimation methods perform better in terms of having a low bias and variability under relevant conditions for empirical practice. Further, little is known about the robustness of the specific estimation methods towards misspecification of the model. This knowledge is important to optimize a time series research design, and to select a low-variability, low-bias, and robust method for estimating an AR(1) model in empirical practice.

In this chapter, we discuss two studies to assess the relative performance of several estimators of the AR(1) model. We focus on short time series, with a length T between 10 and 100. Even though these lengths are relevant, for example in psychological research, they are not thoroughly studied yet for all estimators we compare. For the autocorrelation we use values between -1 and 1, and hence consider stationary time series. Our first study provides the information needed to make an informed choice between the estimation methods for the AR(1) model. To this end, we selected five popular and/or promising estimation methods. In a simulation study, we compare these methods with regard to bias, standard error, the bias of the standard error, the rejection rate for $\phi = 0$, the power for $\phi \neq 0$, and the point and 95% interval estimates.

Our second study focuses on the issue of robustness. Robustness, as used in this chapter, is the resilience to misspecification with regard to the number of parameters. The effects of misspecification of the ARMA(1,1), AR(1) and AR(2) model have been studied for the least squares estimator. For an underspecified model, the parameters become more biased when the unspecified parameters are further from zero (Tanaka & Maekawa, 1984). Overspecification of the model gives a larger prediction mean squared error for the estimation of the score at y_t (Kunitomo & Yamamoto, 1985). To study the robustness with regard to misspecification, we use the two estimation methods that showed the lowest bias in the first study. In the misspecification study we generate the data using an ARMA(1,1) model, but estimate the parameters as if the data was generated using the same AR(1) = ARMA(1,0) model as used in Study 1.

In the next section, we describe the selection process of the estimation methods used in this chapter, followed by a short introduction to the estimators. Then, we present the design, performance criteria and the results of the first simulation study, which aims at comparing various estimators when applied to short time series following an AR(1) model. We continue with the design and results of the second simulation study, which aims at exploring the effect of underspecifying a short time series following an ARMA(1,1) model as data following an AR(1) model. We conclude this chapter with a discussion of the simulation studies and the implications of the results.

2.2 Selection of estimation methods

To start, we performed a literature search towards estimation methods for AR(1) models. Our selection criteria for the papers were as follows: 1) it must discuss one or more simulation studies that compare different estimators of the AR(1) model; 2) it must include conditions with less than 50 time points and a range of values between -1 and 1 for the autocorrelation ϕ. This literature search revealed the five papers shown in Table 2.1.

The earliest discussed estimator is the r_1 estimator (Walker, 1931), as implemented in the Yule-Walker model (Yule, 1927; Box & Jenkins, 1976). However, since several studies have found that the bias of r_1 for small samples is large, espe-

paper	estimators	length	outcome measures
Huitema and McKean (1991)	r_1, r_1^+, r_1*, r_1^c, r_1^e, OLS	6, 10, 20, 50, 100, 500	bias (th & emp), MSE (av+), α, power, σ_e^2 (th & emp)
DeCarlo and Tryon (1993)	r_1, r_1^+, C	6, 10, 20, 30, 50	bias (emp), MSE (av+), α, power
Huitema and McKean (1994)	r_{q1}, r_{q2}, r_{q3}	6, 10,20, 50, 100, 500	bias (emp, av & ϕ=0.9), MSE (av), α, power, σ_e^2 (emp)
Arnau and Bono (2001)	r_1, r_1^+, r_1'	6, 10, 20, 30, 50	bias (th & emp), MSE (av), α, power
Solanas, Manolov, and Sierra (2010)	r_1,r_1^+, r_1*, r_1^c, r_1^e, C, r_1^f, OLS, r_1^{fb},r_1^δ	5, 6, 7, 8, 9, 10, 15, 20, 50, 100	bias (emp), MSE, α, power

Table 2.1: List of papers with considered estimators, the lengths of the time series, and the outcome measures, where ϕ = autocorrelation, th = theoretical, emp = empirical, av = averaged over all ϕ, and av+ = averaged over all positive ϕ. All papers used a range of simulated autocorrelations of [-0.9 (0.1) 0.9], the estimators with 'r' in their name are derived from r_1 estimator.

cially for data with a positive autocorrelation, various alternatives were proposed (Huitema & McKean, 1991, 1994; DeCarlo & Tryon, 1993; Arnau & Bono, 2001; Solanas et al., 2010). A selection of these is given by name in Table 2.1. Note that most alternatives are based on the original r_1, as can be deduced from the names using 'r' or 'r_1' and a sub- or superscript. In general, the modifications of r_1 showed a smaller bias than r_1 itself, but a larger variability of the estimated autocorrelation (Huitema & McKean, 1991, 1994; Arnau & Bono, 2001; Solanas et al., 2010), except for the estimators r_1^+ and the C-statistic. In direct comparisons between r_1^+ and the C-statistic, it was shown that the C-statistic had a smaller average bias and a smaller average mean square error, thus a smaller variability, over different values of ϕ than the r_1^+ estimation method.

Apart from the modifications of r_1, another closed form solution may be used. The ordinary least squares (OLS) estimator is used in many different applications, most notably in regression analysis. Since the autocorrelation may be interpreted as a special kind of regression parameter, OLS can be used to find the autocorrelation. In comparisons, the OLS estimator showed a smaller bias than most derivations from the r_1 estimators (Huitema & McKean, 1991; Solanas et al., 2010). However, the OLS estimator also showed a slightly larger mean squared error than most r_1 derivations. These comparisons between estimators reveal a bias-variance tradeoff in the autocorrelation estimator.

Two important methods that are not found in the comparisons listed in Table 2.1, are the frequentist MLE and Bayesian MCMC estimation. Though simulation studies using MLE have been done, those studies did not include the conditions of our primary interest. For example, the studies considered different ARMA(p, q)-

models (Stoica, Friedlander, & Söderstorm, 1986; Pantula & Fuller, 1985; Garcia-Hiernaux, Casals, & Jerez, 2009), had no condition with less than 100 time points (Cox & Llatas, 1991) or were aimed at examining other parts of the estimation process, such as deciding on which ARMA(p, q)-model to use (Watson, Clark, McIntyre, & Hamaker, 1992). This was the same for papers using Bayesian MCMC estimation. Examples of this are studies that have no systematic comparison using different estimators (Price, 2012), use AR(2) models (West & Wilcox, 1996) or use lagged cross-correlation (Zhang, Hamaker, & Nesselroade, 2008). The MLE and Bayesian MCMC have become often-used methods of analysis in different fields and applications.

2.3 Estimation methods

In the next paragraphs we will describe the five different estimation methods used in this chapter.

2.3.1 The r_1 estimator in the Yule-Walker method

The Yule-Walker method for ARMA models (Yule, 1927; Walker, 1931; Box & Jenkins, 1976) may be the best known estimation method in time series analysis. It uses the r_1 estimator to estimate the lag 1 autocorrelation:

$$\hat{\phi}_{r1} = \frac{\sum_{t=1}^{T-1} (y_t - \bar{y}) (y_{t+1} - \bar{y})}{\sum_{t=1}^{T} (y_t - \bar{y})^2},$$

where y_t is the observed score at time t, $(t = 1, 2, ..., T)$ and $\bar{y}$ is the mean score over the T observations. Asymptotically, the autocorrelation function for this series is biased by $-(1 + 4\phi)/T$ (Kendall & Ord, 1990). This bias has empirically been shown to be as large as -0.73 for $T = 6$ and $\phi = 0.90$ (DeCarlo & Tryon, 1993). This empirical bias is surprisingly close to the asymptotic bias of -0.77. To keep the bias within reasonable limits, Box and Jenkins (1976, p. 32-33) advise a minimum length of 50 time points for a time series.

The standard error of the $\hat{\phi}_{r_1}$ is calculated as:

$$SE_{r_1} = \sqrt{\frac{\hat{\sigma}_e^2}{(T-1)\hat{\sigma}_y^2}}, \tag{2.3}$$

where $\hat{\sigma}_y^2$ is the estimated variance of y_t and $\hat{\sigma}_e^2$ is the estimated variance of e.

In comparison studies, several other proposals were done to replace the r_1 estimator (Huitema & McKean, 1991, 1994; Young, 1941). One of these, which outperformed the r_1 estimator and some of the other estimators in several studies, was the C-statistic (Young, 1941; DeCarlo & Tryon, 1993; Solanas et al., 2010).

2.3.2 C-statistic

The C-statistic (Young, 1941) compensates the bias of the r_1 estimator by adding a factor to $\hat{\phi}_{r1}$ as:

$$\hat{\phi}_C = \hat{\phi}_{r1} + \frac{(y_T - \bar{y})^2 (y_1 - \bar{y})^2}{2\sum_{t=1}^{T} (y_t - \bar{y})^2}.$$

The $\hat{\phi}_C$ is asymptotically unbiased. The $\hat{\phi}_C$ has been shown to be a better estimator than $\hat{\phi}_{r1}$ for ϕ for short time series and a positive ϕ (DeCarlo & Tryon, 1993; Solanas et al., 2010). However, the bias still remains quite large (e.g., -0.38 for $\phi = 0.60$ and $T = 5$) and the power remains quite low (e.g., ≤ 0.09 for $\phi = 0.60$ and $T = 5$) for short time series (Solanas et al., 2010).

The standard error associated with $\hat{\phi}_C$ is:

$$SE_C = \sqrt{\frac{T-2}{(T-1)(T+1)}}, \tag{2.4}$$

which is obviously only dependent on the number of observations.

2.3.3 Ordinary Least Squares

The Ordinary Least Squares (OLS) for an AR(1) model is:

$$\hat{\phi}_{ols} = \frac{\sum_{t=1}^{T-1} (y_t - \bar{y})(y_{t+1} - \bar{y})}{\sum_{t=1}^{T-1} (y_t - \bar{y})^2}.$$

The asymptotic standard error for $\hat{\phi}_{ols}$ is:

$$SE_{ols} = \sqrt{\frac{T - (T-1)\phi^2 - 1}{T^2 - T - Ty_T^2}}. \tag{2.5}$$

The OLS estimation is capable of handling non-stationary data under certain restrictions. This means that it is possible to obtain a non-stationary estimate (i.e., $|\hat{\phi}_{ols}| > 1$). To identify possible different behaviours, we distinguish two types of OLS analysis results: OLS-A will refer to the complete results, where OLS-S will refer to the results where the non-stationary results are left out.

2.3.4 Maximum Likelihood Estimation

The iterative Maximum Likelihood Estimation (MLE) used to estimate the autocorrelation, shares asymptotic properties with the OLS estimation (Lütkepohl, 1991, p. 368-370). The MLE method uses a collection of algorithms to find the maximum likelihood for a parameter or model (Durbin & Koopman, 2012). In this study, we will compute the MLE with the 'Broyden-Fletcher-Goldfarb-Shanno' algorithm (Byrd, Lu, Nocedal, & Zhu, 1995). An asymptotic standard error for $\hat{\phi}_{mle}$ may be estimated in the same way as for $\hat{\phi}_{r_1}$, using Equation 2.3. The asymptotic bias for an AR(1) model with population mean assumed to be zero, is $-2\phi/T$. For an AR(1) model with the mean estimated, the asymptotic bias is $(-3\phi+1)/T$ (Tanaka, 1984).

2.3.5 Bayesian Markov Chain Monte Carlo

The Bayesian MCMC is the only non-frequentist estimation method considered in this chapter. Bayesian analysis uses a prior probability distribution for the parameters, set up before the analysis. This is combined with the observed likelihood, as computed from the observed data, to form the posterior probability of the parameters. This posterior probability can be expressed through Bayes' theorem: $p(\phi|Y) \propto (Y|\phi)p(\phi)$. For the Bayesian analyses we will use MCMC sampling to find the combination of parameter values which gives the highest likelihood.

In these simulation studies we will consider two weak informative Bayesian priors. Since we assume stationarity we restrict ourselves to prior distributions with non-zero probabilities for $|\phi| \leq 1$. That is, we consider a flat prior, giving all values of ϕ between -1 and 1 an equal probability:

$$\pi_{f(\phi)} = \tfrac{1}{2}, \qquad \text{for } -1 \leq \phi \leq 1.$$

Further, we consider the symmetrized reference prior defined by Berger and Yang (1994), which is specifically tailored to autoregressive processes. The symmetrized reference prior is given as:

$$\pi_{sr(\phi)} = 1/[2\pi\sqrt{1-\phi^2}], \qquad \text{for } -1 \leq \phi \leq 1.$$

This symmetrized reference prior gives a higher probability to higher values of $|\phi|$ and has a narrower posterior distribution and a smaller mean square error than the flat prior or Jeffrey's prior in the case of AR(1) models (Berger & Yang, 1994). We will denote these methods as B_f and B_{sr}, respectively.

2.4 Research design Study 1: Comparison of estimators

To compare the various estimators for the autocorrelation (ϕ), we simulate according to an AR(1) model (see Equation 2.2). In the generation of the data we vary the length of the time series T and the autocorrelation ϕ. For T we use five

Parameter	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7
Priors used							
μ	N(0,2)	N(1,2)	N(0,5)	N(1,5)	N(0,2)	N(0,2)	N(0,2)
σ_e	$\Gamma(2,2)$	$\Gamma(2,2)$	$\Gamma(2,2)$	$\Gamma(2,2)$	$\Gamma(1,1)$	$\Gamma(1,2)$	$\Gamma(2,1)$
Mean estimated parameters and their standard deviation in brackets for $\phi = -0.50$							
B$_f$: ϕ	-0.33 (0.29)	-0.32 (0.29)	-0.31 (0.30)	-0.31 (0.30)	-0.32 (0.29)	-0.34 (0.29)	-0.30 (0.29)
B$_f$: μ	0.00 (0.22)	0.05 (0.23)	0.01 (0.25)	0.03 (0.25)	0.00 (0.22)	0.00 (0.22)	0.01 (0.22)
B$_f$: σ_e	1.06 (0.24)	1.06 (0.24)	1.06 (0.24)	1.06 (0.24)	1.07 (0.26)	0.99 (0.23)	1.16 (0.28)
B$_{sr}$: ϕ	-0.37 (0.34)	-0.37 (0.34)	-0.34 (0.37)	-0.34 (0.37)	-0.37 (0.34)	-0.39 (0.34)	-0.34 (0.34)
B$_{sr}$: μ	0.01 (0.22)	0.07 (0.24)	0.01 (0.27)	0.06 (0.28)	0.00 (0.22)	0.00 (0.22)	0.01 (0.22)
B$_{sr}$: σ_e	1.06 (0.24)	1.06 (0.24)	1.07 (0.24)	1.07 (0.24)	1.08 (0.26)	1.00 (0.23)	1.16 (0.28)
Mean estimated parameters and their standard deviation in brackets for $\phi = 0$							
B$_f$: ϕ	0.05 (0.29)	0.05 (0.30)	0.08 (0.31)	0.08 (0.31)	0.05 (0.29)	0.04 (0.30)	0.07 (0.28)
B$_f$: μ	0.00 (0.32)	0.12 (0.33)	-0.00 (0.40)	0.06 (0.40)	0.00 (0.32)	0.00 (0.33)	0.00 (0.32)
B$_f$: σ_e	1.05 (0.23)	1.06 (0.23)	1.07 (0.23)	1.07 (0.23)	1.07 (0.25)	0.99 (0.22)	1.15 (0.26)
B$_{sr}$: ϕ	0.08 (0.36)	0.09 (0.36)	0.14 (0.38)	0.14 (0.38)	0.08 (0.36)	0.06 (0.36)	0.10 (0.35)
B$_{sr}$: μ	0.00 (0.31)	0.18 (0.33)	0.00 (0.43)	0.13 (0.44)	0.00 (0.31)	0.00 (0.32)	0.00 (0.30)
B$_{sr}$: σ_e	1.07 (0.23)	1.07 (0.23)	1.09 (0.24)	1.09 (0.24)	1.09 (0.25)	1.01 (0.22)	1.17 (0.27)
Mean estimated parameters and their standard deviation in brackets for $\phi = 0.50$							
B$_f$: ϕ	0.38 (0.25)	0.39 (0.25)	0.42 (0.26)	0.42 (0.26)	0.38 (0.25)	0.37 (0.26)	0.38 (0.24)
B$_f$: μ	-0.00 (0.56)	0.23 (0.56)	-0.01 (0.74)	0.13 (0.74)	0.00 (0.55)	0.00 (0.57)	0.00 (0.54)
B$_f$: σ_e	1.02 (0.22)	1.02 (0.22)	1.03 (0.23)	1.03 (0.23)	1.03 (0.24)	0.96 (0.22)	1.11 (0.25)
B$_{sr}$: ϕ	0.46 (0.28)	0.47 (0.28)	0.53 (0.28)	0.53 (0.28)	0.46 (0.28)	0.45 (0.29)	0.47 (0.27)
B$_{sr}$: μ	-0.00 (0.51)	0.34 (0.51)	-0.01 (0.75)	0.26 (0.76)	-0.00 (0.51)	-0.00 (0.53)	-0.00 (0.49)
B$_{sr}$: σ_e	1.03 (0.22)	1.04 (0.22)	1.05 (0.23)	1.05 (0.23)	1.05 (0.24)	0.97 (0.22)	1.12 (0.25)

Table 2.2: Different combinations of priors tested to see their influence on the posterior results, with the used prior distributions (top) and parameters as estimated (with the empirical standard deviation) with these distributions (bottom).

different sizes, namely 10, 25, 40, 50 and 100. For ϕ, we use an autocorrelation of -0.90 to 0.90 inclusive, taking steps of 0.10. Earlier studies show that there is a difference between the bias for the negative and positive ϕ for several estimators, including r_1 and the C-statistic (DeCarlo & Tryon, 1993; Solanas et al., 2010). This indicates that a thorough test is required to include both positive and negative autocorrelations. Finally, the number of replications must be set. All of the studies in Table 2.1 have a minimum of 10,000 replications per condition. However, a pilot study showed that the maximum standard deviation of the mean $\hat{\phi}$ over 5,000 to 10,000 replications was 0.0007, when $T = 10$ and $\phi = 0.7$, for all estimators. Therefore we use $N = 5,000$ replications per condition. Considering a fully crossed experimental design, this yields $19 \times 5 \times 5,000 = 475,000$ simulated data sets.

Across all conditions, μ is set to zero and σ_e^2 to one, which can be done without loss of generality. This results in a standard normal distribution for y_t given ϕ.

Priors We performed a small simulation study to decide on the values for the hyperparameters of the priors in our Bayesian analyses. In the model we use, only the prior distributions for μ and σ_e have such hyperparameters. We used 3 conditions, with $\phi = -0.50,$ 0 and 0.50, using 1,000 replications per condition and

2,000 iterations per analysis. We set $T = 10$, since shorter series provide less data, and will therefore be more strongly influenced by the choice of the prior. For μ we used a normal prior with mean and standard deviation as given, and for σ_e we used a Γ prior with shape and rate as given in the top part of Table 2.2.

As can be seen in Table 2.2, the differences in the estimated parameters are small, especially when taking into account the uncertainty added by the small T. As a result, we based our choice of priors on theoretical grounds. To reduce the influence of the priors, we choose our priors close to the distributions used for the data generation: $\mu \sim N(0, 2)$ and $\sigma_e \sim \Gamma(2, 2)$.

Outcome measures For each data set we obtain different estimators: r_1, C-statistic, OLS, MLE, B$_f$ and B$_{sr}$. To compare the estimators, we consider the bias of the various estimators of ϕ, their empirical standard error, the bias of the estimated standard error, the rejection rate for $\phi = 0$, power for $\phi \neq 0$, and the point and 95% interval estimates of ϕ. All outcome measures are calculated for each condition and each estimation method.

2.4.1 Bias

The bias is computed as:

$$\text{bias} = \left(\frac{1}{N}\sum_{n=1}^{N}\hat{\phi}_n\right) - \phi,$$

where n $(n = 1, 2, ..., N)$ refers to the replication number.

2.4.2 Variability

To compare the variability of the different estimators over the different conditions, we consider two estimators: the empirical standard error and the bias of the estimated standard error. The empirical standard error shows the variability of the $\hat{\phi}$ across replications. The bias of the estimated standard error shows to what extent the standard error estimated by the estimation method, resembles the empirical standard error.

Empirical standard error: $SD(\hat{\phi})$

The empirical standard error of $\hat{\phi}$ is calculated by:

$$SD(\hat{\phi}) = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}\left(\hat{\phi} - \bar{\hat{\phi}}\right)^2},$$

where $\bar{\hat{\phi}}$ is the mean estimated $\hat{\phi}$ over all replications within a condition.

Bias of the estimated standard error

For the frequentist estimators, the estimated standard error $SE(\hat{\phi})$ is calculated using Equations 3, 4 and 5, and for the Bayesian estimation, the estimated standard error is obtained through MCMC. To estimate the expected value of $\overline{SE}(\hat{\phi})$ for each estimator, we compute the average $SE(\hat{\phi})$ over all replications within a condition:

$$\overline{SE}(\hat{\phi}) = \frac{1}{N}\sum_{n=1}^{N} SE(\hat{\phi}).$$

To assess the bias of the estimated standard error with regard to the observed standard error, we substract the observed standard error, $SD(\hat{\phi})$ from the mean estimated standard error, $\overline{SE}(\hat{\phi})$:

$$\text{Bias of } SE(\hat{\phi}) = \overline{SE}(\hat{\phi}) - SD(\hat{\phi}).$$

2.4.3 Rejection rate and power

For each estimation method and condition, we compute the empirical probability (EPr) for rejecting $H_0 : \phi = 0$, with $\alpha = 0.05$. In the condition with $\phi = 0$, the EPr indicates the rejection rate or actual α, in all other conditions the EPr equals the actual power. For the r_1, MLE, OLS-S and C-statistic methods, first a p-value is obtained using a t-distribution. Considering the t-statistic for a correlation coefficient:

$$t_{\text{all}} = \frac{\hat{\phi}\sqrt{T-2}}{\sqrt{1-\hat{\phi}^2}}, \quad df_{\text{all}} = T - 3.$$

For the OLS-A method, a t-test based on the estimated standard error of $\hat{\phi}$ is applied, since the possibility of $\hat{\phi}$ having a higher value than one in absolute value renders the t-statistic for correlations inapplicable:

$$t_{\text{ols}} = \frac{\hat{\phi}}{SE_{\hat{\phi}}}, \quad df_{\text{ols}} = T - 3.$$

For the Bayesian estimation methods, we consider the percentage of datasets for which the 95% credible interval (CrI) does not hold zero.

For each condition and method, we then calculate the EPr of rejecting H_0 :

$\phi = 0$ as:

$$\begin{array}{ll} \text{for } r_1, \text{C-statistic, MLE, OLS-A, OLS-S:} & EPr = \#(H_0 \text{ is rejected})/N, \\ \text{for } B_f, B_{sr}: & EPr = \#(\text{CrI does not hold } 0)/N. \end{array}$$

2.4.4 Point and interval estimates for ϕ

To illustrate the joint effects of bias and variability we consider the two estimation methods with the smallest bias, using the point and interval estimates of ϕ. As point estimate we use the mean of $\hat{\phi}$ per condition, for the interval estimation we use the mean 95 percentile of the $\hat{\phi}$ over all replications per condition.

2.4.5 Procedure

For the simulations and analyses we use the program 'R' (R Core Team, 2015). The C-statistic was computed directly with the basic functions available. For the Yule-Walker, OLS and MLE methods we use the command 'ar' from the software package 'stats'. The Bayesian analyses are done with the program 'Rstan' (Stan Development Team, 2014).

2.5 Results Study 1

The OLS estimator rendered estimates of ϕ that were higher than one in absolute value, and thus non-stationary, as expected. The highest percentage of non-stationary estimates, 15.1%, was found for the shortest series, $T = 10$ and the highest autocorrelation, $\phi = 0.90$. For $T = 10$ and $\phi = -0.90$ to $\phi = 0.80$, up to 6.8% of the estimates per condition were non-stationary, with higher percentages associated with higher values of $|\phi|$. For $T = 25$ to 50 and $\phi = 0.50$ to 0.90 in absolute value, up to 2.3% of the estimates were non-stationary. However, the difference in the results was quite small. Thus we will discuss only the OLS-A results for the OLS, which includes all measurements, unless the OLS-S shows a strong deviation from OLS-A.

For the Bayesian analysis, non-convergence is expressed in the potential scale reduction factor, $\hat{R}$. The potential scale reduction factor shows the ratio of how much the estimation may change when the number of iterations is doubled, with a perfect 1 indicating that no change is expected (Gelman & Rubin, 1992; Stan Development Team, 2014). For each estimated parameter ϕ, μ and σ_e, less than 0.39% of the estimates showed a $\hat{R}$ above 1.02. Furthermore, a maximum of 2.8%, found for μ as estimated with B$_f$, showed a $\hat{R}$ above 1.01.

2.5.1 Bias

The bias of the six estimators as a function of ϕ for $T = 10, 25$, and 50 is presented in Figure 2.1. The conditions for $T = 40$ and $T = 100$ are not shown due to their uninformative nature: $T = 40$ yields results highly similar to $T = 50$, and $T = 100$ yields results with hardly any differences between the estimators. As can be seen in Figure 2.1, the bias becomes smaller as T increases for all methods, which is to be expected. The relation between the bias and ϕ is roughly linear for all methods, being positive for negative values of ϕ and negative for positive values of ϕ. Further, the bias for positive values of ϕ is larger than the bias for their negative counterparts (i.e. $-\phi$). This holds for all values of T and for all methods, except for the C-statistic.

With regard to the ordering of the estimation methods, differences are found between negative and positive values of ϕ and between short time series, $T = 10$, and longer time series, $T \geq 25$. For the shortest time series with $T = 10$, the differences between the methods with regard to bias are strongly dependent on ϕ. For low, negative values of ϕ, the smallest bias is shown by the OLS, MLE and, to a lesser extent, the r_1. For positive values of ϕ, the smallest bias is shown by the B$_{sr}$, followed by the B$_f$. The largest bias for $T = 10$ is associated with the C-statistic for negative values of ϕ, and the r_1 for positive values of ϕ.

For $T \geq 25$ and any ϕ, B$_{sr}$ consistently shows the smallest bias. Just as for the shortest series, the largest bias for $T \geq 25$ is associated with the C-statistic for negative values of ϕ, and with the r_1 for positive values of ϕ.

2.5.2 Variability

With regard to variability, the results for $T \geq 40$ are highly similar to the results for $T = 25$ with regard to pattern of the variability and the order of the estimation methods. The only difference is the decline in absolute size. This prompted us to only explicitly show the results for $T = 10$ and $T = 25$ for the empirical standard error and the bias of the estimated standard error.

Empirical standard error: $SD(\hat{\phi})$

The empirical standard error ($SD(\hat{\phi})$) as a function of ϕ is shown in Figure 2.2 for $T = 10$ (panel a) and $T = 25$ (panel b). For all frequentist estimators, the $SD(\hat{\phi})$ for positive values of ϕ is larger than the $SD(\hat{\phi})$ for their negative counterparts (i.e. $-\phi$), implying that the variability is higher for positive values of ϕ than for negative values of ϕ. For the Bayesian estimators, this differs between values of T and $|\phi|$.

With regard to the ordering of the estimation methods for the $SD(\hat{\phi})$, small differences are found between the short time series, $T = 10$, and longer time series,

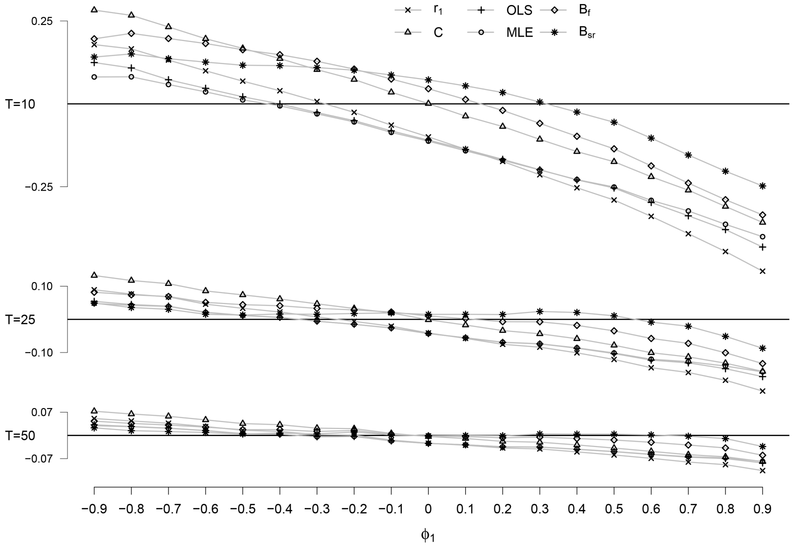


Figure 2.1: Bias for the six estimators and time series lengths $T = 10, 25$, and 50 as a function of ϕ.

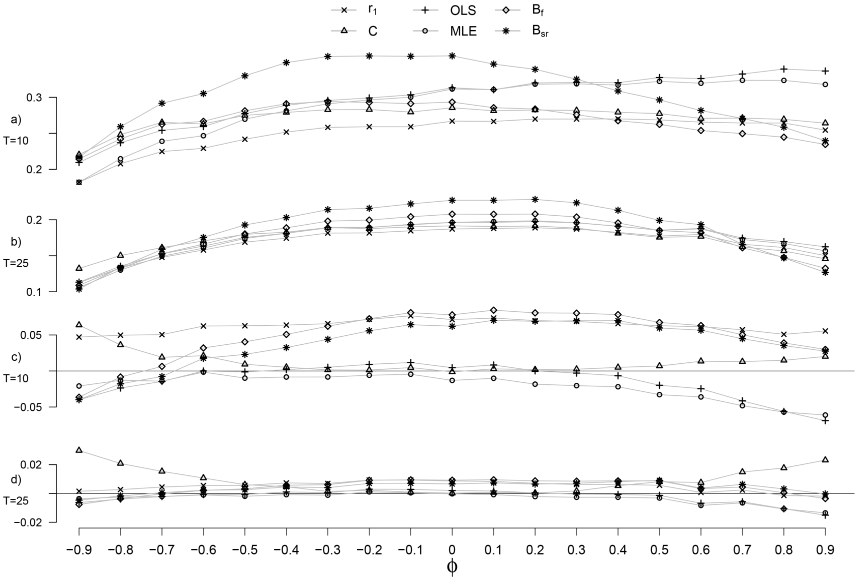


Figure 2.2: The empirical standard error for $T = 10$ (panel a) and $T = 25$ (panel b), and the bias of the estimated standard error for $T = 10$ (panel c) and $T = 25$ (panel d), as a function of ϕ by estimation method.

$T \geq 25$. For the shortest time series, $T = 10$, and ϕ below 0.40, the lowest $SD(\hat{\phi})$ is shown by r_1, for ϕ above 0.40 this is shown by B$_f$. The highest $SD(\hat{\phi})$ for ϕ below 0.30 is shown by B$_{sr}$, for ϕ above 0.30 this is shown by the OLS estimator. The OLS and MLE stand out due to the continuing increase in the $SD(\hat{\phi})$ for higher values of ϕ.

For $T \geq 25$, the B$_{sr}$ shows an distinct pattern. The B$_{sr}$ shows the lowest $SD(\hat{\phi})$ for ϕ below -0.80 and above 0.80, but the highest $SD(\hat{\phi})$ for ϕ between -0.6 and 0.60. The lowest $SD(\hat{\phi})$ for ϕ between -0.70 and 0.40 is shown by the r_1. The highest $SD(\hat{\phi})$ for ϕ below -0.70 is shown by the C-statistic, for ϕ above 0.70 this is shown by the OLS followed by the MLE. When T increases, the empirical standard error of the different methods become smaller and more similar to each other. For $T = 100$, the size of the $SD(\hat{\phi})$ is between 0.05 and 0.10 for all values of ϕ and all estimators.

Bias of the estimated standard error

The bias of $SE(\hat{\phi})$ as a function of ϕ is shown in Figure 2.2 for $T = 10$ (panel c) and for $T = 25$ (panel d). In general, the bias of $SE(\hat{\phi})$ decreases when T becomes larger, indicating a smaller difference between the estimated and the empirical standard errors. For $T = 100$, the bias of $SE(\hat{\phi})$ is between -0.01 and 0.04 for all values of ϕ and all estimators. With regard to the ordering of the estimation methods, small differences are found between $T = 10$ and longer time series. Differences were also found for different values of ϕ.

The direction of the bias of $SE(\hat{\phi})$ differs between the methods and the value of ϕ. For r_1 and the C-statistic, the bias of $SE(\hat{\phi})$ is positive for all ϕ, indicating an overestimation of the standard error. The OLS shows a positive bias of $SE(\hat{\phi})$ for ϕ between -0.70 and 0.20, and a negative bias of $SE(\hat{\phi})$ for other values of ϕ. For the MLE the bias of $SE(\hat{\phi})$ is negative for all ϕ. Both B$_f$ and B$_{sr}$ show a negative bias of $SE(\hat{\phi})$ for $\phi < -0.70$, and a positive bias of $SE(\hat{\phi})$ for higher values of ϕ.

For T=10, the smallest bias of $SE(\hat{\phi})$ for ϕ below -0.70 is shown by the OLS method. The smallest bias of $SE(\hat{\phi})$ for ϕ above -0.50 is shown by the C-statistic, closely followed by the OLS and the MLE. The largest bias of $SE(\hat{\phi})$ for ϕ above -0.80, is shown by r_1, which is joined in this regard by B$_f$ and B$_{sr}$ for ϕ between -0.20 to 0.60.

The bias of $SE(\hat{\phi})$ for $T \geq 25$ is smaller than the bias of $SE(\hat{\phi})$ for $T = 10$ and the different methods are closer together. The domain of ϕ for which the C-statistic shows the largest bias of all estimators increases when T becomes larger; for $T = 10$ this is when ϕ is below -0.50 and above 0.70, for $T = 100$ this is when ϕ is below -0.30 and above 0.20. The other estimators show the same pattern and order in the bias of $SE(\hat{\phi})$ as for $T = 10$.

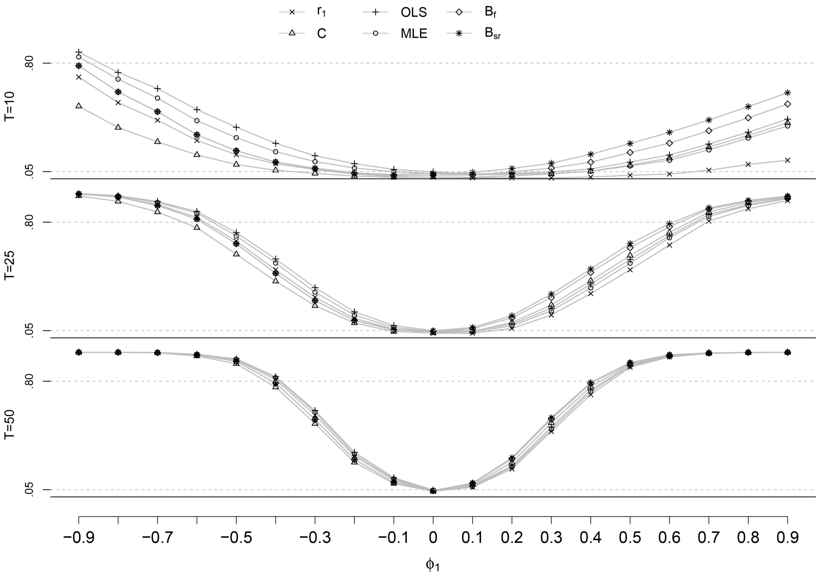


Figure 2.3: Power as a function of ϕ for the different estimation methods and $T = 10, 25$ and 50

2.5.3 Rejection rate and power

The EPr of the different methods is presented in Figure 2.3 for $T = 10$, 25 and 50, where the EPr at $\phi = 0$ indicates the empirical rejection rate and the EPr at $\phi \neq 0$ the empirical power. As with the bias, the EPr for $T = 40$ and $T = 100$ are not shown due to their uninformative nature. When looking at the rejection rate, the empirical rejection rate approaches the nominal α as the length of the time series increases, as to be expected. The rejection rate for $T = 10$ is between 0.01 and 0.04 and for $T \geq 25$ between 0.03 and 0.05, for most estimators. The only exception is the rejection rate for OLS-A at $T = 40$, which is 0.06. At $T = 100$, the MLE, B$_{sr}$, OLS-S and B$_f$ show a rejection rate of 0.050, which is equal to the nominal α of 0.05. For all practical purposes, the difference in rejection rates between estimation methods is negligable.

The power of the estimated ϕ shows a positive relation to the size of T and the absolute value of ϕ, as expected. When we would consider a minimal power of 0.80, for $T = 10$ this is only found for the estimators OLS and MLE, and at very low values of ϕ, i.e. $\phi \leq -0.90$. For $T = 25$ and negative ϕ, the power is above 0.80 for $\phi \leq -0.60$ for all estimators except for the C-statistic, which has a power above 0.80 for $\phi \leq -0.70$; for positive values of ϕ, the B$_{sr}$ shows a power above 0.80 for $\phi \geq 0.60$, for the other estimators this is for $\phi \geq 0.70$. For larger T, the power reaches 0.80 at lower values of ϕ; for $T = 100$, the power is 0.80 for

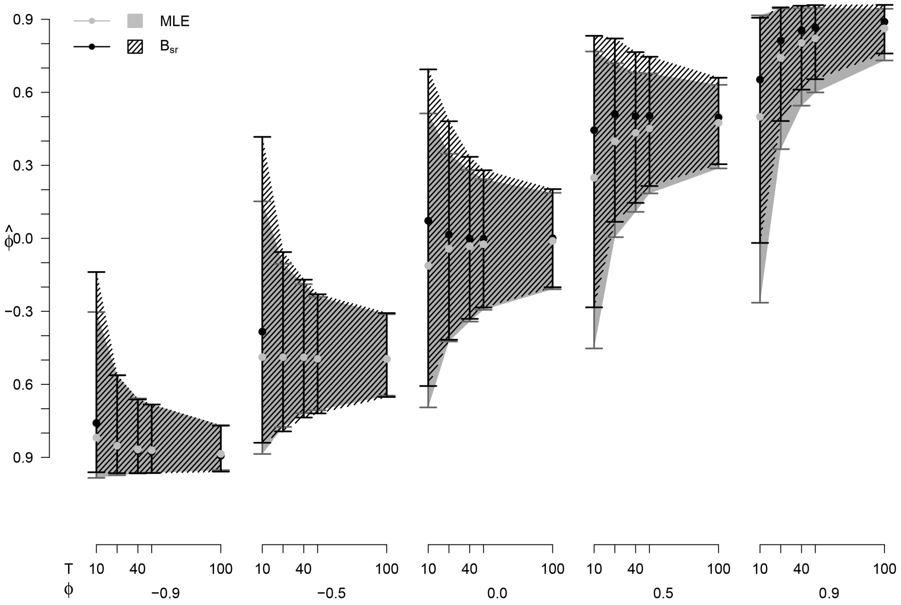


Figure 2.4: Mean of $\hat{\phi}$ (points) with a 95% percentile interval (lines) for different values of ϕ and T.

$|\phi| \geq 0.30$.

The order of the estimation methods with regard to the power is consistent for the different values of T. The highest power for negative ϕ is shown by the OLS, for positive ϕ this is B$_{sr}$. The lowest power for negative ϕ is shown by the C-statistic, for positive ϕ this is r_1. In general, the difference in power between the methods becomes smaller as T becomes larger.

2.5.4 Point and interval estimates for ϕ

The mean $\hat{\phi}$ with a 95% estimation interval for the MLE and B$_{sr}$ estimations can be seen in Figure 2.4. We only present this for the MLE and B$_{sr}$, since these methods show the lowest bias. As can be seen in Figure 2.4, the 95% intervals are larger for smaller values of T, indicating a larger variability in the $\hat{\phi}$, as would be expected. The strongest decrease in both variability and bias is from $T = 10$ to $T = 25$: for B$_{sr}$ the bias decreases with up to 89% and the variability with 29% to 51%, for the MLE the bias decreases by 82% and the variability by 34% to 51%.

2.5.5 Conclusion

In general, it may be concluded that the Bayesian B_{sr} and the frequentist MLE perform best in terms of bias, but not in terms of variability. With regard to empirical variability, the r_1 performs best. For the bias of the estimated variability, the MLE performs best. Furthermore, the B_{sr} is favorable with regard to power for positive ϕ, showing only slight differences with the OLS estimator for a negative B_{sr}. This leads us to continue with the MLE and the B_{sr} estimators for the misspecification study of this chapter.

2.6 Research Design Study 2: Robustness

To study the robustness of the estimation methods, we misspecify the model. The data is still analysed as if they stem from an AR(1) model, but we generate the data using an ARMA(1,1) model. We generate data sets for two different sizes of T, namely 25 and 50. For ϕ and θ, we use parameters of -0.90 to 0.90 inclusive, taking steps of 0.15. Every condition consists of 5,000 replications. Considering a fully crossed design, this yields $13 \times 13 \times 2 \times 5,000 = 1,690,000$ datasets.

Again, across all conditions, μ is set to zero and σ_e^2 to one. We consider the same outcome measures for Study 2 as we did for Study 1.

2.7 Results Study 2

We successively present the results on the bias, empirical standard error, bias of the estimated standard error, rejection rate, power, and point and 95% interval estimates. Note that when θ is zero, the simulated data follows an AR(1) model, rendering the results equal to the results discussed in the first study of this chapter, apart from small deviations resulting from simulation variability.

As with the first study, we checked the $\hat{R}$ of the estimated parameters ϕ, μ and σ_e of B_{sr}. For each of the parameters, less than 0.14% showed an $\hat{R}$ above 1.02, and less than 1.69% showed an $\hat{R}$ above 1.01.

2.7.1 Bias

In Figure 2.5, heatmaps for the bias of B_{sr} and MLE for $T = 25$ and $T = 50$ are presented, expressing the bias depending on the combination of ϕ and θ. The θ influences the bias in two ways: first, the bias is smaller when θ is close to zero, second, the bias gets larger when θ is further from ϕ.

The bias is also influenced by the value of T and the estimation method. When looking at T, in the MLE the bias for $T = 50$ is larger than the bias for $T = 25$, unless both θ and ϕ are negative. The difference between the bias of $T = 50$ and

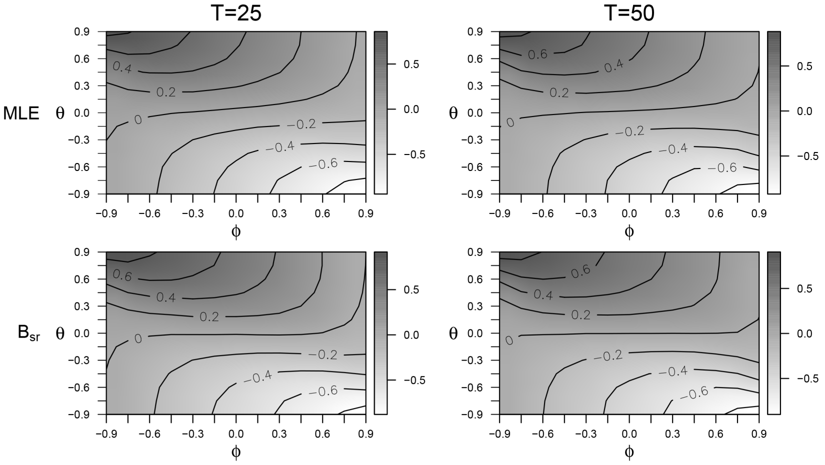


Figure 2.5: Heatmaps for the bias of $\hat{\phi}_{mle}$ for T=25 and T=50 (top panes) and the bias of $\hat{\phi}_{B_{sr}}$ for T=25 and T=50 (bottom panes).

the bias of $T = 25$ ranges for MLE from -0.03 to 0.12 per condition. For the B$_{sr}$, the bias for $T = 50$ is smaller than for $T = 25$, unless ϕ has a value above 0.30. The difference between the bias of $T = 50$ and $T = 25$ ranges for B$_{sr}$ from -0.07 to 0.03 per condition. Comparing the estimation methods reveals that the bias is small, and in general slightly larger for the B$_{sr}$ than for the MLE, with the largest difference being 0.11 for $\phi = 0.60$, $\theta = 0$, and $T = 25$. The difference between the estimation methods is larger for $T = 25$ than for $T = 50$.

2.7.2 Variability

Close inspection of the results for the variability and EPr for the B$_{sr}$ and MLE estimators and the two lengths of T, revealed that the patterns are very similar across methods and different lengths of T. This prompted us to only present the results of B$_{sr}$ and $T = 25$ in Figure 2.6. However, we discuss any quantitative differences between the methods. For comparison purposes, we also plotted the $SD(\hat{\phi})$, the bias of $SE(\hat{\phi})$ and the EPr for B$_{sr}$ and $T = 25$ of Study 1.

Empirical standard error: $SD(\hat{\phi})$

The empirical standard error ($SD(\hat{\phi})$), for B$_{sr}$ with $T = 25$ and $\theta = -0.45, 0.00,$ and 0.45, can be seen in Figure 2.6 (panel a). Some differences between the $SD(\hat{\phi})$ over different values of θ, ϕ and T are found. First, the $SD(\hat{\phi})$ shows a positive slope over ϕ for negative values of θ, and a negative slope over ϕ for positive values of θ. Second, the $SD(\hat{\phi})$ is smaller for $T = 50$ compared to $T = 25$. For the MLE

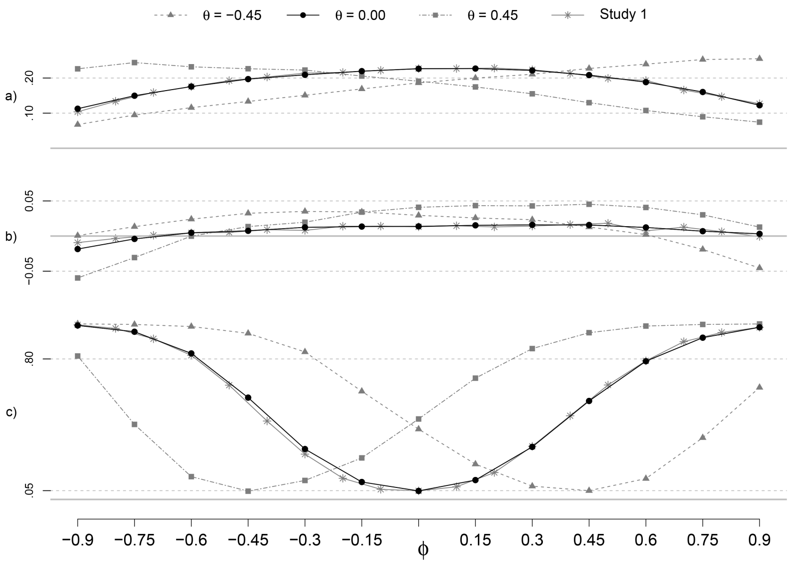


Figure 2.6: Empirical standard error (panel a), bias of the estimated standard error (panel b) and EPr (panel c) for B$_{sr}$ with T=25 as a function of ϕ.

estimator the $SD(\hat{\phi})$ is up to 0.07 smaller for $T = 50$, for the B$_{sr}$ the $SD(\hat{\phi})$ is up to 0.08 smaller for $T = 50$. When comparing methods of estimation, the $SD(\hat{\phi})$ of the B$_{sr}$ is larger than the $SD(\hat{\phi})$ of the MLE, for both values of T. For $T = 25$, this differs up to 0.03, for $T = 50$ this differs up to 0.01 per condition.

Bias of the estimated standard error

The bias of $SE(\hat{\phi})$ for B$_{sr}$ with $T = 25$ and $\theta = -0.45, 0.00$, and 0.45, can be seen in Figure 2.6 (panel b). The bias of $SE(\hat{\phi})$ for most combinations of ϕ and θ, where $\theta \neq 0$, is positive and higher than the bias for the AR(1) data. Only for low values of $|\theta|$ combined with high values of $|\phi|$, the bias of $SE(\hat{\phi})$ is negative. The bias of $SE(\hat{\phi})$ is larger for $T = 25$ than for $T = 50$, for all methods and conditions, with differences up to 0.02 for both methods. Furthermore, the bias of $SE(\hat{\phi})$ is slightly larger for the B$_{sr}$ estimation than for the MLE, with a maximum absolute difference of 0.01 for $T = 25$ and 0.03 for $T = 50$.

2.7.3 Rejection rate and power

The EPr for B$_{sr}$ with $T = 25$ and $\theta = -0.45, 0.00$, and 0.45, can be seen in Figure 2.6 (panel c). When $\theta \neq 0$, the curve of the EPr shifts relative to the curve of the AR(1) data. For a negative θ, the curve shifts to the right, for a positive θ,

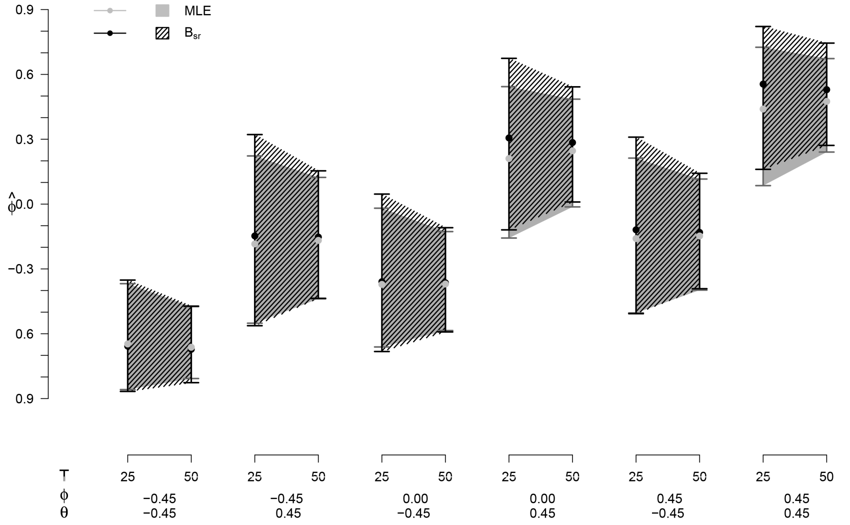


Figure 2.7: Mean of $\hat{\phi}$ (points) with a 95% percentile interval (lines) for different values of ϕ and T.

the curve shifts to the left. In both cases, the amount the curves shifts is roughly equal to the absolute value of θ. This is the same for both methods, with the shape of the curve being dependent on T, as in Study 1. The differences between the methods are negligible.

2.7.4 Point and interval estimates for ϕ

The mean $\hat{\phi}$ with a 95% estimation interval for the MLE and B$_{sr}$ estimations are presented in Figure 2.7. As can be seen in Figure 2.7, the 95% intervals are larger for smaller values of T, indicating a larger variability in the $\hat{\phi}$. For both methods, the decrease in the 95% estimation interval is between 33% and 44% from $T = 25$ to $T = 50$ for the different conditions. In most conditions, the 95% estimation interval is larger and the mean $\hat{\phi}$ is slightly higher for the B$_{sr}$ than for the MLE.

2.7.5 Conclusion

We found that the further the θ deviates from zero, the larger the difference between the ϕ and $\hat{\phi}$ is. The B$_{sr}$ shows a larger bias than the MLE when θ is further from zero, showing a larger influence of the θ parameter in the estimated ϕ. Furthermore, the observed variability is slightly smaller for the MLE, with the difference between B$_{sr}$ and MLE being larger for $T = 25$ than for $T = 50$.

2.8 Discussion

We compared five estimation methods for the autocorrelation: the r_1, C-statistic, ordinary least squares, maximum likelihood estimation and Bayesian MCMC estimation. For the Bayesian MCMC estimation we used both a flat prior and a symmetrized reference prior, giving a total of six autocorrelation estimators. We compared these estimators with regard to bias, variability, rejection rate, power, and point and 95% estimation interval estimates. After this comparison, we selected the Bayesian estimation with symmetrized reference prior and the maximum likelihood estimator to use in a second study. In this small study we assessed the robustness of the methods against underspecification.

The results we found in the first study largely complied with the results from previous studies. For the bias for positive values of ϕ, we found the bias of the C-statistic and the OLS to be smaller than the bias of r_1, as did previous studies (DeCarlo & Tryon, 1993; Huitema & McKean, 1991; Solanas et al., 2010). For the empirical standard error, we found smaller values for r_1 than for OLS, as did Huitema and McKean (1991). The low rejection rate we found for the r_1 and the C-statistic confirms to earlier studies (Huitema & McKean, 1991; DeCarlo & Tryon, 1993; Solanas et al., 2010). The power we found for the C-statistic is similar to the power found by Arnau and Bono (2001). However, the results of Solanas et al. (2010) with regard to power were only partly confirmed by our study: for negative ϕ we indeed found a higher power for OLS followed r_1, than for the C-statistic. But for positive ϕ, we found a higher power for the C-statistic than for r_1.

The first study showed a strong improvement in all measures for all methods between $T = 10$ and $T = 25$. This improvement continued, be it not as strong, for higher values of T. When comparing methods, B_{sr} showed the smallest bias. For the frequentist methods, this was MLE followed by the C-statistic. The smallest empirical standard error is shown by r_1, the smallest bias of the estimated standard error is shown by the C-statistic, the OLS and the MLE. We further found that a small bias is often paired with a high empirical standard error. The power was rather low for all methods at the lengths of time series we considered. For $T = 25$, the power is below 80% for all methods for ϕ between -0.5 and 0.5, for $T = 100$, the power is below 80% for ϕ between -0.2 and 0.2. The differences between methods with regard to power are negligible. In research areas where effect sizes are small, this may pose a problem. Some studies use moving windows to assess the stability of parameter estimates over time. For these moving windows, these results indicate that a moving window of at least 50 time points is advisable, especially when the differences in parameters over time are small.

The first study was conducted to explore the differences between estimation methods for the autocorrelation in a single subject design. However, this is only a small step in a large research area. The next step may be to explore these results in multilevel or group analyses, thus when there is not one but multiple subjects per dataset. Another issue may be how the different methods respond to non-stationary data, i.e., $|\phi| > 1$.

In the second study, the robustness of the MLE and B_{sr} to underspecification was examined. In general, we confirmed the notion that the further the unmodelled parameter is from zero, the larger the influence of this parameter is on $\hat{\phi}$ (Tanaka & Maekawa, 1984). As with the first study, the empirical standard error decreased when T became larger. However, the bias reacted differently for both methods: for the MLE, the bias became slightly smaller for most conditions, where the bias of B_{sr} became slightly larger for a larger T. The difference in performance for all measurements between the MLE and the B_{sr} was small for both values of T. It was shown that the bias, variability, rejection rate and power were all highly dependent on the value of the non-modeled parameter in the data, θ. This can be related to the fact that the autocorrelation of the error also has an influence on the autocorrelation of the total score.

The robustness study was rather small and specific, looking into only one possible way to misspecify the ARMA(1,1) model. More options within misspecification should be explored to find how robust the estimation methods are with regard to under-, over- and misspecification. Important points are the influence of a misspecified error distribution or overspecification of the model.

In conclusion, we found that the best performing methods for autocorrelation estimation are the Bayesian estimator with symmetrized reference prior and the maximum likelihood estimator. The difference in performance between these two is, for all practical purposes, negligible. The results for the measurements improving greatly between $T = 10$ and $T = 25$ and continue to do so, but in a less spectacular fashion. For the misspecification study, we found the size of θ, the non-modelled parameter, to be vital for the performance of the estimation methods. The differences between lengths of the series and estimation methods was of lesser influence on the results.

Chapter 3

Comparison of Estimation Procedures for Multilevel AR(1) Models

Abstract

To estimate a time series model for multiple individuals, a multilevel model may be used. In this chapter we compare two estimation methods for the autocorrelation in Multilevel AR(1) models, namely Maximum Likelihood Estimation (MLE) and Bayesian Markov Chain Monte Carlo. Furthermore, we examine the difference between modeling fixed and random individual parameters. To this end, we perform a simulation study with a fully crossed design, in which we vary the length of the time series (10 or 25), the number of individuals per sample (10 or 25), the mean of the autocorrelation (-0.6 to 0.6 inclusive, in steps of 0.3) and the standard deviation of the autocorrelation (0.25 or 0.40). We found that the random estimators of the population autocorrelation show less bias and higher power, compared to the fixed estimators. As expected, the random estimators profit strongly from a higher number of individuals, while this effect is small for the fixed estimators. The fixed estimators profit slightly more from a higher number of time points than the random estimators. When possible, random estimation is preferred to fixed estimation. The difference between MLE and Bayesian estimation is nearly negligible. The Bayesian estimation shows a smaller bias, but MLE shows a smaller variability (i.e., standard deviation of the parameter estimates). Finally, better results are found for a higher number of individuals and time points, and for a lower individual variability of the autocorrelation. The effect of the size of the autocorrelation differs between outcome measures.

This chapter is published as: Krone, T., Albers, C. J., & Timmerman, M. E. (2016d). Comparison of estimation procedures for multilevel AR(1) models. *Frontiers in Psychology*, *7*(486). doi:10.3389/fpsyg.2016.00486

3.1 Introduction

The electronic revolution allows for new and exciting research possibilities. One such possibility that has become increasingly easy to use is ecological momentary assessment (c.f., Shiffman et al., 2008; Bos et al., 2015) through electronic devices such as the mobile phone. This advancement allows, with little hassle for the individuals, multiple measurements per individual per day at the researcher's discretion (Bolger & Laurenceau, 2013). The data provided through ecological momentary assessment, often denoted as intensive longitudinal data (Hamaker, Ceulemans, Grasman, & Tuerlinckx, 2015), give ample opportunities for studying complex processes, involving the trends and dynamics of human behavior and experience. The latter pertains to studying how aspects of behavior and/or experience evolve across time, and how aspects mutually influence each other. Using these kinds of data, studies have been done pertaining to, for example, emotional complexity and age (Brose, de Roover, Ceulemans, & Kuppens, 2015), dynamics of depression (Kuppens, Allen, & Sheeber, 2010; Kashdan & Farmer, 2014; Erbas, Ceulemans, Pe, Koval, & Kuppens, 2014), and the relation between affect and stress (Scott, Sliwinski, Mogle, & Almeida, 2014).

Intensive longitudinal data of several individuals fall under the category of multilevel data. Multilevel data are collected according to a nested sampling design, resulting in data with a hierarchical structure (e.g., Snijders & Bosker, 1999; Hox, 2010). A two-level example is univariate longitudinal data of multiple individuals, where the time points at level 1 are nested within the individuals at level 2. In psychological sciences, momentary assessment data pertain to longitudinal series of limited length, collected among a limited number of individuals, creating a multilevel data set. To analyze these data, one can use multilevel models. In the analysis of longitudinal data, we can discern two different focuses: the trend and the dynamics across time. To study the trend across time, a multilevel regression model for repeated measures can be used. Herewith, one could use either dummy-variables (also known as indicator or design variables), to indicate effects pertaining to each time point, or time itself as a predictor (e.g., Snijders & Bosker, 1999).

To study the dynamics across time, a model is needed for describing the relationships between scores at successive measurements. This is typically done using an autoregressive model (Box & Jenkins, 1976). The simplest variant is an autoregressive model of the first order, the AR(1) model for short. For multilevel data with multiple individuals, Suls, Green, and Hills (1998) used an autoregressive component in a multilevel model with random coefficients to assess change in mood over time. In this model, the autoregressive parameter was composed of a population parameter, a parameter dependent on the predictor neuroticism, and a subject dependent noise parameter. The same approach was used by Kuppens, Allen, and Sheeber (2010), who included self-esteem as a predictor for the autocorrelation. Both authors interpreted the autoregressive parameter as reflecting the degree of inertia, which is the tendency to retain the status-quo over time. An often encountered problem in time series analysis is the violation of the assumption of independent errors, due to autocorrelated noise. To account for this effect, a multilevel model including autocorrelated noise was proposed by Goldstein, Healy, and Rasbash (1994). Note that although Goldstein et al. (1994) denote this model as an autoregressive model, it is actually a moving average model, according to the common terminology (Box & Jenkins, 1976).

At the moment, it is unclear how efficient the estimation methods of different multilevel model variants are for intensive longitudinal data. Several simulation studies have been conducted to compare the different estimators for single case AR(1) models (Krone et al., 2016a; Huitema & McKean, 1991; DeCarlo & Tryon, 1993; Arnau & Bono, 2001; Solanas et al., 2010). While the empirical standard error is lowest for the classical estimation method denoted by r_1 (Walker, 1931), the bias is lowest for iterative estimators (Krone et al., 2016a). For all methods, the empirical power is low for series with less than 50 time points. For a true autocorrelation below $|0.40|$, the power is below 80% for all compared estimation methods (Krone et al., 2016a). This is consistent with the advice of a lower bound of 50 time points for any time series modelled with an AR(1) model, as given by Box and Jenkins (1976).

In this chapter, we focus on the AR(1) model in a multilevel setting, for relatively short time series and numbers of individuals. We do so because these characteristics are typical for intensive longitudinal data, and the properties of multilevel AR(1) model estimators have been investigated scarcely. Furthermore, the inclusion of multiple individuals may have a profound effect on the bias, variability and power of the estimators. In a recent paper, Jongerling, Laurenceau, and Hamaker (2015) compared the Maximum Likelihood Estimation (MLE) and the Bayesian multilevel AR(1) estimators. Their simulation design included manipulations of the intercept variance and of the covariance between the autocorrelation and the error variance. However, their design lacked manipulations of the mean and variance of the autocorrelation, central to the current chapter. Further, they only used person centering in MLE models and only used a random effect for the error variance in the Bayesian model, which means that their design was not fully-crossed. Jongerling et al. (2015) concluded that the estimation may be improved by including a random effect for the error variance and by refraining from person-centering. The differences in bias they found are small and inconsistent; in certain conditions increasing sample size and time series length also seems to increase rather than decrease the bias. As such, their model estimates may be biased. While they raise an interesting point with regard to individual error variances and person-centering the data, we will first consider a more basic comparison between estimation methods using the same model.

For multilevel models, several of the estimation methods used in single subject designs are unavailable. Two closed form estimators that can be used for multilevel models are generalized least squares (GLS) and generalized estimation equations (GEE) (Liang & Zeger, 1986). Although these methods have the benefit of being faster than iterative methods, i.e., MLE and Bayesian Markov-Chain Monte Carlo (Bayesian MCMC), the resulting estimates show bias and need a large amount of data points to achieve an acceptable standard error. (Hox, 2010). Better fitting estimators for the ML-AR(1) model are iterative estimators, specifically the MLE and the Bayesian MCMC estimation (Hox, 2010). In an earlier study, we also found this for single subject data, which leads us to use MLE and Bayesian MCMC in this chapter (Krone et al., 2016a).

In this chapter we use a simulation study to quantify the differences between two model variants for multilevel autocorrelated data, and between two estimation methods, being MLE and Bayesian MCMC estimation. In the next part of this chapter, we discuss the multilevel model and the estimation methods. This is followed by an explanation of the simulation study design, the results of the simulation study, and a discussion on the implications for designing empirical studies involving intensive longitudinal data and properly modeling the resulting data.

3.2 The Multilevel Autoregressive Lag 1 Model

The ML-AR(1) model we use is a random coefficients model (e.g., Snijders & Bosker, 1999; Hox, 2010). The model has two levels: the first level holds the time points, as the second level holds the individuals. The level 1 model is based on the AR(1) model for a single individual (Box & Jenkins, 1976):

$$y_{t,n} = \mu_n + \phi_n(y_{t-1,n} - \mu_n) + e_{t,n}, \qquad e_{t,n} \sim N(0, \sigma_e), \tag{3.1}$$

where $y_{t,n}$ is the score of individual n ($n = 1, 2, ..., N$) at time t ($t = 1, 2, ..., T$), μ_n the intercept, ϕ_n the autocorrelation, and $e_{t,n}$ is the error term. The error terms follow a normal distribution with mean zero and standard deviation σ_e and are independent of each other and of the observations $y_{t,n}$.

In this chapter we compare two ways of modeling multilevel data: the random model and the fixed model. The difference between these models is based on the theory behind the sampling of individuals, and is expressed in the level 2 model. In the random model, as used in the random coefficients approach, the individuals are assumed to be drawn randomly from a certain population. As such, the parameters of the individuals are assumed to be drawn randomly from the population distribution of the parameter concerned. It is common, but not required, to assume a normal distribution for the individual parameters. We will use the normality assumption in this chapter.

The fixed model makes no assumption with regard to the sampling of the individuals. To reflect this, the parameters of the fixed model are estimated freely. This implicitly defines the level 2 model, as the joint distribution of the individually estimated parameters for all individuals is hereby defined. Due to the free parameter estimation, these model estimates would be the same as when the time series of each individual were modeled separately. This implies that the standard deviation of the error is $\sigma_{e,n}$, and hence may vary across individuals.

For the random model, a level 2 model must be defined which captures the assumed population distributions of the parameters. The level 2 model we use is:

$$\mu_n = \gamma_{0,0} + U_{0,n}, \tag{3.2}$$

$$\phi_n = \gamma_{0,1} + U_{1,n}, \tag{3.3}$$

with:

$$U_{0,n} \sim N(0, \sigma_{U_{0,n}}), \tag{3.4}$$

$$U_{1,n} \sim N(0, \sigma_{U_{1,n}}). \tag{3.5}$$

where $\gamma_{0,0}$ is the population intercept, $U_{0,n}$ is the individual specific deviation from the population intercept for individual n, $\gamma_{0,1}$ is the population autocorrelation and $U_{1,n}$ is the individual specific deviation from the population autocorrelation. Note that the standard deviation of the error, σ_e, is assumed to be equal across the population of individuals (unlike the fixed model), and independent of both $U_{0,n}$ and $U_{1,n}$. The composite model, expressing both levels in one model, is:

$$\begin{aligned} y_{t,n} = {} & \gamma_{0,0} + \gamma_{0,1}(y_{t-1,n} - \gamma_{0,0} - U_{0,n}) \\ & + U_{0,n} + U_{1,n}(y_{t-1,n} - \gamma_{0,0} - U_{0,n}) + e_{t,n}, \qquad e_{t,n} \sim N(0, \sigma_e). \end{aligned} \tag{3.6}$$

3.2.1 Estimation methods

MLE

For MLE, the distinction can be made between full maximum likelihood (FML) and restricted maximum likelihood (RML, also known as REML). The difference lies in how the likelihood is estimated: FML includes both the regression coefficients and the variance components in the likelihood, whereas RML only includes the variance components. The regression coefficients for RML are estimated in a secondary step (Hox, 2010). In general, the FML is easier to calculate. Furthermore, the FML allows for an overall chi-square test for two models that differ in the fixed part, which the RML generally does not. However, when estimating the variance, the FML model is biased since it does not take into account the number of fixed parameters (Bryk & Raudenbush, 1992, p. 46), while the RML has asymptotically unbiased variance estimates.

For the random model using MLE (henceforth denoted as MLE-R), we will use RML (Harville, 1977) with the 'Bound Optimization BY Quadratic Approximation' algorithm (Powell, 2009). The method we use estimates the random parameters under the assumptions of normality, in line with typical applications in social sciences (Hox, 2010; Goldstein, 2011). The multilevel implementation of the MLE we use is not specifically made for autocorrelation measures, and may thus produce non-stationary autocorrelation values, i.e., $|\hat{\phi}_n| > 1$. The number of non-stationary results obtained will be touched upon in the results section.

For the fixed model using MLE (henceforth denoted as MLE-F), we will use the 'Broyden-Fletcher-Goldfarb-Shanno' algorithm (Byrd et al., 1995). The estimation method we use is especially programmed for autocorrelation estimation and, as such, produces stationary autocorrelation estimates. For both MLE approaches, the algorithm may fail to reach convergence. The number of non-convergent results will be touched upon in the results section. Furthermore, both MLE approaches are unable to handle missing data, other than by removing the whole case from the analysis. To retain the data, an Expectation-Maximization algorithm (Dempster, Laird, & Rubin, 1977), also used in latent variable modeling, may be used. However, in this chapter we will assume that the full data is available.

Bayesian MCMC

Estimation through Bayesian MCMC is very versatile with respect to the models and distributions that can be estimated. The MCMC-method we use for both the fixed and random (denoted as BAY-F and BAY-R, respectively) Bayesian estimators is Hamiltonian Monte Carlo (HMC), a generalization of the Metropolis-Hastings algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Hastings, 1970) that allows for an efficient estimation of the parameters (Gelman, Carlin, et al., 2013). An added advantage of the Bayesian approach is the possibility to deal with missing data optimally, i.e. without casewise deletion. For AR(1) models it is possible to apply the autoregressive model on the estimated score of the missing time point, instead of on the observed score itself. This allows the estimation to continue past the missing data points, adjusting the estimation as soon as the next time point is observed again.

3.2.2 Procedure

In this study, we aim to examine the comparative quality of MLE and Bayesian MCMC estimation for the autocorrelation parameter in random and fixed ML-AR(1) models. This results in four estimators which will be compared: MLE-F, MLE-R, BAY-F and BAY-R. For the Bayesian MCMC estimations, we use the

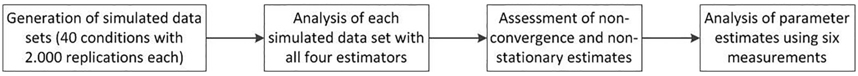

Generation of simulated data sets (40 conditions with 2.000 replications each)
Analysis of each simulated data set with all four estimators
Assessment of non-convergence and non-stationary estimates
Analysis of parameter estimates using six measurements

Figure 3.1: Flowchart of the study design of the simulation study.

program Rstan (Stan Development Team, 2014). For the estimation of the MLE-R, we use the package lme4 for R (Bates, Maechler, Bolker, & Walker, 2015). All other analyses, including data generation, are done using the functions available in the base installation of the program R (R Core Team, 2015).

3.3 Simulation study

3.3.1 Simulation design

To compare the four estimators for the autocorrelation, we set up a simulation design with data generation, data analyses, assessment of computational issues and analyses of the results as shown in Figure 3.1, with 40 conditions in total. The conditions stem from a fully crossed experimental design, including the following factors, with number of factor levels between parentheses: the length of the time series T (2), the number of individuals per dataset N (2), the standard deviation $\sigma_{U_{1,n}}$ (2), and mean $\gamma_{0,1}$ (5) of the autocorrelation distribution, as used in Equations 3.5 and 3.3. Both T and N are either 10 or 25, $\sigma_{U_{1,n}}$ is either 0.25 or 0.40, $\gamma_{0,1}$ is set from -0.60 up to 0.60 inclusive, taking steps of 0.30 for the values in between.

The time series were generated according to Equation 3.6. The mean and standard deviation of the error of each series in each replication is set to zero and one, respectively. The values of ϕ_n were then drawn from a truncated and rescaled normal distribution with range -1 to 1, to ensure the resulting time series were stationary:

$$\phi_n \propto N(\gamma_{0,1}, \sigma_{U_{1,n}})\tau(-1, 1). \tag{3.7}$$

Parameter priors

We performed a small simulation study to examine the sensitivity for the choice of the hyperparameters of the priors of our Bayesian model. We considered $2,000$ replications of a single simulation condition, using $5,000$ iterations, taking $\gamma_{0,0} = 0.00$, $\sigma_{U_{1,n}} = 0.40$, $T = 10$ and $N = 10$ (see Equation 3.6). This condition is one where the prior is expected to have the most influence, due to the high variability across individuals and the small amount of data. The prior we use for $\hat{\gamma}_{0,1}$ for BAY-R and $\hat{\phi}_n$ for BAY-F is Berger's symmetrized reference prior (Berger & Yang,

	Fixed model		Random model			
Test	μ_n	σ_e	μ	σ_μ	σ_e	σ_ϕ
1	$N \sim (0,2)$	$\Gamma \sim (2,2)$	$N \sim (0,2)$	$\Gamma \sim (2,2)$	$\Gamma \sim (2,2)$	$\Gamma \sim (2,2)$
2	$N \sim (0,5)$	$\Gamma \sim (2,2)$	$N \sim (0,5)$	$\Gamma \sim (2,2)$	$\Gamma \sim (2,2)$	$\Gamma \sim (2,2)$
3	$N \sim (1,2)$	$\Gamma \sim (2,2)$	$N \sim (1,2)$	$\Gamma \sim (2,2)$	$\Gamma \sim (2,2)$	$\Gamma \sim (2,2)$
4	$N \sim (1,5)$	$\Gamma \sim (2,2)$	$N \sim (0,2)$	$\Gamma \sim (1,1)$	$\Gamma \sim (2,2)$	$\Gamma \sim (2,2)$
5	$N \sim (0,2)$	$\Gamma \sim (1,1)$	$N \sim (0,2)$	$\Gamma \sim (1,2)$	$\Gamma \sim (2,2)$	$\Gamma \sim (2,2)$
6	$N \sim (0,2)$	$\Gamma \sim (1,2)$	$N \sim (0,2)$	$\Gamma \sim (2,2)$	$\Gamma \sim (1,1)$	$\Gamma \sim (2,2)$
7	$N \sim (0,2)$	$\Gamma \sim (2,1)$	$N \sim (0,2)$	$\Gamma \sim (2,2)$	$\Gamma \sim (1,2)$	$\Gamma \sim (2,2)$
8			$N \sim (0,2)$	$\Gamma \sim (2,2)$	$\Gamma \sim (2,2)$	$\Gamma \sim (1,1)$
9			$N \sim (0,2)$	$\Gamma \sim (2,2)$	$\Gamma \sim (2,2)$	$\Gamma \sim (1,2)$

Table 3.1: Different combinations of priors tested to see their influence on the posterior results

1994), which has shown to better perform than the flat prior for single case AR(1) models (Krone et al., 2016a). This prior does not need hyperparameters.

We tested several hyperparameters for the prior distributions of μ_n and σ_e for the fixed model, and $\gamma_{0,0}$, $\sigma_{U_{0,n}}$, σ_e and $\sigma_{U_{1,n}}$ of the random model, as shown in Table 3.1. Our parameter of primary interest, $\hat{\gamma}_{0,1}$, showed small differences across the various tests. For the random model, the estimates ranged from 0.017 (test 5) to 0.026 (test 9). For the fixed model, the estimates ranged from 0.033 (test 6) to 0.109 (test 4).

For the random estimator, the estimated $\gamma_{0,0}$ showed small differences across the various priors, resulting in estimates ranging from 0.000 (test 9) to 0.004 (test 3). For the fixed estimator, the estimated $\bar{\mu}_n$ ranged from 0.000 (test 7) to 0.192 (test 2). The effect of the different priors is most notable for the posterior of the parameter for which the prior was changed. For the simulation study, we use the priors of test 1 of Table 3.1, as these gave the best results.

Number of iterations

A preliminary study was performed to decide on the number of iterations needed for the Bayesian MCMC. Because of the more complicated model of BAY-R compared to BAY-F, we only tested the number of iterations for BAY-R. Ten datasets per condition were used to find the convergence rate as expressed through the potential scale reduction factor $\hat{R}$, as can be seen in Table 3.2. The potential scale reduction factor shows the ratio of how much the estimation may change when the number of iterations is doubled, with a value of 1 indicating that no change is expected (Gelman & Rubin, 1992; Stan Development Team, 2014). We deemed the improvements brought by a higher number of iterations negligible, thus we continued using 3,000 total iterations, of which 1,500 were burn-in.

Iterations		mean	Percentage of $\hat{R}$ above:			
Total	burn-in	$\hat{R}$	1.05	1.1	1.5	1.7
3,000	1,500	1.01	2.53	0.89	0.02	0.00
4,000	2,000	1.01	1.97	0.68	0.02	0.00
10,000	5,000	1.00	1.54	0.75	0.01	0.00
10,000	8,000	1.01	2.13	0.70	0.01	0.00
Final analyses: 3,000 iterations with 1,500 burn-in						
BAY-R		1.00	0.94	0.35	0.01	0.00
BAY-F		1.00	0.02	0.00	0.00	0.00

Table 3.2: Values of $\hat{R}$ for different amounts of iterations for tests with 10 replications per condition using the BAY-R method and for the final analyses.

Number of replications

A preliminary study using $N = 10$, $T = 10$, $\sigma_{U_{1,n}} = 0.40$, and $\gamma_{0,1} = -0.30$, with the priors and number of iterations as specified, showed that the outcome measures (to be introduced in the next section) started stabilizing after around 1,500 replications for all used methods, being stable for all at 2,000 replications. For example, the standard deviations of the estimated mean $\gamma_{0,1}$ or ϕ_n, depending on estimation method, over replications was lower than 0.01 at 2,000 replications for all used estimators. Therefore, the number of replications per condition is set to $R = 2,000$. Given that we have 40 conditions, this amounts to $40 \times 2,000 = 80,000$ datasets generated.

Summary

Using this simulation design, we can define our study using the classification for intensive longitudinal data designs as discussed by Hamaker et al. (2015). We analyze multi-subject data (where the single-subject case can be seen as a special case). Since we use the classic AR(1) model, we model a univariate, stationary, linear process in discrete time. Our variable has a continuous distribution and is based in the time-domain. Finally, we model the process and are primarily interested in the parameters characterizing the process, rather than the descriptive statistics.

In our simulation study, we consider two measures of computational problems (i.e., non-convergence and non-stationary estimates), and six different outcome measures for the autocorrelation: the bias of $\hat{\gamma}_{0,1}$, the bias of $\hat{\sigma}_{U_{1,n}}$, the empirical standard deviation of $\hat{\gamma}_{0,1}$, the bias of the standard error of $\hat{\gamma}_{0,1}$, the empirical rejection rate (EPr) of $\hat{\gamma}_{0,1}$ and the point and interval estimates of $\hat{\gamma}_{0,1}$. For each outcome measure, we offer a short explanation of the measurement and the obtained results.

3.3.2 Results

We start with discussing the rates of non-convergence (MLE-F) and non-stationarity (MLE-R), followed by the outcome measures for the autocorrelation. We will only discuss the conditions where an effect was found; thus if the random estimator is named but not the fixed estimator, the condition discussed does not influence the result of the fixed estimator and vice versa. The graphs presented in this section show the outcome measures as a function of N, T, $\sigma_{U_{1,n}}$ and $\gamma_{0,1}$. The model parameters will be discussed in the notation used in Equation 3.6, the statistics obtained with the random and fixed estimators in their respective notations as in Equations 3.6 and 3.1.

Computational problems: Non-convergence and non-stationary estimates

The MLE-F is occasionally unable to reach convergence in the estimation of the model, which is connected to the inability to estimate values outside the range of -1 to 1. Of the 40 conditions, 28 converged for all analyses performed. In total, 0.002% of the estimates did not reach convergence. The highest percentage of non-convergence for individual time series is 0.01% for the condition with $N = 10$, $T = 25$, $\sigma_{U_{1,n}} = 0.25$, and $\gamma_{0,1} = 0.6$. Apart from the condition with the highest number of non-convergence, higher numbers of non-convergence are found for conditions with larger values of $|\phi|$ and conditions with the highest value of $\sigma_{U_{1,n}}$.

Out of the 40 conditions, only three had purely stationary estimates. In total 0.33% of the estimates were non-stationary. The highest percentage of non-stationary values for the MLE-R was 1.23%, for the condition with $N = 10$, $T = 10$, $\sigma_{U_{1,n}} = 0.40$, and $\gamma_{0,1} = -0.60$. As expected, higher numbers of non-stationary estimates were found for higher values of $|\gamma_{0,1}|$ and for the highest value of $\sigma_{U_{1,n}}$.

Thus, although we found non-convergence and non-stationarity in some cases, their low occurrence indicate that the problems caused by these issues are minor.

Bias of $\hat{\gamma}_{0,1}$

The bias of the $\hat{\gamma}_{0,1}$ indicates whether a systematic under- or overestimation of $\gamma_{0,1}$ is found. The bias is computed as:

$$\text{bias} = \left(\frac{1}{R} \sum_{r=1}^{R} \hat{\gamma}_{0,1_r} \right) - \gamma_{0,1}, \tag{3.8}$$

where r ($r = 1, 2, ..., R$) refers to the replication number. The random estimators estimate $\hat{\gamma}_{0,1}$ directly. For the fixed estimators, $\hat{\gamma}_{0,1}$ is estimated as $\frac{1}{N}\sum_{n=1}^{N} \hat{\phi}_n$.

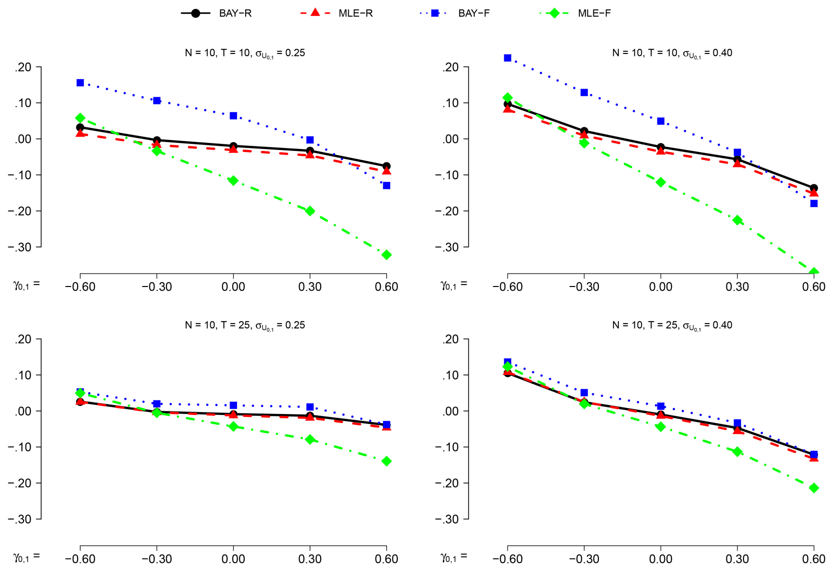


Figure 3.2: The bias of $\hat{\gamma}_{0,1}$ for $N = 10$ for the different estimators, different time series length T and different values of $\sigma_{\gamma_{0,1}}$ as a function of $\gamma_{0,1}$.

The bias decreases marginally for $N = 25$ compared to $N = 10$, with the largest difference being -0.05 for MLE-R, in the conditions with $T = 25$, $\sigma_{U_{1,n}} = 0.25$ and $\gamma_{0,1} = 0.6$. This prompted us to only show the results for $N = 10$, see Figure 3.2. The bias decreases for $T = 25$ compared to $T = 10$ for the fixed methods. For $\sigma_{U_{1,n}} = 0.25$ compared to $\sigma_{U_{1,n}} = 0.40$, the bias decreases for all methods. A trend is present, where the value of the bias of $\hat{\gamma}_{0,1}$ decreases as $\gamma_{0,1}$ increases. The bias is, in general, positive for negative values of $\gamma_{0,1}$, and negative for positive values of $\gamma_{0,1}$.

As can be seen in Figure 3.2, the random estimators, BAY-R and MLE-R, show a smaller bias than the fixed estimators, MLE-F and BAY-F. This difference is larger when $T = 10$ compared to $T = 25$. The difference between MLE-R and BAY-R is very small and inconsistent over conditions. For $\gamma_{0,1}$ above 0.00, the bias of MLE-F is larger than the bias of BAY-F; for $\gamma_{0,1}$ below 0.00, this is the other way around.

Bias of $\hat{\sigma}_{U_{1,n}}$

The bias of $\hat{\sigma}_{U_{1,n}}$ indicates whether $\hat{\sigma}_{U_{1,n}}$ is systematically under- or overestimated, and is calculated as:

$$\text{bias} = \left(\frac{1}{R} \sum_{r=1}^{R} \hat{\sigma}_{U_{1,n}} \right) - \sigma_{U_{1,n}}. \tag{3.9}$$

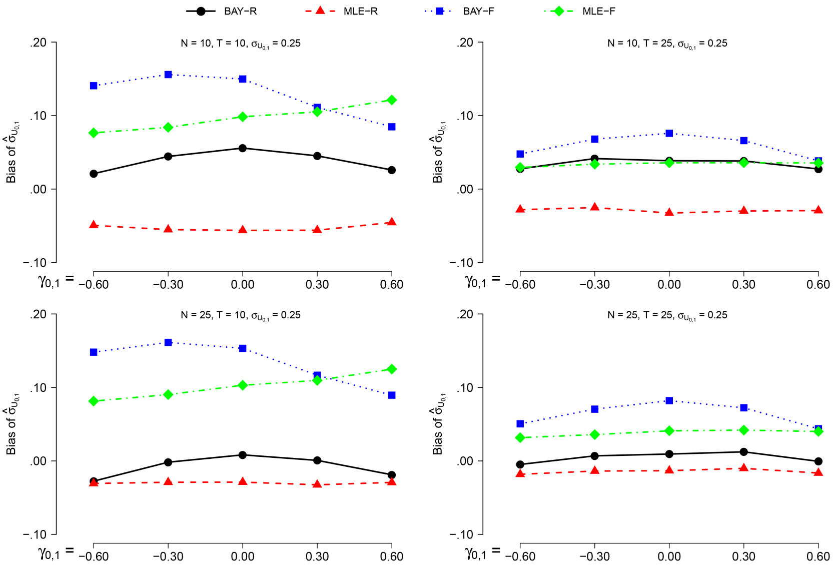


Figure 3.3: The bias of $\hat{\sigma}_{U_{1,n}}$ for $\sigma_{\phi_n} = 0.25$ for the different estimators and different group sizes N for timeseries of different length T as a function of $\gamma_{0,1}$.

The random estimators estimate $\hat{\sigma}_{U_{1,n}}$. For the fixed estimators, $\hat{\sigma}_{U_{1,n}}$ is calculated per replication r as $SD(\hat{\phi}_n)$.

The bias of $\hat{\sigma}_{U_{1,n}}$ is smaller for $\sigma_{U_{1,n}} = 0.40$ than for $\sigma_{U_{1,n}} = 0.25$ for all estimators. As the pattern over the other conditions stays the same, we only show the results for $\sigma_{U_{1,n}} = 0.25$, as depicted in Figure 3.3. For the random estimators, the bias for $N = 25$ is smaller than the bias for $N = 10$. The bias is smaller for $T = 25$ than for $T = 10$, with a more pronounced effect for the fixed estimators. The effect of $\gamma_{0,1}$ is small and inconsistent between conditions and estimators.

BAY-R shows the lowest bias, followed by MLE-R, except for the combination of $\sigma_{U_{1,n}} = 0.40$, $N = 10$ and $T = 25$, where MLE-F shows a smaller bias than both MLE-R and BAY-R. For all conditions, the bias of $\hat{\sigma}_{U_{1,n}}$ is largest for BAY-F.

Empirical $SD(\hat{\gamma}_{0,1})$

The empirical, or observed, standard deviation ($SD(\hat{\gamma}_{0,1})$) indicates the variability of $\hat{\gamma}_{0,1}$. The empirical SD is computed as the standard deviation of $\hat{\gamma}_{0,1}$ over the R replications for the random estimators, and as the standard deviation of $\frac{1}{N}\sum_{n=1}^{N}\hat{\phi}_n$ over replications for the fixed estimators.

The empirical $SD(\hat{\gamma}_{0,1})$ is larger for $\sigma_{U_{1,n}} = 0.40$ than for $\sigma_{U_{1,n}} = 0.25$, on average by a factor of 1.2. The effect of all other parameters is equal for both values of $\sigma_{U_{1,n}}$, prompting us to only display the $SD(\hat{\gamma}_{0,1})$ for $\sigma_{U_{1,n}} = 0.40$, as can be seen in Figure 3.4. The $SD(\hat{\gamma}_{0,1})$ is smaller for $N = 25$ compared to $N = 10$,

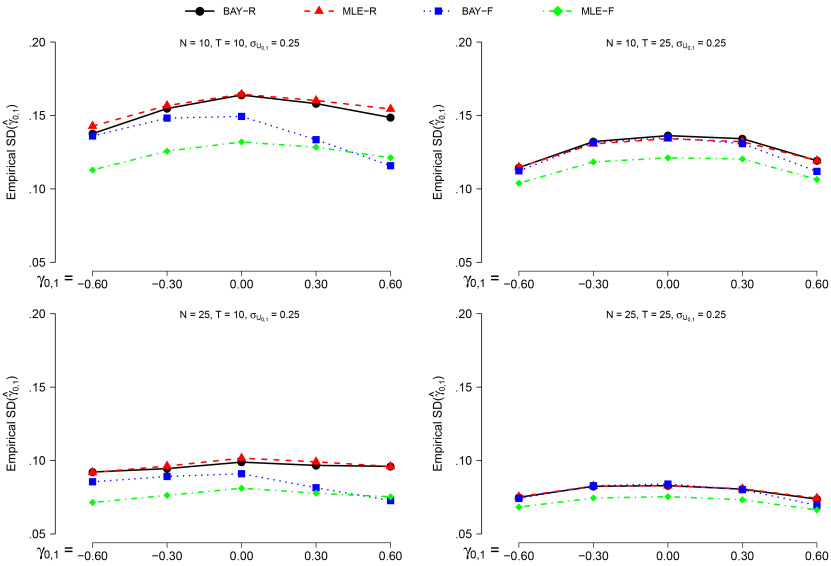


Figure 3.4: The empirical $SD(\hat{\gamma}_{0,1})$ for $\sigma_{\phi_n} = 040$ for the different estimators and different group sizes N for timeseries of different length T as a function of $\gamma_{0,1}$.

and for $T = 25$ compared to $T = 10$. Extreme values of $\gamma_{0,1}$ give a lower $SD(\hat{\gamma}_{0,1})$, but only marginally.

The random estimators show a larger $SD(\hat{\gamma}_{0,1})$ than the fixed estimators. The smallest empirical SD is shown by the MLE-F, followed by the BAY-F. The difference between the MLE-R and BAY-R is small and practically negligible.

Bias of $SE(\hat{\gamma}_{0,1})$

The bias of the standard error indicates how well the methods estimate the standard deviation of $\hat{\gamma}_{0,1}$. The bias of $SE(\hat{\gamma}_{0,1})$ is calculated as:

$$\text{bias of SE } (\hat{\gamma}_{0,1}) = \left(\frac{1}{R} \sum_{r=1}^{R} SE(\hat{\gamma}_{0,1_r}) \right) - SD(\hat{\gamma}_{0,1}), \tag{3.10}$$

where $SE(\hat{\gamma}_{0,1_r})$ is the standard error of $\hat{\gamma}_{0,1}$ in replication r. For the random estimators, the $SE(\hat{\gamma}_{0,1_r})$ is the standard error as calculated by the estimator. For the fixed estimators, the SE is taken as $\frac{1}{N} \sum_{n=1}^{N} SE(\hat{\phi}_n)$.

The bias of $SE(\hat{\gamma}_{0,1})$ is smaller when $\sigma_{U_{1,n}} = 0.40$ than when $\sigma_{U_{1,n}} = 0.25$. However, the effect of all other parameters on the bias of $SE(\hat{\gamma}_{0,1})$ is equal for both values of $\sigma_{U_{1,n}}$, prompting us to display the results for $\sigma_{U_{1,n}} = 0.25$ only, as can be seen in Figure 3.5. For the random estimators, $N = 25$ gives a smaller bias than $N = 10$, for the fixed estimators this is the other way around. The effect of

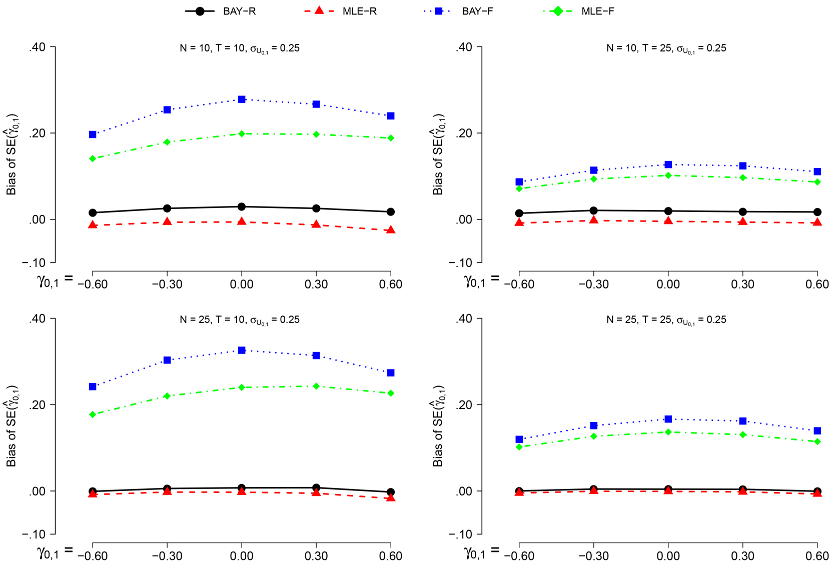


Figure 3.5: The bias of $SE(\hat{\gamma}_{0,1})$ for $\sigma_{\phi_n} = 0.25$ for the different estimators and different group sizes N for timeseries of different length T as a function of $\gamma_{0,1}$.

T is only present for the fixed estimators, which show a smaller bias of $SE(\hat{\gamma}_{0,1})$ for $T = 25$ than for $T = 10$. For the fixed estimators, this effect is stronger than the effect of N. The different values of $\gamma_{0,1}$ only influence the estimations of the fixed estimators, which show a slightly smaller bias for higher values of $|\gamma_{0,1}|$.

The MLE-R shows the smallest bias of $SE(\hat{\gamma}_{0,1})$ for all conditions, and is the only estimator which shows a constant negative bias. For higher values of N, the difference between MLE-R and BAY-R disappears. For all conditions, the bias of $SE(\hat{\gamma}_{0,1})$ is larger for the fixed estimators than for the random estimators.

Empirical rejection rate and power

For each estimator and condition, we compute the empirical probability (EPr) for rejecting $H_0 : \gamma_{0,1} = 0$ in favor of $H_\alpha : \gamma_{0,1} \neq 0.00$, with $\alpha = 0.05$. Using frequentist terminology, the EPr equals the actual α in the condition with $\gamma_{0,1} = 0.00$; and the power in all other conditions.

For frequentist methods, testing $H_0 : \gamma_{0,1} = 0$ versus a two-sided alternative at significance level α, is equivalent to checking whether the $(1 - \alpha)$ confidence interval (CI) includes zero or not. The CI per replication per condition and per estimator is calculated as follows:

$$\hat{\gamma}_{0,1} \pm t^*_{(1-\alpha);df=N-2} SE(\hat{\gamma}_{0,1}), \tag{3.11}$$

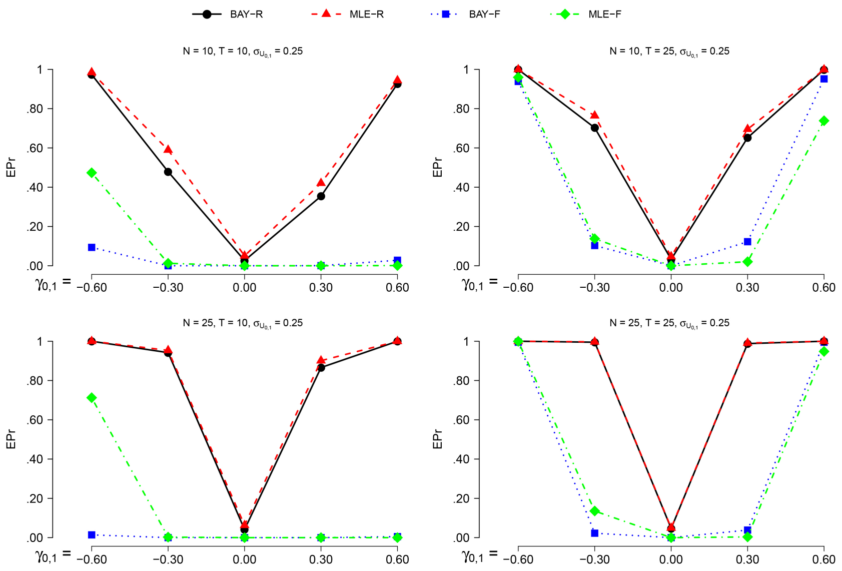


Figure 3.6: The EPR for $\sigma_{U_{1,n}} = 0.25$ for the different estimators, different group sizes N, and for different timeseries length T as a function of $\gamma_{0,1}$.

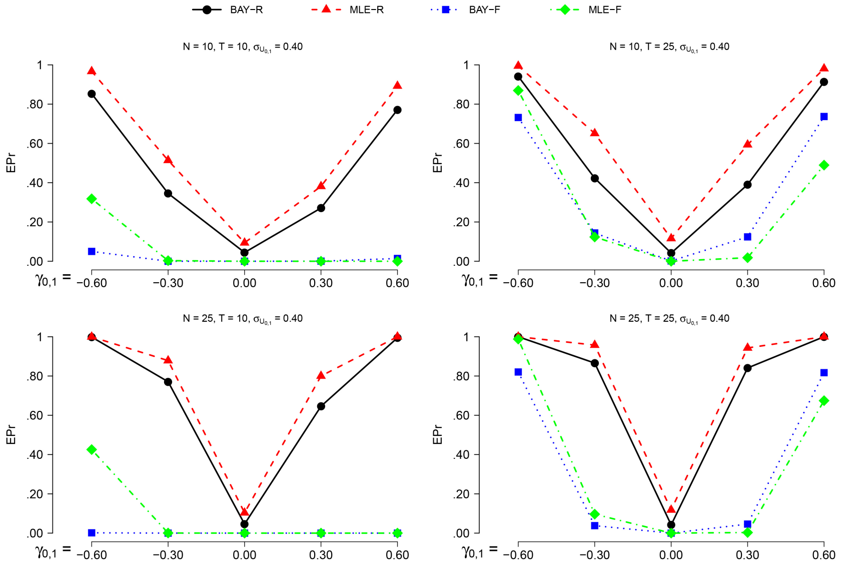


Figure 3.7: The EPR for $\sigma_{U_{1,n}} = 0.40$ for the different estimators, different group sizes N, and for different timeseries length T as a function of $\gamma_{0,1}$.

where $SE(\hat{\gamma}_{0,1})$ is obtained as explained in Section 3.3.2. The proportion of replications per condition for which the corresponding confidence interval does not contain zero is the EPr.

For the Bayesian estimators, the EPr is the proportion of replications per condition for which the credible interval (CrI) as obtained through MCMC does not hold zero. For the BAY-R, we consider the CrI of $\hat{\phi}$, for BAY-F we use the average scores of the CrI's of $\hat{\phi}_n$ within each replication.

The power is higher for $N = 25$ than for $N = 10$ and for $T = 25$ compared to $T = 10$, as can be seen in Figures 3.6 and 3.7. The actual α shows no such effect. The EPr shows lower values for $\sigma_{U_{1,n}} = 0.40$ compared to $\sigma_{U_{1,n}} = 0.25$, except for the actual α of MLE-R. When $|\gamma_{0,1}|$ is higher, the EPr becomes higher. For the fixed estimators, this effect is strongly dependent on T: for $T = 10$, the EPr only increases for $\gamma_{0,1} < -0.30$.

The highest power is found using BAY-R when $\sigma_{U_{1,n}} = 0.25$, and using MLE-R when $\sigma_{U_{1,n}} = 0.40$. For the fixed estimators, the BAY-F shows a higher power than the MLE-F. The BAY-R has an actual α consistently at or around 0.05, while the MLE-R has an actual α that is too high for $\sigma_{U_{1,n}} = 0.40$, namely at 0.10. The fixed estimators have an actual α at or even below 0.01, rather than the desired 0.05.

Point and interval estimates of $\gamma_{0,1}$

To illustrate the joint effects of bias and variability we consider BAY-R and MLE-R, using the point and interval estimates of $\gamma_{0,1}$. As point estimate we use the mean of $\hat{\gamma}_{0,1}$ per condition. For the interval estimation we present the 2.5 and 97.5 percentiles of the $\hat{\gamma}_{0,1}$ across all R replications per condition as the lower and upper bounds.

The point estimates and interval estimates can be seen in Figure 3.8 for $\sigma_{U_{1,n}} = 0.40$. The interval is larger for $N = 10$ and for $T = 10$ than for $N = 25$ and $T = 25$. The effect of N is slightly larger. $\sigma_{U_{1,n}} = 0.25$ effectuates a smaller estimation interval than $\sigma_{U_{1,n}} = 0.40$, the latter being 1.2 to 1.3 times the former. The influence of $\gamma_{0,1}$ on the estimation interval is negligible, as are the differences between BAY-R and MLE-R.

3.3.3 Combined conclusions of the different measures

We found that the use of random estimators as opposed to fixed estimators improves all measurements considerably, except for the empirical SD, which is larger for the random estimators. The BAY-R shows a slight advantage over the MLE-R with respect to the bias of $\hat{\sigma}_{U_{1,n}}$ and the bias of $SE\hat{\gamma}_{0,1}$. As expected, higher values of N and T improve the estimation. Further, as expected, a lower value of

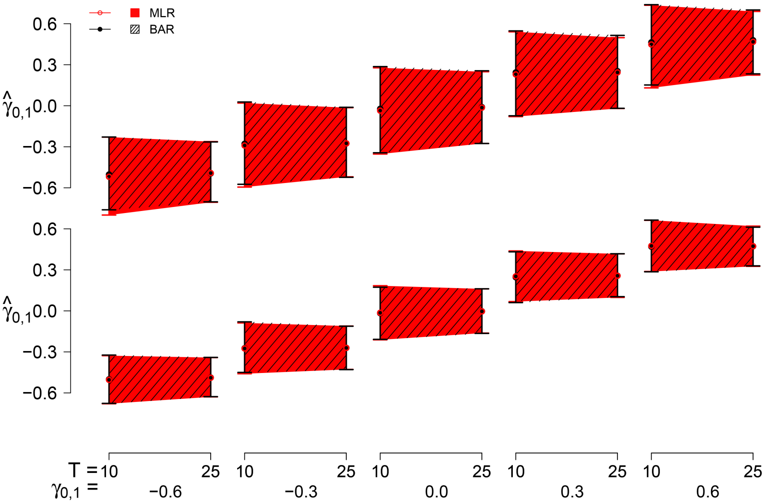


Figure 3.8: The point and interval estimates for $\sigma_{\phi_n} = 0.40$ for the different estimators and different group sizes N ($N = 10$ top pane, $N = 25$ bottom pane) for time series length $T = 10$ and $T = 25$ as a function of $\gamma_{0,1}$.

$\sigma_{U_{1,n}}$ lowers the bias of $\hat{\gamma}_{0,1}$, lowers the $SD(\hat{\gamma}_{0,1})$ and increases the power, but also increases the bias of $\hat{\sigma}_{U_{1,n}}$ and the bias of $SE(\hat{\gamma}_{0,1})$.

3.4 Discussion and conclusions

In this chapter we studied the performance of four models for multilevel time-series data. We compared two estimation methods, namely maximum likelihood estimation and Bayesian MCMC, as previous work indicates that these methods perform best for single case designs (Krone et al., 2016a). We combined this with two model variants, a random model and a fixed model, to obtain four estimators: MLE-F, MLE-R, BAY-F and BAY-R. We compared their estimates in different conditions, where we varied the time series lengths, number of subjects and the mean and standard deviation of the autocorrelation distribution. As outcome measures, we considered the bias, the bias of the standard deviation, the empirical standard deviation, the bias of the standard error, the empirical rejection rate, and the point and interval estimates of the autocorrelation.

We found substantial differences between the fixed and the random estimators. When compared to the fixed estimators, the random estimators show better results for the bias, the bias of the standard deviation, the bias of the standard error and the power. Furthermore, the actual α as obtained with the fixed estimators, appears to be between 0.00 and 0.01, in stead of 0.05. The fixed estimators show a better empirical standard deviation than the random estimators. In general, the random estimators are clearly preferred over the fixed estimators.

Smaller differences were found between the estimation methods. In general, the Bayesian MCMC shows a smaller bias than the MLE. The bias of the standard deviation is smaller for BAY-R than for MLE-R, but smaller for MLE-F than for BAY-F. The empirical standard deviation is smaller for the MLE-F than for the BAY-F, but the difference between BAY-R and MLE-R is negligible. The bias of the standard error is smaller for the MLE. The power is higher for the MLE estimators, but the actual α is better for the Bayesian MCMC. In general, the bias of the estimated autocorrelation is smaller for the Bayesian MCMC, but the variability is smaller for the MLE estimators.

The effect of the different conditions depends on the model variant. A higher sample size N improves all outcome measurements for the random estimators. For the fixed estimators, a higher N marginally improved the bias, the empirical standard deviation and the power of the autocorrelation. However, although the increase in N decreased the empirical standard deviation, it did not influence the estimation of the standard error, thus increasing the bias of the standard error for $N = 25$.

The time series length T influences the estimations for both model variants. A higher value of T showed small but positive effects on the outcome measures for the random estimators. However, the improvement was smaller than for an equal increase in N. For the fixed estimators, the results were more profound, showing stronger improvements in all outcome measures than obtained for an equal increase in N.

The standard deviation of ϕ_n influenced the results for all estimators and conditions. A higher $\sigma_{U_{1,n}}$ gave less favorable results for the autocorrelation with regard to bias, empirical standard deviation and power, but more favorable results for the bias of the standard deviation and the bias of the standard error. The effect of the mean of ϕ_n, $\gamma_{0,1}$, differs per estimator and per condition, not showing a clear pattern between estimators and conditions. Earlier studies showed a negative relation between the bias and $\gamma_{0,1}$: a negative $\gamma_{0,1}$ gave a positive bias, and the other way around (e.g., Huitema & McKean, 1991; DeCarlo & Tryon, 1993; Solanas et al., 2010). This result was replicated.

An important question in time series analysis is how many individuals and time points are needed to obtain acceptable estimates for a given model. In choosing between a random or a fixed approach to modeling, the random modeling is clearly favored when the assumptions associated with the model do hold. In this case, more individuals can be used to make up for a smaller number of time points, and the other way around. When $\sigma_{U_{1,n}}$ is up to 0.25, the random model may produce results with an acceptable size of bias when T or N is at least 25, and the other one of the two is at least 10. When $\sigma_{U_{1,n}}$ is up to 0.40, both are required to be higher than 25. The number of individuals only has a small effect on the results for the fixed model. Here, the number of time points is the strongest criteria. In this study, we still found a sizable bias for 25 time points, which is stronger for $\sigma_{U_{1,n}} = 0.40$. This is confirmed in single subject studies, where a T of 50 is advised (Box & Jenkins, 1976; Krone et al., 2016a).

The aim of this chapter was to compare the four estimators MLE-F, MLE-R, BAY-F and BAY-R using a multilevel AR(1) model. For the single subject AR(1) model, several issues and important factors are discussed in the literature. These may be just as relevant for a multisubject model, such as our multilevel model. The AR(1) model, though very often used, is not sophisticated enough for various empirical applications. This is because the error term ($e_{t,n}$ in Equation (3.6)) is also affected by the auto-correlation. Schuurman, Houtveen, and Hamaker (2015) demonstrates that including so-called white noise (i.e., error not carried over to the next time point) in the model, leads to improved empirical model fit. Lacking this term leads to underestimation of the absolute autocorrelation. Studying how various estimators perform under such an extension to the (multilevel) AR(1) model is an interesting step in future research.

The literature on the single subject AR(1) model discusses several other factors that influence the estimation of the autocorrelation. In our models we kept the error variance equal for all datasets, but this does influence the estimation of the AR(1) model (Schuurman et al., 2015), as does the error distribution (Solanas et al., 2010). This may also influence the performance of the different estimators as used in this chapter. Another issue is misspecification, where the model used may not be equal to the one underlying the data. Earlier studies showed that this influences the estimation of the autocorrelation (Tanaka & Maekawa, 1984; Kunitomo & Yamamoto, 1985; Krone et al., 2016a). For the multilevel model, the inclusion of a random error covariance may improve estimation, while person-centering may have a negative effect on the estimation of the parameters (Jongerling et al., 2015). The effect of these factors on the different estimators in a multilevel model is also an interesting topic for further studies.

We chose a well-known multilevel framework for our estimators, which is often used in longitudinal analyses. An alternative framework to model an AR-model is a State Space Model (SSM) (Durbin & Koopman, 2012). The versatility of the SSM means that it can be used for a vast range of models and any distribution for which a link-function with the normal distribution exists. Furthermore, the implementation of measurement error parameters is straightforward in a SSM. SSM can be modeled to allow for a multilevel AR(1) structure for different kinds of distributions; implementations have been made for normally distributed data (Lodewyckx, Tuerlinckx, Kuppens, Allen, & Sheeber, 2011) and data following a Poisson distribution (Terui, Ban, & Maki, 2010). However, the theoretical framework to estimate a SSM with any distribution in the exponential family is available

(Durbin & Koopman, 2012; Petris, Petrone, & Campagnoli, 2009). A Bayesian interpretation of the state space model is found in the Bayesian dynamic model (West & Harrison, 1997).

We compared several estimators, but many other possibilities remain. Future studies may look into the effect of data properties, such as the error variance or misspecification, and different ways of modeling the data, using for example a SSM framework. Finally, we did not assess how the estimators handle missing data, and what the effect of missing data is on the outcome measures. As missing data occurs often in the social sciences, this is an interesting and important topic for further studies.

Chapter 4

Bayesian Dynamic Modeling to assess differential treatment effects on panic attack frequencies

Abstract

To represent the complex structure of intensive longitudinal data of multiple individuals, we propose a hierarchical Bayesian Dynamic Model (BDM). This BDM is a generalized linear hierarchical model where the individual parameters do not necessarily follow a normal distribution. The model parameters can be estimated on the basis of relatively small sample sizes and in the presence of missing time points. We present the BDM and discuss the model identification, convergence and selection. The use of the BDM model is illustrated using data from a randomized clinical trial to study the differential effects of three treatments for panic disorder. The data involves the number of panic attacks experienced weekly (73 individuals, 10 to 52 time points) during treatment. Presuming that the counts are Poisson distributed, the BDM considered involves a linear trend model with an exponential link function. The final model included a moving average parameter, and an external variable (duration of symptoms pre-treatment). Our results show that cognitive behavioral therapy is less effective on the reduction of panic attacks than serotonin selective re-uptake inhibitors or a combination of both. Post-hoc analyses revealed that males show a slightly higher number of panic attacks at the onset of treatment than females.

This accepted for publication as: Krone, T., Albers, C. J., & Timmerman, M. E. (2016c). Bayesian dynamic modeling to assess differential treatment effects on panic attack frequencies. *Statistical Modelling (in press)*. doi:10.1177/1471082X16650777

4.1 Intensive longitudinal data to study psychological processes

In recent years the interest in studying the course of psychological processes has increased. One may think of, for example, the effects of psychological interventions across time and the role of external variables on these effects (Elkins & Moore, 2011; Kellett, 2007; Wild et al., 2006). To study these processes, intensive longitudinal data is obtained: data of one or more individuals are gathered repeatedly over time, in such a frequency and over such a time span that it characterizes the trends and dynamics of interest (Hamaker et al., 2015). The measurements can pertain to questionnaires that are administered on a weekly or daily basis, or even multiple times a day. The latter is typically referred to as ecological momentary assessment (EMA) (Larson & Csikszentmihalyi, 1983; Shiffman et al., 2008).

Intensive longitudinal data typically have specific characteristics, which may yield selecting a proper statistical model a challenging task. First, the amount of data available is often rather limited, in terms of the number of observed time points per individual and the number of observed individuals.

Second, missing data may easily occur. Commonly, the amount of incomplete data substantially increases with larger numbers of scheduled time points. The data can be incomplete because of completely missing individuals, or because of incidental missing values, where observations on one or a few time points within a series are lacking, or because of drop-out, where a series lacks observations after a specific time point. The latter may coincide with the drop-out of an intervention (e.g., therapy), but this is not necessarily the case. It is typically wise to use all available data in modeling the intensive longitudinal data, to reduce bias and uncertainty in the model estimates.

Third, intensive longitudinal data is usually collected among multiple individuals. Then the interest is to capture the intra-individual processes as well the interindividual differences in these processes, and possibly to relate these to the external variables. This requires a model that covers the hierarchical structure in the data, where time points are nested within individuals. A popular approach is the multilevel regression model for repeated measures (Bryk & Raudenbush, 1992; Goldstein, 2011; Snijders & Bosker, 1999). However, this model may be overly restrictive in empirical practice, because of its normality assumption of the individual parameters.

Fourth and final, measurements of psychological processes are typically made on discrete scales. Examples include binary scales, such as indicating the absence or presence of a certain behavior, ordinal polytomous scales, such as the well-known Likert-scale, and counts, such as the number of times a certain behavior occurred. The model fit can improve considerably by using a proper distribution for the scales at hand, rather than the often applied normal distribution. The latter can be an approximation to the discrete scales at best.

These characteristics result in a couple of requirements for a proper intensive longitudinal data analysis method. The method must be able to build a model upon a relatively small amount of data, even though more data will improve the estimation and allow for more complex models to be estimated reliably. The method must be able to deal with missing data. The method should allow for achieving insight into the intra-individual processes and their inter-individual differences. Herewith it is important that the distribution of the individual parameters is not necessarily restricted to normality. Finally, the method must allow for the typically occurring discrete scales (i.e., binary, ordinal polytomous and counts). These requirements lead us to the Bayesian Dynamic Model (BDM) (West & Harrison, 1997), which can handle the combination of requirements mentioned. The remainder of the chapter is organized as follows. In the next section, we will introduce the general BDM-framework. To illustrate the usefulness of the BDM for modeling psychological processes, we present an empirical application to intensive longitudinal count-data from multiple individuals, who participated in a randomized clinical trial. We introduce three variants of the BDM and the three fitted BDMs will be interpreted, and the results will be compared to the previous modeling endeavor which used a frequentist multilevel model. We will conclude with a discussion pertaining to the general use of the BDM for intensive longitudinal data.

4.2 Bayesian Dynamic Model

The BDM, including its generalization the Bayesian Dynamic Generalized Linear Model (West & Harrison, 1997), is a Bayesian interpretation of the state space model. The BDM includes a latent score that is connected to the observed score using the so-called latent state vector. The BDM comprises three equations, namely the link function, the observation equation and the system equation. We will successively present these three equations, in view of jointly modeling intensive longitudinal data of multiple individuals.

Link function We model the distribution p of observed score $y_{t,n}$ of individual n $(n = 1, ..., N)$ at time t $(t = 1, 2, ..., T_n)$ using a latent score $y^*_{t,n}$. The link function we use is equal to the one used in generalized linear models for observed data (Nelder & Wedderburn, 1972; McCullagh & Nelder, 1989) and allows for modeling any distribution from the exponential family:

$$\text{Link function :} \qquad p(y_{t,n}) = g^{-1}(y^*_{t,n}).$$

Observation and system equations The observation equation connects the latent score $y^*_{t,n}$ of individual n at time t to the latent state vector $\boldsymbol{\theta}_{t,n}$:

$$\text{Observation equation :} \qquad y^*_{t,n} = \boldsymbol{f}_{t,n}\boldsymbol{\theta}_{t,n} + \varepsilon_{t,n}, \tag{4.1}$$

where $\boldsymbol{f}_{t,n}$ $(1 \times r)$ is the design vector at time t for individual n, $\boldsymbol{\theta}_{t,n}$ $(r \times 1)$ is the latent state vector at time t for individual n and $\varepsilon_{t,n}$ is the white noise. The system equation models the evolution of the latent state vector over time:

$$\text{System equation :} \qquad \boldsymbol{\theta}_{t,n} = \boldsymbol{G}_{t,n}\boldsymbol{\theta}_{t-1,n} + \boldsymbol{\eta}_{t,n},$$

where $\boldsymbol{G}_{t,n}$ $(r \times r)$ is the innovation matrix and $\boldsymbol{\eta}_{t,n}$ $(r \times 1)$ is the innovation noise vector, at time t for individual n.

Each instance of the vectors $\boldsymbol{f}_{t,n}$ and matrices $\boldsymbol{G}_{t,n}$ must be defined by the researcher. When there is no reason to assume that $\boldsymbol{f}_{t,n}$ and $\boldsymbol{G}_{t,n}$ differ over time and/or over individuals, it is advised to take $\boldsymbol{f}_{t,n}$ and $\boldsymbol{G}_{t,n}$ invariant over time and/or over individuals, to simplify the model and its interpretation. In what follows, we presume that both the design vector and the innovation matrix are invariant over time and individuals. This results in the use of a single vector $\boldsymbol{f}$ and a single matrix $\boldsymbol{G}$ per model.

Noise parameters Typically, it is assumed that the noise, as expressed through $\varepsilon_{t,n}$ and $\boldsymbol{\eta}_{t,n}$, follows a normal distribution with mean zero and standard deviation $\sigma_{\varepsilon_{t,n}}$ and covariance matrix $\boldsymbol{\Sigma}_{\eta_{t,n}}$, respectively. Varying (co)variances of the noise over time and/or individuals may be necessary to achieve a proper fitting model (Jongerling et al., 2015). However, it also complicates the model and may give rise to estimation difficulties due to identification issues and/or lack of data. To avoid these difficulties, the distribution is often taken to be invariant over time and individuals, giving σ_ε and $\boldsymbol{\Sigma}_\eta$ instead of $\sigma_{\varepsilon_{t,n}}$ and $\boldsymbol{\Sigma}_{\eta_{t,n}}$, respectively. As an alternative, advanced techniques such as variance discounting (West & Harrison, 1997, p. 194-195), may be employed. Though typically assumed, the noise distribution is not restricted to normality. Note that when the link function already includes the variance of $y^*_{t,n}$ (as is the case for Poisson data where the variance equals the mean, which is given by the link function), $\varepsilon_{t,n}$ can be omitted from Equation 4.1.

Priors As any Bayesian model, the BDM requires priors chosen by the researcher. The choice of priors generally depends on the expected posterior distribution. If there is little information on what is to be expected, a weak informative prior may be used. An example is a symmetrized reference prior for autocorrelations (Berger & Yang, 1994), which can essentially be used for all instances when it is only known that the parameter to be estimated is an autocorrelation.

When more information is present, an informative prior can be used. For example, when a parameter appeared to be between -0.2 and 0.2 in earlier, well-conducted and trusted studies, a normal prior with mean zero and standard deviation 0.1 may be used. Finally, a non-informative prior may be used, generally an uniform distribution encompassing all possible values for the indicated parameter. However, for parameters with theoretically an infinite range of possibilities, these slow down calculations. Further, a weak informative prior is often available or can be derived from what is known about the expected range of the posterior estimation of the parameters.

Missing data As stated, the BDM can handle both incidental missing data and drop-outs. Incidental missing data can be handled by not linking the observed score to the latent state when the observed score is missing. In this case, the observation equation and the system equation are estimated, but the link between the observed score and the latent score is not made. For drop-outs, the analysis will stop at T_n, the final observed time point for individual n.

External variables The BDM allows for inclusion of external variables in two ways. First, the external variable can be included as an active covariate. This can be done as a direct effect, for example by considering an external variable as an element of $\boldsymbol{F}_{t,n}$ in Equation 4.1, and as a moderator effect, for example letting an element of $\boldsymbol{\theta}_{t,n}$ be dependent on the level of a covariate. Second, inactive covariates can be implemented post-hoc, by examining the relation between any model parameter and an external variable after the model estimation. This can be done, for instance, by using partial correlations or linear regression, thereby accounting for confounding variables. In our empirical example, we will demonstrate both approaches.

Model estimation The BDM is estimated using Bayesian Markov Chain Monte Carlo (MCMC) estimation. For the MCMC estimation, we use Hamiltonian Monte Carlo (HMC), a generalization of the Metropolis-Hastings algorithm that allows for an efficient estimation of the parameters (Gelman, Carlin, et al., 2013). This is incorporated in the software RStan (R Core Team, 2015; Stan Development Team, 2015), which we used in our modeling.

Model convergence The BDM can be a fairly complex model, especially when parameters are allowed to differ over time and individuals. This may lead to estimation problems due to identification issues, where more than one solution fits the data equally well, or due to too little data in comparison to the complexity of the model. Both of these will result in non-convergence, which implies that the estimation procedure has failed to find the single, optimal solution.

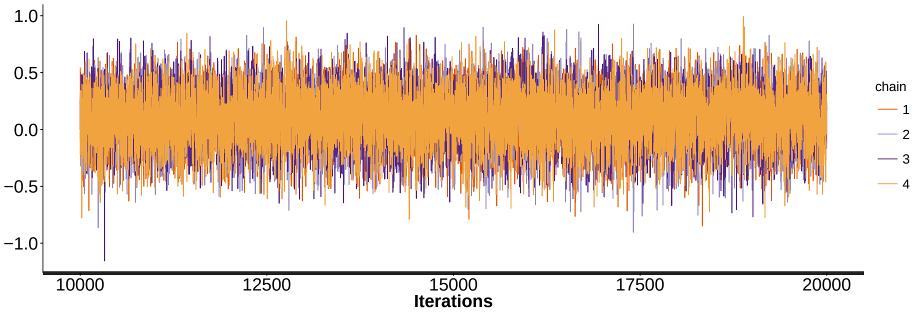


Figure 4.1: Traceplot of a converged parameter with a $\hat{R}$ of 1.00 and posterior mean 0.10

The convergence can be checked through visual inspection of the trace plots or through assessment of the potential scale reduction factor, $\hat{R}$. A trace plot shows the Bayesian MCMC estimates for each parameter at each iteration. If a parameter reaches convergence, the estimates over iterations are highly similar across chains. As a result, the trace plot looks like a fat caterpillar with all chains completely overlapping, except at the fringe of the caterpillar (as shown in Figure 4.1). The $\hat{R}$ expresses the ratio of how much the estimation may change when the number of iterations is doubled; the (ideal) value of 1 indicates that no change is expected (Gelman & Rubin, 1992; Stan Development Team, 2016).

Model selection To select a BDM from a series of competing variants, one may use two strategies. First, the model variants can be compared using the estimated parameters to see which parameters show the most preferable properties. For example, the noise variance is preferred to be small, indicating a proper model fit. Further, model parameters which are close to zero may be superfluous, thereby unnecessary complicating the model.

Second, competing models can be compared considering their fit (i.e., log-likelihood) and number of parameters. To this end, several information criteria can be used, such as the Deviance Information Criterion (Spiegelhalter, Best, Carlin, & Van Der Linde, 2002), the Bayesian Information Criterion (Schwarz, 1978) and the Watanabe-Aikake Information Criterion (WAIC) (Watanabe, 2010). Compared to other information criteria, the WAIC most closely follows the Bayesian methods, as it takes into account the whole posterior distribution, as opposed to just the point estimates (Gelman, Hwang, & Vehtari, 2013). As is usual with information criteria, a lower WAIC indicates a better predictive model accuracy.

4.3 Empirical application: differential treatment effects on panic attack frequencies

In this chapter, we re-analyze data described in Van Apeldoorn, Van Hout, Timmerman, Mersch, and Den Boer (2013). In a randomized clinical trial involving panic disorder patients, the differential rate of improvement across three types of therapy was examined. The three treatments involved were Cognitive Behavioural Therapy (CBT), Serotonin Selective Re-uptake Inhibitors (SSRI) and a combination of both (BOTH). As a measure that reflects symptom severity and that is feasible to be measured intensively, the frequency of panic attacks is used. Each patient recorded the number of panic attacks experienced during the previous week. They did so on a weekly basis for the period of one year in which treatment was delivered, including medication taper. Here, we only consider the patients who completed the therapy according to protocol and who provided scores on at least 10 consecutive time points. This results in 73 out of 178 patients, of which $n = 28$ with CBT, $n = 22$ with SSRI and $n = 23$ with BOTH.

4.3.1 Model specification

The core research questions of the study were how the frequencies of panic attacks develop during and after intervention and to what extent this differs across the three treatments, and possibly relates to individual characteristics. The initial state is captured via an individual intercept, and the course across time via an individual slope. Following Van Apeldoorn et al. (2013), we included treatment (CBT, SSRI, BOTH) and level of agoraphobia (no/mild versus moderate/severe) as individual predictors. Earlier research has shown that a panic disorder, when remaining untreated, may become more severe and change in nature (Altamura, Santini, Salvadori, & Mundo, 2005; Federici & Tommasini, 1992). As a result, the intensity of symptoms may increase over time when no treatment is received. To test this, we include the duration of symptoms pre-treatment as a predictor for the panic attack frequency at the start of the treatment.

In empirical data, the noise terms are often not independent for subsequent time points (Goldstein et al., 1994). To assess the presence of autocorrelated noise in our model, we include an moving average mechanism in the noise of the system equation (Box & Jenkins, 1976).

To assess the importance of each of these elements, we will compare three models. Model 1 will include only the predictors for the slope, being treatment and presence of agoraphobia. In Model 2, we will add the moving average mechanism to the system equation. In Model 3, we will add the duration of symptoms pre-treatment as a predictor for the intercept.

4.3.2 Model design

We elicit three different models for the data set.

Link function and observation equation To map our observed count data $y_{t,n}$ to the continuous latent score $y^*_{t,n}$, we consider a Poisson distribution and use an exponential link function:

$$y_{t,n} \sim \text{Poisson}(g^{-1}(y^*_{t,n})), \qquad g(y^*_{t,n}) = \exp(y^*_{t,n}).$$

Earlier studies found an exponential decay in symptoms over time (e.g., Bandelow et al., 2004; Ross, Klein, & Uhlenhuth, 2010; Toni, Perugi, Frare, Mata, & Akiskal, 2004), which is also seen in the observed scores in our data set. Since our link function uses an exponential transformation, we can use a linear function of t on $y^*_{t,n}$ to model the exponential decay in $y_{t,n}$.

As variation is implicitly included in the link function, we can have an observation equation without white noise:

$$y^*_{t,n} = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} \mu_n \\ \delta_{t,n} \\ \beta_n \end{bmatrix}.$$

where μ_n is the intercept of individual n, roughly corresponds to $y^*_{1,n}$, $\delta_{t,n}$ is difference between the intercept and $y^*_{t,n}$, and β_n is the slope. The link function and observation equation are taken equal for all three models.

System equation For Model 1, the system equation depicts a linear growth model, with slope β_n and intercept μ_n:

$$\begin{bmatrix} \mu_n \\ \delta_{t,n} \\ \beta_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} \mu_n \\ \delta_{t-1,n} \\ \beta_n \end{bmatrix} + \begin{bmatrix} 0 \\ \eta_{t,n} \\ 0 \end{bmatrix}, \qquad \eta_{t,n} \sim N(0, \sigma_\eta), \tag{4.2}$$

where $\eta_{t,n}$ is the innovation error at time t for individual n with standard deviation σ_η.

For Models 2 and 3, we add a moving average mechanism to the system equation, Equation (4.2):

$$\begin{bmatrix} \mu_n \\ \delta_{t,n} \\ \beta_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} \mu_n \\ \delta_{t-1,n} \\ \beta_n \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \eta_{t-1,n} & \eta_{t,n} \\ 0 & 0 \end{bmatrix} \times \begin{bmatrix} \psi_{1,n} \\ 1 \end{bmatrix}, \qquad \eta_{t,n} \sim N(0, \sigma_\eta),$$

where $\psi_{1,n}$ is the lag 1 moving average parameter of individual n.

Predictor variables Including the combination of the treatment group and presence of agoraphobia as predictor for the individual slope yields the following expression for β_n in all three models:

$$\beta_n = \beta_0 + \begin{bmatrix}\beta_{1,1} & \dots & \beta_{1,5}\end{bmatrix} \times \begin{bmatrix} d_{s,m} \\ d_{b,m} \\ d_{c,a} \\ d_{s,a} \\ d_{b,a} \end{bmatrix} + \xi_{\beta,n}, \qquad \xi_{\beta,n} \sim N(0, \sigma_\beta), \tag{4.3}$$

where β_n is the slope for individual n in group $g_{tr,ag}$ with *tr* (*tr*: CBT=c, SSRI=s, BOTH=b) being treatment and *ag* (*ag*: m = none or mild agoraphobia, a = moderate or severe agoraphobia) being presence of agoraphobia. Furthermore, β_0 is the estimated slope-coefficient for an individual in $g_{c,m}$, $\beta_{1,\cdot}$ is the added effect of the group depicted in the accompanying dummy variable $d_{tr=\cdot,ag=\cdot}$, where $d_{tr=\cdot,ag=\cdot}$ is the 0/1-coded dummy variable indicating whether the patient is in group $g_{tr,ag}$, $\xi_{\beta,n}$ is the individual deviation in β_n and σ_β is the standard deviation of $\xi_{\beta,n}$.

In Models 1 and 2, the intercept μ_n is presumed to follow a normal distribution:

$$\mu_n = \mu_0 + \xi_{\mu,n}, \qquad \xi_{\mu,n} \sim N(0, \sigma_\mu).$$

where μ_0 is the mean estimated intercept, $\xi_{\mu,n}$ is the individual deviation in the intercept and σ_μ is the standard deviation of $\xi_{\mu,n}$.

In Model 3, we include the duration of symptoms pre-treatment x_n as predictor variable for μ_n:

$$\mu_n = \mu_0 + \mu_1 \times x_n + \xi_{\mu,n}, \qquad \xi_{\mu,n} \sim N(0, \sigma_\mu), \tag{4.4}$$

where μ_0 is the estimated intercept when x_n is zero and μ_1 is the effect of x_n on μ_n.

Priors The priors of the model are aimed to be weak-informative. The observed number of panic attacks in this data set ranges from 0 to 12 attacks per week, giving an expected range of $y^*_{t,n}$ between $-\infty$ and 2.5. For the standard deviations (σ_η, σ_β and σ_μ) we set the prior at N(0.5,5) with lower bound zero, creating a prior similar to a half-Cauchy prior (Stan Development Team, 2016). Taking into account the expected range of the latent score, we believe that a wider prior would only delay calculations without improving the model estimates. For β_0, β_1, μ_0 and μ_1 we set the prior at N(0,5), since these parameters are expected to have relatively small values, not exceeding an absolute value of 2. Finally, for ψ_n we used Berger's symmetrized reference prior (Berger & Yang, 1994), as this prior has shown to be a better prior for autocorrelation parameters than an uniform $[-1, 1]$ prior (Krone et al., 2016a). This prior does not require hyperparameters.

4.4 Results

For each model, we discuss the convergence and the posterior estimates. Further, we interpret the estimated model parameters and compare the models, considering the noise standard deviations and the WAIC. Finally, we show the results for a post-hoc comparison of males and females with regard to the intercept.

4.4.1 Convergence

Each model is estimated using four MCMC chains of 20,000 iterations, of which half the iterations were used for burn-in. Before the results can be interpreted, we must check whether the chains converge. In all three models, all model parameters for all subjects ($\beta_{\cdot}, \mu_{\cdot}, \psi_n$ and $\sigma_{\cdot}$) showed good to very good $\hat{R}$-values (i.e., all below 1.01, except for σ_ε with $\hat{R} = 1.011$ for Models 1 and 2, and 1.012 for Model 3). The traceplots were all proper, as was already expected from the $\hat{R}$-values, that is, fat caterpillars similar to those in Figure 4.1.

4.4.2 Parameters

The slope β_n

The boxplots for the slopes β_n $(n = 1, ..., N = 73)$ as estimated for each of the three models, are shown in pane (A) of Figure 4.2. For all three models, the mean β_n in the sample is -0.14, with range -0.23 to -0.05 for Model 1, and range -0.24 to -0.02 for Models 2 and 3. In Table 4.1 the posterior means (and standard deviations) of the slope using Equation (4.3) are presented. In Table 4.2 the estimated mean slopes for each treatment/agoraphobica combination are presented. The differential effects are highly similar over the models, showing only small differences in estimated size of effect. The steepest mean slopes are estimated for $g_{s,m}$ and $g_{b,a}$, followed by $g_{s,a}$ and $g_{b,m}$. The shallowest mean slopes are estimated for $g_{c,m}$ and $g_{c,a}$, for which the mean slope fall outside, or are on the edge of, the 95% credible interval (CrI) of the mean estimated slopes for the other groups.

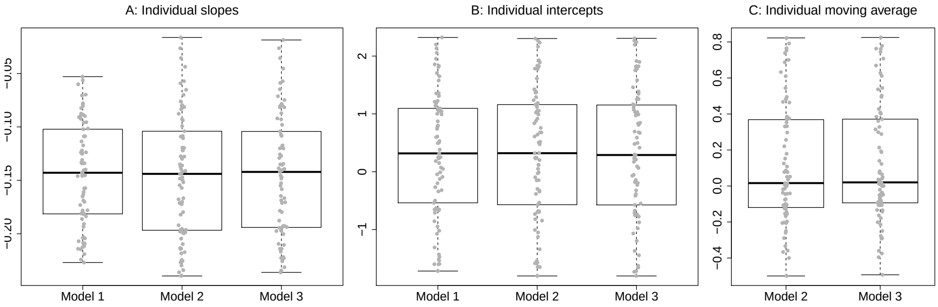


Figure 4.2: Boxplots of individual parameter values (shown by the dots) within the sample per model for (A) the slope β_n, (B) the intercept μ_n and (C) the moving average parameter ψ_n

	Model 1		Model 2		Model 3	
	mean (sd)	95% CrI	mean (sd)	95% CrI	mean (sd)	95% CrI
μ_0	0.32 (0.17)	-0.02; 0.65	0.31 (0.17)	-0.04;0.63	0.10 (0.23)	-0.36; 0.54
μ_1					0.03 (0.02)	-0.01; 0.07
σ_μ	1.24 (0.16)	0.97; 1.57	1.27 (0.15)	1.00;1.60	1.27 (0.15)	1.00; 1.60
β_0	-0.10 (0.02)	-0.15; -0.06	-0.09 (0.03)	-0.15;-0.05	-0.09 (0.02)	-0.14; -0.05
$\beta_{1,1}$	-0.09 (0.04)	-0.17; -0.01	-0.11 (0.04)	-0.20;-0.02	-0.11 (0.04)	-0.20; -0.02
$\beta_{1,2}$	-0.04 (0.03)	-0.11; 0.03	-0.06 (0.04)	-0.13;0.02	-0.06 (0.04)	-0.13; 0.02
$\beta_{1,3}$	0.01 (0.03)	-0.05; 0.08	0.01 (0.03)	-0.05;0.08	0.01 (0.03)	-0.05; 0.08
$\beta_{1,4}$	-0.06 (0.04)	-0.14; 0.01	-0.08 (0.04)	-0.17;-0.00	-0.08 (0.04)	-0.16; -0.00
$\beta_{1,5}$	-0.09 (0.04)	-0.17; -0.02	-0.10 (0.04)	-0.18;-0.03	-0.10 (0.04)	-0.18; -0.03
σ_β	0.05 (0.02)	0.01; 0.08	0.06 (0.02)	0.02;0.09	0.06 (0.02)	0.02; 0.09
σ_η	0.33 (0.03)	0.27; 0.38	0.26 (0.03)	0.21;0.32	0.26 (0.03)	0.21; 0.31

Table 4.1: Posterior mean (and standard deviation) with 95% credible interval (CrI) estimates for the model parameters of Models 1, 2 and 3.

The Intercept

In pane (B) of Figure 4.2 the boxplots of the estimated μ_n per model can be seen. For Models 1 and 2, the mean estimated μ_n is 0.32, with a range of -1.7 to 2.3 and of -1.8 to 2.3, respectively. For Model 3, the mean estimated μ_n is 0.30 with a range of -1.8 to 2.3. The posterior means (and standard deviations) for μ_0 and σ_μ for all three models can be seen in Table 4.1.

In Model 3, μ_n is estimated using Equation (4.4). As can be seen in Table 4.1, the posterior mean of μ_1 is smaller than the accompanying standard deviation and the 95% CrI for the posterior mean of μ_1 includes zero, suggesting that the duration of symptoms pre-treatment is not or only weakly related to the frequency of panic attacks at the start of therapy.

	Model 1	Model 2	Model 3
$g_{c,m}$	-0.10 (0.02)	-0.09 (0.03)	-0.09 (0.02)
$g_{s,m}$	-0.20 (0.05)	-0.21 (0.05)	-0.20 (0.05)
$g_{b,m}$	-0.14 (0.04)	-0.15 (0.05)	-0.15 (0.04)
$g_{c,a}$	-0.09 (0.04)	-0.08 (0.04)	-0.08 (0.04)
$g_{s,a}$	-0.16 (0.05)	-0.18 (0.05)	-0.17 (0.05)
$g_{b,a}$	-0.20 (0.05)	-0.20 (0.05)	-0.20 (0.05)

Table 4.2: Mean slope per group for Models 1, 2 and 3, calculated by adding β_0, which is the mean slope of $g_{c,m}$, to the $\beta_{1,\cdot}$ for the relevant condition.

The Moving Average

The moving average parameter is included in Models 2 and 3. Panel (C) of Figure 4.2 shows the distribution of the posterior means of ψ_n within the sample. For both models, the ψ_n has a mean of 0.12 and median of 0.02 in the sample, with a range of -0.50 to 0.82 for Model 2 and a range of -0.49 to 0.83 for Model 3. The estimated standard deviation of the individual ψ_n ranges from 0.23 to 0.71 for both models. Out of the 73 individuals, for both models only five 95% CrIs did not include zero.

4.5 Comparison of models

We compare the three models to see whether the model is improved by including the duration of symptoms pre-treatment as an external variable, and the moving average. When considering σ_μ, σ_β and σ_η, the differences between the models are small. For σ_μ and σ_β, the posterior means are slightly smaller for Model 1 than for Models 2 and 3. The σ_η of Model 1 falls above the 95% CrI of the posterior estimate of σ_η for Model 2 and 3, for which σ_η is similar in size. This implies that Model 1 shows slightly less noise when estimating μ_n and β_n, but more noise for the estimation of $\delta_{t,n}$ and thus for the estimated latent score.

Second, we compare the models using the Watanabe-Aikake Information Criterion (Watanabe, 2010) with the functions as provide by the package 'loo' in R (Vehtari, Gelman, & Gabry, 2015). A lower WAIC indicates a better predictive model accuracy. The WAIC is 30,312 for Model 1, 27,823 for Model 2 and 27,616 for Model 3, thereby favouring Model 3.

Combining the error standard deviations and the likelihood estimates, we can infer that the effect of including the moving average term (ψ_n) on the model fit is stronger than the effect of including the duration of symptoms pre-treatment.

To give a visual reference of the resulting fit to the sample data, Figure 4.3 shows the mean observed score and the mean estimated score per condition across time for Model 3.

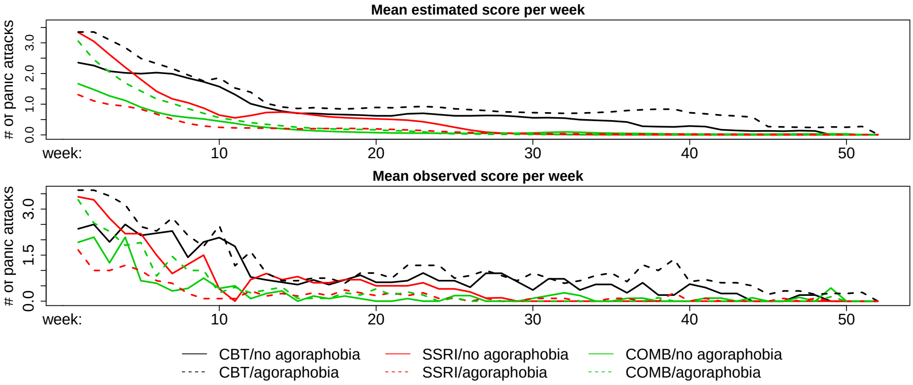


Figure 4.3: Mean observed score and mean estimated score using Model 3 per condition per week

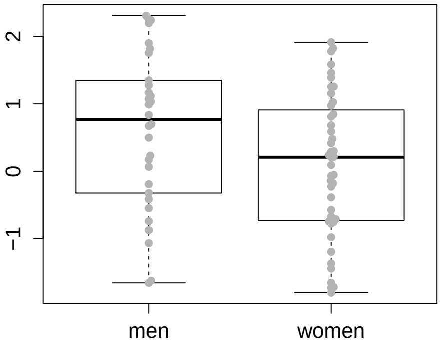


Figure 4.4: Boxplots of individual estimates (shown by the dots) for μ_n with Model 3 per gender

4.6 Post-hoc: intercept and gender

Earlier studies found that compared to women, men wait longer before asking for help and are less able in recognizing symptoms of mental stress (Khlat, Legleye, & Sermet, 2013; Swami, 2012). This may result in a larger number of symptoms experienced before treatment is sought. To see if this is reflected in our sample we compare the intercept, μ_n as found in Model 3, between men and women post-hoc.

As can be seen in Figure 4.4, the difference between the groups is small, and the variation is larger within each group than between groups. To test the difference, we use a Bayesian *t*-test (Morey & Rouder, 2011; Morey, Rouder, & Jamil, 2015), with H_0 : the true difference in mean equals 0, and H_a : the true difference in mean is unequal to zero. The prior of the effect size of the difference in mean is a Cauchy$(1/\sqrt{2})$ distribution. We found only anecdotal evidence in favor of H_a, with a Bayes Factor of 1.32. We conclude that men may have equal to slightly higher numbers of panic attacks than females at the start of the treatment.

4.7 Discussion

In this chapter, we explained how the BDM can handle relatively small data sets containing missing data points and dropouts, with a multilevel structure and non-normally distributed observed data. The BDM for modeling intensive longitudinal data is showcased by a re-analysis of data from Van Apeldoorn et al. (2013), consisting of the number of panic attacks experienced by 73 patients, measured per week. We used three models to examine the effects of three external variables and a moving average element. Model 1 included the treatment and presence of agoraphobia as predictors for the slope, in Model 2 a moving average parameter was added and in Model 3 the duration of symptoms pre-treatment was added as a predictor for the intercept. For the random effects of our predictor variables and our innovation noise distribution, we choose to use the normal distribution. If necessary, it is possible to use a different distribution, for example a Students-t distribution for smaller datasets.

In all three models the slope depends on the treatment an individual received and on the presence of agoraphobia. The CBT treatment shows a slower decrease in symptoms than the other two treatments, both for the individuals without and with agoraphobia. Furthermore, the SSRI treatment shows a stronger decrease for individuals without than for individuals with agoraphobia. This contrasts to the BOTH treatment, which shows a stronger decrease in symptoms for individuals with agoraphobia. All effects of agoraphobia are only trends. The results pertaining to the effect of the treatment and presence of agoraphobia in individuals on the slope are consistent with those found by Van Apeldoorn et al. (2013).

In Model 2, the moving average parameter was added. As a result, the standard deviation of the error of the parameter estimation increased, but the standard error of the latent variable decreased. The WAIC decreased strongly from Model 1 to 2, which indicates that the moving average parameter is an important part of the model, even though the parameter is close to zero for a large number of individuals in the sample.

In Model 3, the duration of symptoms pre-treatment was included as a predictor for the intercept. Though the WAIC indicated that Model 3 fitted best, the 95% CrI of the estimated posterior mean for the predictor effect included zero. This indicates that it is uncertain whether the duration of symptoms has an effect on the intercept. A post-hoc test was conducted using Model 3 to compare the intercepts for men and women. The Bayesian t-test suggest anecdotal evidence for a difference in means.

An important question is how the BDM improves estimation compared to the frequentist multilevel model used by Van Apeldoorn et al. (2013). First, in contrast to the multilevel model, the BDM allows for adjustment to the estimation during the time series through the noise of the latent state vector. Opposed to the white noise as used in the multilevel model, the innovation noise is included in the current latent vector, which is used to estimate the latent vector at the next time point. As a result, the divergence from an expected score at a certain time point is used in the BDM to adjust estimations of later time points.This larger flexibility of the BDM results in a better fit of the observed data. Second, the BDM estimates the individual parameters, opposed to only estimating the distribution from which they are drawn. This allows for inspection of these individual parameters, for example for post-hoc analyses. Third, the combination of random coefficients (β_n, μ_n) and fixed coefficients (ψ_n) as used in Model 3, would not have been possible within the frequentist multilevel framework.

As an alternative to the BDM, the data could have been analyzed with a State Space Model (SSM), which is the frequentist counterpart to the BDM. The SSM is a highly versatile model for intensively measured, functionally related data, such as intensive longitudinal data (Durbin & Koopman, 2012; Petris et al., 2009; Pole, West, & Harrison, 1994). The SSM allows for modeling of non-normally distributed data (Durbin & Koopman, 2012) and time-dependent parameters, such as found in threshold models (Haan-Rietdijk, Gottman, Bergeman, & Hamaker, 2014; Hamaker & Grasman, 2012). However, to our knowledge no SSM has been developed yet that incorporates both missing data and non-normally distributed observed data simultaneously. The hierarchical Poisson SSM proposed by Terui et al. (2010) allows for non-normally distributed data. However, this estimation approach can be applied to complete data only, implying that scores should be available for all individuals at all time points.

The number of individuals in the data set used for our empirical study is small compared to the number of conditions; 73 individuals divided over six conditions.

By combining multiple individuals, the estimation of the distribution of a parameter improves (Krone et al., 2016d), but it is not yet clear how exactly this relates to each other in complicated models as the one discussed in this chapter. Another important property of the data is the length of the time series; earlier studies stated that a length of at least fifty time points is preferred for a simple single-subject moving average model (Box & Jenkins, 1976; Krone et al., 2016a). As our data set contained time series with a minimum of ten and a maximum of fifty-two data points, the model may at points be too complex for the data.

Models used in empirical studies often include a number of different parameters. For trend models as in this chapter, important parameters are the intercept, the slope parameter and eventually added autoregressive or moving average parameters, of which several parameters are often influenced by external variables. When the dynamics are the main interest, parameters such as the mean score, autoregressive and moving average coefficients, and for multivariate models the cross-lag coefficients, are of importance. The efficiency with which these parameters are estimated in such complex models, both with regard to individuals and to time points, is not yet thoroughly studied. Thus, the amount of data needed for reliable estimates in such complex models is an interesting topic for further studies.

Chapter 5

A multivariate model for emotion dynamics

Abstract

In emotion dynamic research one distinguishes various elementary emotion dynamic features, which are studied using intensive longitudinal data. Typically, each emotion dynamic feature is quantified separately, which hampers the study of relationships between various features. Further, the length of the observed time series in emotion research is limited, and often suffers from a high percentage of missing values. In this chapter we propose a vector autoregressive Bayesian dynamic model, that is useful for emotion dynamic research. The model encompasses six elementary properties of emotions, and can be applied with relatively short time series, including missing data. The individual elementary properties covered are: within person and innovation variability, inertia, granularity, cross-lag correlation and the intensity. The model can be applied to both univariate and multivariate time series, allowing to model the relationships between emotions. Further, it may model multiple individuals jointly. One may include external variables and non-Gaussian observed data. We illustrate the usefulness of the model with an empirical example of three emotions of three individuals (47 to 70 measurements), with missing time points within the series.

This chapter is submitted as: Krone, T., Albers, C. J., & Timmerman, M. E. A multivariate model for emotion dynamics.

5.1 Introduction

Emotions are an important part of our daily lives. The importance of emotions for our health and well-being is recognized more and more (Tugade, Fredrickson, & Feldman Barrett, 2004; Grühn, Lumley, Diehl, & Labouvie-Vief, 2013; Lewis, Haviland-Jones, & Feldman Barret, 2008, Ch. 29). Unlike, perhaps, personality and values, emotions fluctuate across time, changing both within and between days under the influence of external events and internal evaluations. As such, a central function of emotions is to alert us to important events and changes, and to motivate us to deal with them (Larsen, 2000; Frijda, 2007; Scherer, 2009; Kuppens & Verduyn, 2015). Understanding the dynamics of emotions is therefore important, not in the least because it provides a window on how emotions may become dysregulated, which is considered a central feature of several mental disorders (Houben, Van Den Noortgate, & Kuppens, 2015; Wichers, Wigman, & Myin-Germeys, 2015).

To study emotion dynamics, intensive longitudinal data are used, sampled sufficiently frequently to characterize the dynamics of interest (Hamaker et al., 2015). Technological advantages facilitate the collection of such data, both in an experimental setting in a lab, as well as in daily life. In lab studies, for instance, video mediated recall and physiological recording can provide information on the dynamics of emotional episodes. In daily life, the widespread availability of mobile devices, first palmtops and now smartphones, enables researchers to collect multiple measurements per day in so-called ecological momentary assessment (EMA, also known as experience sampling) studies (Larson & Csikszentmihalyi, 1983; Shiffman et al., 2008; Bolger & Laurenceau, 2013; Bos et al., 2015).

When intensive longitudinal data is gathered with a structure as complex as found in emotion data, the choice of a proper analysis is of paramount importance. This choice is far from straightforward, given the diversity of techniques available (Hamaker et al., 2015). The analysis typically focuses on identifying particular elementary features of emotion dynamics, with the aim to reveal distinct information on affective functioning and regulation (Kuppens & Verduyn, 2015). For instance, one may be interested in the level of variability emotions display within an individual, or in how different emotions covary across time. However, choosing a proper analysis is hampered by the fact that these elementary features can often be quantified in different ways. Further, the quantifications of these elementary features are typically considered separately. This implies that relationships between these features are kept hidden.

To provide a good picture of emotion dynamics, we propose to use a single model of which most parameters have a clear interpretation in terms of a number of key features that are considered central to emotion dynamics. To this end, we propose to use Bayesian dynamic modelling (West & Harrison, 1997). The Bayesian

Dynamic Model (BDM) we propose offers a representation of multivariate time series, and may be modeled for multiple individuals simultaneously. Furthermore, the BDM as proposed in this chapter can be conveniently interpreted in terms of six important emotion dynamic features. This offers insight into the dynamics of single emotions, as well as the dynamics between multiple emotions within an individual. Finally, the model can offer insight into interindividual differences in emotion dynamics. Using a Bayesian estimation method offers flexibility with regard to the distributions used in specifying the model.

5.1.1 Emotion Dynamic Features

The patterns and regularities of an individual's expression of emotions across time can be captured by various elementary properties. We denote these elementary properties as emotion dynamic features (EDFs). There is a vast range of EDFs discussed in the literature (Houben et al., 2015; Kuppens & Verduyn, 2015; Grühn et al., 2013; Carstensen, Pasupathi, Mayr, & Nesselroade, 2000; Brose et al., 2015). The taxonomy as discussed by Kuppens and Verduyn (2015) organizes these EDFs into four categories: emotional variability, emotional inertia, emotional cross-lag, and emotional granularity. If we complement these four with emotional intensity, we provide a fairly complete picture of an individual's expression of emotions across time. Our aim is to propose a way to succinctly capture the EDFs from the five categories in one single model. In the following section, we discuss each category and how it is captured in our model.

Emotional variability Emotional variability reflects to what extent the intensity of an emotion as experienced by an individual, varies across time. Emotional variability has been found to increase with increasing stress levels (Scott et al., 2014) and decrease with increasing age (Carstensen et al., 2000; Scott et al., 2014; Brose et al., 2015). High emotional variability has been linked with lower emotional well-being (Houben et al., 2015) and higher prevalence and severity of mood disorders (Kuppens & Verduyn, 2015). To quantify emotional variability, the within person variance or standard deviation is typically used (Carstensen et al., 2000; Röcke, Li, & Smith, 2009; Grühn et al., 2013; Scott et al., 2014; Kuppens & Verduyn, 2015).

The within person variance can be seen as a global summary of the degree of emotional variability. This variability can be decomposed into various model elements (Jahng, Wood, & Trull, 2008). Herewith, it is useful to distinguish the predictable part from the random part. The predictable part can be interpreted in terms of the emotional inertia and cross-lag, as will be discussed in the next paragraphs. The random part covers the instantaneous change, and consists of the innovation variance and the white noise variance.

The innovation variance and the white noise variance express the sizes of the instantaneous changes at each measurement point. The key difference between the two is that the innovation variance captures the part of the change that is carried through to the next measurement point, while the white noise variance captures the part of the change that is not carried through to the next measurement point.

As the global summary measure of within person variability we will use the within person variance. Though our model includes both the innovation variance and white noise variance, we will only interpret the innovation variance, as a measure of innovation variability. We leave aside the white noise variance in our interpretation, because change due to white noise is typically attributed to measurement error.

Emotional inertia Emotional inertia refers to the tendency of an emotion to retain its status quo, reflecting the resistance to change (Cook et al., 1995; Suls et al., 1998; Kuppens, Allen, & Sheeber, 2010). High inertia has been linked to impaired emotion regulation (Kuppens, Allen, & Sheeber, 2010; Suls et al., 1998; Koval et al., 2015; Gross, 2015), inflexibility in adapting emotions (Kashdan & Rottenberg, 2010) and rumination (Koval, Kuppens, Allen, & Sheeber, 2012). Emotional inertia is generally quantified as the autoregression between successive measurements of an emotion. Confusingly, this has been incidentally denoted as the autocorrelation (e.g., Kuppens, Allen, and Sheeber (2010), Kuppens and Verduyn (2015)).

Emotional cross-lag Emotions can be regulated through feedback-loops: the increase of one emotion may infer an increase or decrease in another emotion (Gross, 2015; Kuppens & Verduyn, 2015). Although few studies have yet been conducted on emotional cross-lag, it is an important part of emotion regulation (Gross, 2015; Kuppens & Verduyn, 2015). For example, it has been found to be increased in major depression patients in terms of higher levels of overall emotion network density (Pe et al., 2015). Emotional cross-lag is quantified via the cross-lag regression; the lagged regression between two emotions (Pe & Kuppens, 2012). Analogously to the term auto-regression, it is sometimes mistakenly called the cross-lag correlation (e.g., Kuppens and Verduyn (2015)). When the cross-lag regression is positive, this is called augmentation: the experience of one emotion increases the strength of another emotion on a later time point. A negative cross-lag regression is called blunting: the experience of one emotion decreases the strength of another emotion on a later time point.

Emotional granularity Emotional granularity is the ability of differentiating between different emotions and identifying emotions with specificity and precision. This is also known as emotional differentiation or emotional covariation (Feldman, 1995; Barrett, Gross, Christensen, & Benvenuto, 2001; Barrett & Gross, 2001; Kuppens & Verduyn, 2015). Higher emotional granularity is linked to increased emotion regulation (Barrett et al., 2001) and different, more effective coping mechanisms (Tugade et al., 2004). Furthermore, higher emotional granularity is associated with lower levels of neuroticism (Carstensen et al., 2000), and lower incidence of social anxiety disorder (Kashdan & Farmer, 2014) and depression (Erbas et al., 2014).

Emotional granularity has been quantified in different ways. For example, the differentiation index equals the number of components found by a principal component analysis (PCA) on the covariances between emotions of a single individual (Grühn et al., 2013; Brose et al., 2015). Related to this is the concept of the unshared variance: the percentage of variance unexplained by the first component of such a PCA (Grühn et al., 2013). In practice, the choice between the differentiation index and the unshared variance is a pragmatic one. For a large number of emotions, the differentiation index appears to be more informative, and for a small number of emotions, the unshared variance. For these measures, higher scores indicate a higher granularity.

Other quantifications can be calculated directly from the observed data: the covariance between two emotions within a person (Grühn et al., 2013; Erbas et al., 2014), the correlation between two emotions (Barrett et al., 2001), and the intraclass correlation (ICC) between all emotions (Tugade et al., 2004; Erbas et al., 2014). A higher covariance, correlation and ICC indicate a lower level of differentiation between emotions, and thus a lower granularity. The correlation has the advantage of being standardized, allowing for a direct comparison between pairs of emotions both within and between individuals. However, a low within person variance reduces the size of the absolute correlation, which renders interpretation difficult. This issue is not encountered when using the covariance (Scott et al., 2014). As all named quantifications measure the covariation between emotions, we do not need to include them all. In our model, the granularity will be quantified via both the covariance and correlation.

Emotional intensity The EDFs discussed thus far capture the dynamics of emotions over time. In addition, how strong an emotion is felt on average may also differ, both between emotions within an individual, and between individuals, and can provide important information on people's emotional lives. Emotional intensity for positive emotions is positively related to emotion regulation, (Barrett et al., 2001), as well as with extraversion, agreeableness and conscientiousness, but negatively related with neuroticism (Carstensen et al., 2000). The intensity for negative emotions is higher for individuals with social anxiety disorder (Kashdan & Farmer, 2014) as well as for individuals with depression, high scores on neuroticism, and low scores on self-esteem (Erbas et al., 2014). To assess emotional intensity across time, we take into account the average intensity (Carstensen et al., 2000; Barrett et al., 2001), quantified as the mean score over time (Kashdan & Farmer, 2014; Erbas et al., 2014).

This chapter Each of these features provides unique information on how emotions (co)vary, carry over from one moment to the next, or mutually influence each other, and together they provide insight into many crucial aspects of emotional functioning and flexibility. As such, we propose a BDM that captures within person variability, innovation variability, inertia, cross-lag, granularity, and average intensity for multiple emotions and individuals in a single model. First, we will introduce the model and its possibilities. Then, we will present an empirical application of the model. We will conclude with a discussion on the model, its advantages and disadvantages, and recommendations for future research.

5.2 Model

To combine the aforementioned concepts for multiple variables and multiple subjects in one model, we will use a Bayesian interpretation of a State Space Model, called the Bayesian Dynamic Model (BDM) (West & Harrison, 1997). We will use the BDM to estimate an vector AR(1) model, which can be rewritten into a State Space Model (Harvey, 1990; Durbin & Koopman, 2012).

The BDM has two equations: the observation equation and the system equation. For univariate data, the inclusion of so-called white noise in the observation equation improves estimation of the autoregression (Schuurman et al., 2015). Following this, we include white noise in our multivariate BDM as well. Furthermore, we assume equidistant time points in our model. For adaptions that may be incorporated in this model to allow continuous time series with non-equidistant time points, we refer to Kuppens, Oravecz, and Tuerlinckx (2010) and Oravecz, Tuerlinckx, and Vandekerckhove (2011).

We model $Y_{i,t,n}$, the score on emotion i ($i = 1, 2, ..., I$), at time point t ($t = 1, 2, ..., T_n$), for individual n ($n = 1, 2, ..., N$). The first equation, the observation equation, links the observed score $Y_{i,t,n}$ to the latent variable $\theta_{i,t,n}$. The observation equation for the score vector $\boldsymbol{Y}_{t,n} = [Y_{1,t,n}, Y_{2,t,n}, ..., Y_{I,t,n}]'$ is as follows:

$$\boldsymbol{Y}_{t,n} = \quad \boldsymbol{\mu}_n + \boldsymbol{\theta}_{t,n} + \boldsymbol{\varepsilon}_{t,n}, \quad \boldsymbol{\varepsilon}_{t,n} \sim N(\boldsymbol{0}, \boldsymbol{H}_n) \tag{5.1}$$

where $\boldsymbol{\mu}_n(I \times 1)$ denotes the mean vector of the I emotions, $\boldsymbol{\theta}_{t,n}$ $(I \times 1)$ the latent variable vector, $\boldsymbol{\varepsilon}_{t,n}$ $(I \times 1)$ the white noise vector and $\boldsymbol{H}_n$ the covariance matrix of $\boldsymbol{\varepsilon}_{t,n}$. As $\boldsymbol{\varepsilon}_{i,t,n}$ is assumed to be independent across emotions, $\boldsymbol{H_n}$ is a $I \times I$ diagonal matrix with $\sigma^2_{\varepsilon_{i,n}}$ as diagonal elements.

The system equation models the autoregression and cross-lag regression and

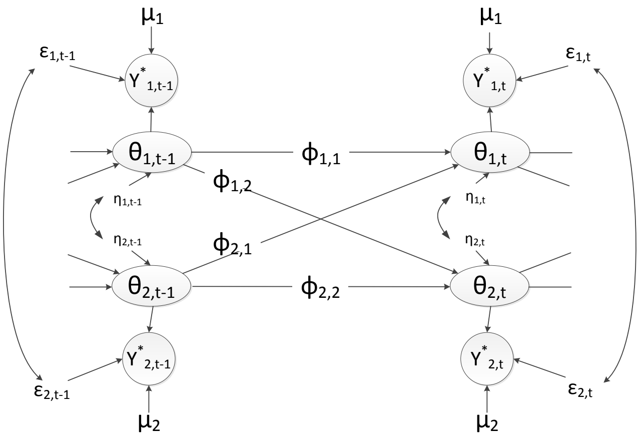


Figure 5.1: Schematic representation of the model as expressed in Equations 6.4 and 6.2 for a single individual n (subscripts left out for clarity reasons) and $i = 1, 2$ emotions.

the innovation over time of the latent variable $\boldsymbol{\theta}_{t,n}$:

$$\boldsymbol{\theta}_{t,n} \quad = \boldsymbol{\Phi}_n \times \boldsymbol{\theta}_{t-1,n} + \boldsymbol{\eta}_{t,n}, \quad \boldsymbol{\eta}_{t,n} \sim N(\mathbf{0}, \boldsymbol{Q}_n), \tag{5.2}$$

where $\boldsymbol{\Phi}_n$ $(I \times I)$ is the autoregression and cross-lag regression matrix, $\boldsymbol{\eta}_{t,n}$ $(I \times 1)$ is the innovation vector and $\boldsymbol{Q}_n$ $(I \times I)$ the covariance matrix of the innovation. All error terms, $\boldsymbol{\eta}_{t,n}$ and $\boldsymbol{\varepsilon}_{t,n}$, are assumed to be mutually independent. A graphical representation of the model can be seen in Figure 6.1.

We remark that $\boldsymbol{\Sigma}_n$ $(I \times I)$ is the model implied variance-covariance matrix of the observed scores for individual n, which is computed through

$$\text{vec}(\boldsymbol{\Sigma}_n) = (\boldsymbol{I} - \boldsymbol{\Phi}'_n \otimes \boldsymbol{\Phi}'_n)^{-1}\text{vec}(\boldsymbol{Q}_n + \boldsymbol{H}_n), \tag{5.3}$$

where $\text{vec}(\boldsymbol{\Sigma}_n)$ is the vectorized version of $\boldsymbol{\Sigma}_n$ and $\otimes$ denotes the Kronecker product. This is an adaptation of the Lyapunov equation used for the traditional VAR model with only one error term, as discussed in Hamilton (1994, p. 265).

Note that all model parameters are assumed to be equal over time. This implies that it is assumed that the emotion dynamics are constant across the time span measured. In case this assumption would be too rigid, alternative models are available. For example, in regime switching or threshold models, the autoregression may change as the state changes (Hamaker & Grasman, 2012; Haan-Rietdijk et al., 2014).

Concept	Quantification	Parameter
Within person variability	Variance of $Y_{i,n}$	$\Sigma_{ii,n}$
Innovation variability	Innovation of $Y_{i,n}$	$Q_{ii,n}$
Inertia	Autoregression	$\Phi_{ii,n}$
Emotional cross-lag	Cross-lag regression	$\Phi_{ij,n}$
Granularity	Covariance of $Y_{i,n}$	$\Sigma_{ij,n}$
	Correlation of $Y_{ij,n}$	$Cor(Y_{ij,n})$
Intensity	Mean estimated score	$\mu_{i,n}$

Table 5.1: Quantification of emotion dynamics features for emotion i, in relation to emotion j where applicable, in the notation of Equations 6.4, 6.2 and 5.3.

With this statistical model, the link can be made between emotion theory and application. We link each of the discussed EDFs to a parameter, as summarized in Table 5.1. The within person variability for individual n and emotion i is expressed via $\Sigma_{ii,n}$ and the innovation variability via $Q_{ii,n}$. The autoregression for individual n and emotion i is $\Phi_{ii,n}$, and the cross-lag regression is $\Phi_{ij,n}$ for $i \neq j$. The covariance and correlation, used for the granularity, is obtained via $\Sigma_{ij,n}$ for $i \neq j$. The intensity for individual n on emotion i is the mean $\mu_{i,n}$. Hence, the model enables the simultaneous study of all discussed emotion dynamics. In the next paragraphs, we will discuss the extended possibilities of this model, and the identification in and application to empirical data.

5.2.1 Multiple individuals

As shown, the model can be defined without any difficulty for multivariate data collected among multiple subjects. When modeling the data of multiple individuals, two possible approaches exist. First, the individuals may be assumed to be drawn at random from a certain population. As such, the parameters of the individuals are assumed to be drawn randomly from the population distribution of the parameter concerned. These assumptions may be expressed in the model via a level 2 model, for example by assuming, as is standard in multilevel modeling, that each $\Phi_{ij,n}$ is drawn from a normal distribution: $\Phi_{ij,n} \sim N(\Phi_{ij}, \sigma_{\Phi_{ij}})$ (Lodewyckx et al., 2011).

Second, there may be no assumption made with regard to the sampling of the individuals. To reflect this, the parameters of the individuals are estimated freely. This implicitly defines the level 2 model, as the joint distribution of the individually estimated parameters for all individuals is hereby defined. Due to the free parameter estimation, these model estimates would be the same as when the time series of each individual were modeled separately. In the current chapter, we follow this latter strategy, and hence make no assumption as to the sampling of the individuals.

When the dynamics of a large number of individuals are studied, it may be of interest to consider the relation between different individual parameters. For instance, one may consider the strength of the relationship between the inertia and the intensity of an emotion.

5.2.2 Non-normally distributed data

The BDM model can be extended to non-normal distributions in two ways. First, when the observations are assumed to be non-Gaussian realizations of an underlying Gaussian process, a link function can be used to transform the latent Gaussian scores into estimated observed non-Gaussian scores. Examples are a probit-link for ordinal data (Chaubert, Mortier, & Saint André, 2008), or the log-link for count data (assuming a Poisson distribution) (Terui et al., 2010). However, this adds more complexity to the model, requiring a larger sample size to estimate the model parameters with reasonable precision. Second, when the underlying process is assumed to be non-Gaussian, the distributions used in the model, for example the white noise and innovation distributions, can be adjusted accordingly (Durbin & Koopman, 2012; West & Harrison, 1997).

5.2.3 Missing data

A link function can be used to accommodate for missing data. The link function links the observed $Y_{i,t,n}$ to a latent $Y^*_{i,t,n}$. When $Y_{i,t,n}$ and $Y^*_{i,t,n}$ are assumed to be equally distributed, an identity-link function is used, indicating that $Y_{i,t,n} = Y^*_{i,t,n}$. Using the link function enables to deal with missing data, since the latent variable $Y^*_{i,t,n}$ is not linked to the observed variable $Y_{i,t,n}$ at time points with missing data. However, the uncertainty increases with more missing data points in a row, since the estimation cannot be checked against the observed data anymore.

5.2.4 External variables

Empirical research often probe how features of emotions are related to, or a function of, other variables, such as experimental manipulation or individual differences. Due to the flexible nature of the model, external variables can be included in two ways. First, the external variable can be included as an active covariate. This can be done as a direct effect, for example letting $Y_{i,t,n}$ being dependent on $\boldsymbol{\mu}_n$, $\boldsymbol{\theta}_{t,n}$ and a covariate, and as a moderator effect, for example letting elements of $\boldsymbol{\Phi}_n$ be dependent on the level of a covariate. Second, inactive covariates can be implemented post-hoc, by examining the relation between any model parameter and an external variable after the model estimation. This can be done, for instance, by using partial correlations or linear regression, thereby accounting for confounding variables.

5.2.5 Model convergence

Estimating the model to empirical data may yield convergence problems. In general, these problems may be due to the identifiability of the model and/or a lack of data. The VAR-BDM expressed in Equations 6.4 and 6.2, is identified. Therefore, when estimation issues arise with this model, this is due to a lack of data. The problem exacerbates when a large number of emotions are included and/or missing data occur. What the minimum requirements are to estimate the model parameters from an empirical data set, is unclear at the moment.

To check the convergence, two methods may be used: assessment of the potential scale reduction factor, $\hat{R}$, and visual inspection of the trace plots. The $\hat{R}$ shows the ratio of how much the estimation may change when the number of iterations is doubled, with an (ideal) value of 1 indicating that no change is expected (Gelman & Rubin, 1992; Stan Development Team, 2014). The trace plots show the Bayesian Markov Chain Monte Carlo (MCMC) estimates for each parameter at each iteration. If a parameter reaches convergence, the estimates over iterations are highly similar across chains. As a result, the trace plot will look like a fat caterpillar where all chains completely overlap, except at the fringe of the caterpillar (as shown in Figure 5.3).

5.2.6 Prior specification

In Bayesian modelling, prior distributions, quantifying the a priori degree of belief in parameter values, have to be specified. In empirical situations where there is relevant context information (such as some results of a pilot study), this information can be incorporated by specifying informative priors. Alternatively, one can aim for weak-informative priors, to reduce the influence of the choice of priors on the estimates.

5.2.7 Model estimation

The model is estimated using Bayesian MCMC. For the MCMC estimates we use Hamiltonian Monte Carlo (HMC), a generalization of the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) that allows for an efficient estimation of the parameters (Gelman, Carlin, et al., 2013). This is incorporated in the software RStan (Stan Development Team, 2014; R Core Team, 2015).

The R-code and Stan-code for the model can be found in Appendices A and B, respectively.

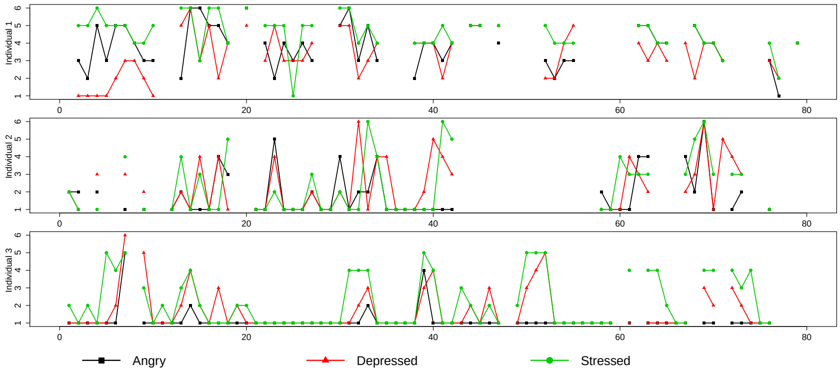


Figure 5.2: Observed time series for the three individuals.

5.3 Empirical example

In this chapter, we re-analyse data described in Brans, Koval, Verduyn, Lim, and Kuppens (2013) and Erbas et al. (2014). As part of a larger study, the emotions of 50 individuals were observed using experience sampling. The data gathering process consisted of three parts. In a first lab session the participants signed an informed consent form and were handed out a palmtop which they would use to record their emotions, along with instructions for its use. Second, during seven days, participants carried the palmtops and recorded their emotions using the Experience Sampling Program (ESP) (Barrett & Barrett, 2000). The waking hours of the individuals on each day were divided into ten intervals. During each interval, at a random time point, the ESP would ask them to rate their emotions in terms of how angry, depressed and stressed, for example, they felt at that moment. Each emotion was rated on a 6-point Likert scale, ranging from 0 to 5, with higher values indicating a stronger feeling of that emotion. Finally, in a second lab session, the participants returned the palmtops and were each given AU$ 40,- for their participation.

Data preparation From the 50 individuals, we selected three. These three individuals had observed scores that were distributed over the entire, or almost entire, 6-point Likert scale for the negative affect emotions considered, angry, depressed and stressed. Furthermore, they had at least seven consecutive measurement days with nine measurements sampled per day. However, the scores were not all completed by the participants: they had 36%, 37% and 8% missing data, resulting in 50, 47 and 70 time points, respectively. In Figure 5.2 the different observed scores for the individuals are depicted. As can be seen, the first individual shows missing data all through the sample, where the second individual shows large patches of

Bayes factor		Strength
In favor of H_0	In favor of H_1	of evidence
$< 1/100$	> 100	Decisive evidence
$1/100 - 1/30$	$30 - 100$	Very strong evidence
$1/30 - 1/10$	$10 - 30$	Strong evidence
$1/10 - 1/3$	$3 - 10$	Substantial evidence
$1/3 - 1$	$1 - 3$	Anecdotal evidence
1	1	No evidence

Table 5.2: Strength of evidence in favor of H_0 or H_1 for difference values of the Bayes factor

complete and missing data. The third individual shows little missing data.

Prior specification For the elements of $\boldsymbol{\Phi}_n$ we use a symmetrized reference prior (Berger & Yang, 1994). The scale parameter of $\boldsymbol{H}_n$ was given a half-Cauchy(0, 2.5) prior, and the correlation matrix of $\boldsymbol{H}_n$ was given a Lewandowski-Kurowicka-Joe (LKJ) correlation prior (Lewandowski, Kurowicka, & Joe, 2009). We set $\boldsymbol{Q}_n$ to a diagonal matrix, to simplify the model slightly and make the estimation easier, with an $\lambda(3, 3)$ prior for each element $Q_{ii,n}$. For $\mu_{i,n}$ we used a normal prior with mean 0 and variance 4.

5.4 Results

We start by presenting results concerning the convergence of the parameter estimates. Then, we present the posterior parameter estimates related to the emotion dynamic features in the same order as presented in the introduction, i.e., the parameters related to the within person variability, innovation variability, inertia, cross-lag, granularity and intensity.

For several EDFs, correlations are calculated or estimated. Each correlation is accompanied by a Bayes factor (BF), calculated and interpreted along the lines of Wetzels and Wagenmakers (2012). In all cases, BF< 1 is relative evidence for H_0: ‘the correlation is zero’, and BF> 1 is relative evidence for H_1: ‘the correlation is non-zero’. In Table 5.2 the interpretation of the BF as given by Wetzels and Wagenmakers (2012) is provided. The BF is highly dependent on sample size, with stronger evidence for the same correlation when the sample is larger. As such, the difference in BF for similar correlations of different individuals can, largely, be explained by the difference in sample size: the individuals have a total of 50, 47 and 70 observed time points, respectively. In this chapter we only use the BF to assess the relative evidence for H_0 and H_1, and not to assess the evidence for inter-individual differences.

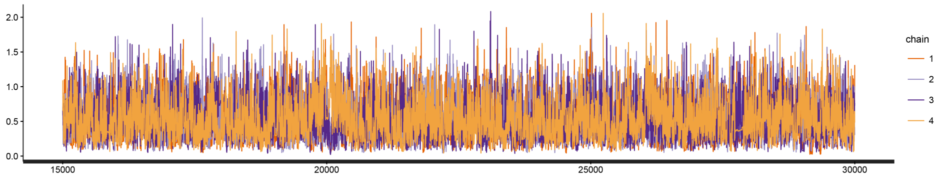


Figure 5.3: Trace plot of $Q_{11,1}$ as estimated in the empirical data, using six chains and 10.000 iterations (5.000 warm-up).

5.4.1 Convergence

In our analysis, we used four MCMC-chains of 30.000 iterations each. Further, we modeled each individual separately, due to time constraints. Modeling all individuals in one analysis takes more time per iteration and more iterations to reach convergence. The model we used for this data set, as shown in Equations 6.4 and 6.2, estimates each individual seperately, as such separating the individuals does not influence the results. To check whether convergence was reached, we used the $\hat{R}$ and the trace plots. All elements of the model parameters ($\boldsymbol{\Phi}_n$, $\boldsymbol{Q}_n$ and $\boldsymbol{\mu}_n$) reached convergence for all individuals and emotions, with an $\hat{R}$ below 1.02 for all estimated model parameters. One of the trace plots, representable for all relevant trace plots, is given in Figure 5.3 which, as can be seen, gives the expected fat caterpillar.

5.4.2 Emotional variability

Emotional variability is quantified using the EDFs within person variability and innovation variability. The individual values of the within person variability, quantified as $\Sigma_{ii,n}$, for each emotion is shown in Figure 5.4, panel A (continuous lines). As can be seen, Individual 2 shows a higher within person variability than the other two subjects, on all three emotions. For Individual 1 the within person variability of depressed is highest, followed by the within person variability of angry. For Individuals 2 and 3 the within person variability of stressed is highest, followed by the within person variability of depressed.

The innovation variability, quantified as the estimated $Q_{ii,n}$, is also shown in Figure 5.4, panel A (dashed lines) along with its 95% credible interval (CrI). For angry, the within person variability and innovation variability are nearly identical for Individual 3. The relations between the individuals and between the emotions within individuals are similar to the relations found for within person variability. As with the within person variability, the innovation variability is higher for Individual 2 than for the other two individuals. This suggests that individual differences in emotion dynamics may at least to some extent be timescale-invariant, i.e., individuals with a high variability over the whole time period, as indicated

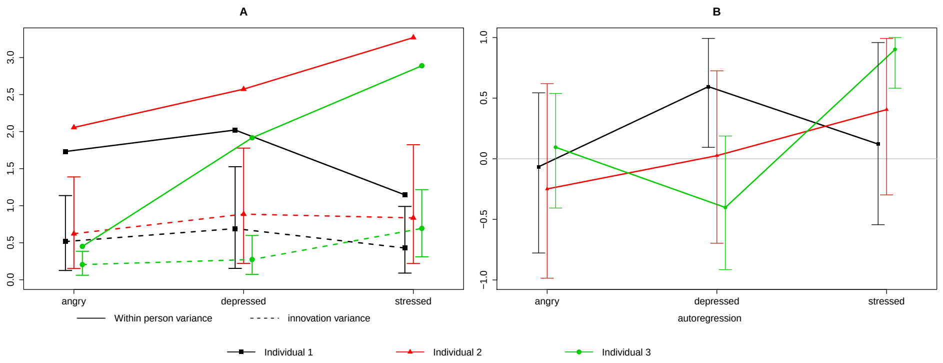


Figure 5.4: Individual scores for (A) within person variability (continuous line) and innovation variability (dashed line), and (B) inertia with the 95% credibility interval for the innovation variability and inertia.

by the within person variability, may also show high variability between two time points, as indicated by the innovation variability. This has been found elsewhere as well (Kuppens, Oravecz, & Tuerlinckx, 2010).

5.4.3 Emotional inertia

The inertia is expressed through the autoregression. As can be seen in Panel B of Figure 5.4, the autoregression shows differences, both in size and in pattern over emotions, across the individuals. However, for all emotions the 95% CrI show overlap. The autoregression for angry is small for all three individuals, as shown by the BFs in the top part of Table 6.2. For depressed and stressed, the differences between the individuals are larger. For depressed, Individual 1 shows a high positive autoregression and Individual 3 shows a medium negative autoregression. For Individual 2, there seems to be no or only a small autoregression for depressed. The autoregression of stressed is positive for all three individuals, with a medium autoregression for Individual 2, but a very high autoregression for Individual 3. Taken together, these results provide evidence that emotions are strongly self-related over time, be it positive or negative. This is reflected in the literature. Indeed, most previous research has found that emotional states tend to be mildly or strongly predictive over time in daily life (e.g., Suls et al., 1998; Kuppens, Allen, & Sheeber, 2010; Koval et al., 2012).

5.4.4 Emotional cross-lag

The impact of one emotion on another is measured through the cross-lag regressions. The cross-lag regressions and their 95% CrI are shown in Figure 5.5 and the lower part of Table 6.2. Individual shows 1 small values for all cross-lag regressions,

	Individual 1	Individual 2	Individual 3
	Estimated autoregressions		
A	-0.07 (0.12)	-0.25 (0.47)	0.10 (0.13)
D	0.60 (3332.48)	0.03 (0.12)	-0.40 (34.88)
S	0.12 (0.16)	0.41 (6.04)	0.90 (2.86×10^{23})
	Estimated cross-lag regressions		
A_{t-1} on D_t	-0.13 (0.16)	-0.48 (41.43)	0.44 (108.87)
A_{t-1} on S_t	-0.16 (0.20)	-0.37 (2.93)	0.40 (32.68)
D_{t-1} on A_t	0.40 (7.55)	0.01 (0.11)	-0.13 (0.16)
D_{t-1} on S_t	0.05 (0.12)	0.41 (6.72)	-0.68 (1.33×10^8)
S_{t-1} on A_t	0.32 (1.40)	0.33 (1.44)	0.18 (0.29)
S_{t-1} on D_t	0.07 (0.13)	0.43 (10.07)	0.70 (7.75×10^8)

Table 5.3: Autoregressions and cross-lag regression for the three individuals on angry (A), depressed (D) and stressed (S), with the accompanying Bayes factors in parentheses.

except for D_{t-1} on A_t and S_{t-1} on A_t, which show medium sized augmenting effects of depressed and stressed on angry. Individuals 2 shows a negative cross-lag regression effect of medium size for A_{t-1} on both D_t and S_t, indicating that angry has a blunting effect on both depressed and stressed. For Individual 3, angry augments both depressed and stressed, with a medium sized, positive cross-lag regression effect. The effect of depressed on angry is small for both Individual 2 and 3, but positive and medium sized for Individual 1. For Individual 2, depressed has an medium sized augmenting effect on stressed, while for Individual 3 depressed has a high negative cross-lag effect on stressed, indicating a blunting effect. For both Individuals 1 and 2, stressed has an augmenting cross-lag effect on angry. Finally, stressed augments depressed for both Individual 2 and 3, with a medium effect for Individual 2 and a high effect for Individual 3. The substantial individual differences that are found in the extent to which different emotions augment or blunt each other across time, are consistent with previous research (Pe & Kuppens, 2012).

5.4.5 Emotional granularity

The granularity is used to express the covariation between emotions and quantified via the covariance and the bivariate correlation between the different couples of emotions per individual. Here, higher values indicate a lower granularity. As can be seen in Panel B of Figure 5.5, all correlations (and thus covariances) between the paired emotions are positive, for all three individuals. This strongly resonates with previous research showing generally positive relations between like-valenced emotional states across time within individuals (e.g., Vansteelandt, Van Mechelen,

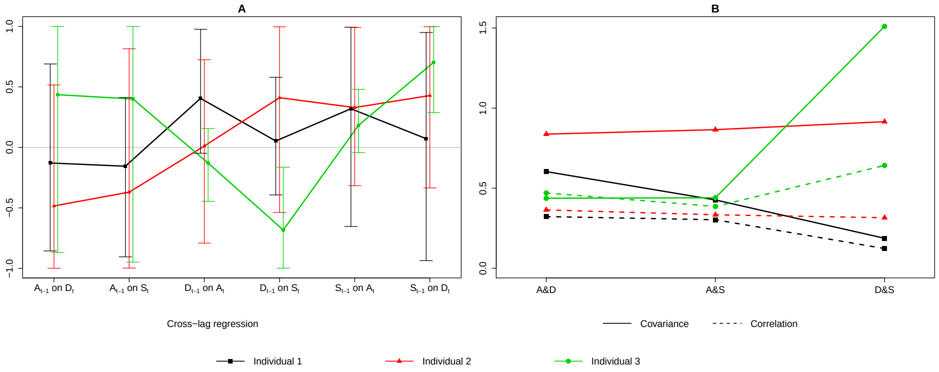


Figure 5.5: Individual scores for (A) cross-lag regressions with the estimated 95% credible interval and (B) granularity in covariance (continuous line) and correlation (dashed lines).

& Nezlek, 2005; Brose et al., 2015; Carstensen et al., 2000)

The relation between angry and stressed ($A\&S$) and angry and depressed ($A\&D$) is similar for all three individuals, apart from scale differences acquired through the variance. While Individual 2 shows similar correlations in size between all three emotions, Individuals 1 and 3 show clear differences in the correlations for $A\&D$ and $A\&S$ on the one hand, and $D\&S$ on the other hand. As such, the findings show individual differences in the level of emotion differentiation or granularity, thought to be indicative of differences in emotion regulation and functioning (e.g., Barrett et al., 2001; Erbas et al., 2014).

5.4.6 Cross-lag regression and variability

The relation between the autoregression and variance has been discussed and quantified before (Box & Jenkins, 1976). The autoregressions and cross-lag regressions increase the overall variance, $\Sigma_{ii,n}$, but not the variance of the innovation, $Q_{ii,n}$. As a result, emotions with large autoregressions and cross-lag regressions show a larger difference between the innovation variability and the within person variability. A clear example of this is visible for the emotion stressed of Individual 2 and 3. Both individuals show a large difference between the within person variability and the innovation variability (see Figure 5.4), and have high autoregressions for stressed, and large cross-lag regressions for both depressed and angry at $t-1$ on stressed at t (see Table 6.2). This identifies a clear advantage of the proposed modeling approach. When calculating autoregression and variability separately, the intrinsic relation between both may obfuscate relations with third variables (for a more detailed discussion see Koval, Pe, Meers, and Kuppens (2013)). Here, the innovation variability is not confounded with the autoregression, allowing to

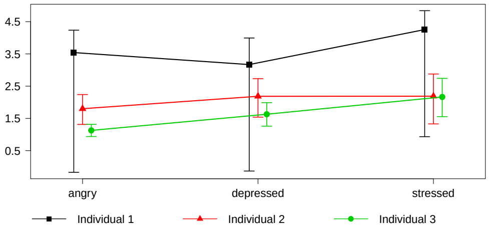


Figure 5.6: Estimated mean score per individual per emotion with bars to indicate the 95% credible interval.

study relations between these two EDFs and third variables in a more balanced way.

5.4.7 Emotional Intensity

The intensity, quantified as $\mu_{i,n}$, is shown in Figure 5.6. As can be seen, Individual 1 clearly shows a higher intensity than Individuals 2 and 3, accompanied by a higher variability. The intensity estimated for Individuals 2 and 3 is similar in pattern, with an overlap in credible intervals for depressed and virtually no difference in intensity for stressed. Where Individuals 2 and 3 show a higher intensity for emotions with a higher variability and higher autoregression, Individual 1 shows the opposite relation. As found in earlier studies, a higher average intensity over all emotions seems related to a lower average correlation over all emotion couples (Carstensen et al., 2000; Kashdan & Farmer, 2014; Erbas et al., 2014).

5.5 Discussion

In this chapter we proposed to use a BDM to analyze intensive longitudinal data to capture the patterns and regularities of an individual's expression of emotions across time. To accommodate for missing data, a link function was introduced, linking the observed to the latent variable only when the data is indeed observed. With this VAR-BDM we analyzed a data set consisting of three emotions for three individuals.

The study of emotion dynamics and the relation between EDFs is an important topic in psychological research. Most earlier studies regarding emotion dynamics have computed summary statistics for several EDFs (e.g., Carstensen et al., 2000; Barrett et al., 2001; Erbas et al., 2014; Scott et al., 2014). A complete model that encompasses all EDFs capturing the essential dynamics of multivariate, multisubject data and that can be applied to EMA data with its inevitable limitations, was lacking so far. Models that do capture multiple EDFs at once, often require rather large sample sizes of at least 100 time points without missing values (e.g., Hamaker & Grasman, 2012; Haan-Rietdijk et al., 2014). Another option, which does deal with the small sample size, is the use of a multilevel model. However, these generally require a large number of individuals measured, and are based on distributional assumptions on the individual parameters (e.g., Kuppens, Allen, & Sheeber, 2010; Lodewyckx et al., 2011; Bringmann et al., 2013). The multilevel VAR model of Schuurman, Ferrer, de Boer-Sonnenschein, and Hamaker (2016) lacked an essential element, namely the white noise. Further, they did not link the model elements to the EDFs as we did.

The results found in our empirical study largely concur with the results in previous literature. Our sample of three individuals is too small to infer the relations between EDFs among individuals, as done in earlier studies (e.g., Carstensen et al., 2000; Grühn et al., 2013; Kashdan & Farmer, 2014). One point where our results disagree with the literature, is the direction of the autoregression found, which is generally positive in other studies (Kuppens, Allen, & Sheeber, 2010; Suls et al., 1998). However, the current sample size is clearly too small to draw any conclusions without further replication on a larger scale. Clearly, this finding calls for further investigation.

With our model, we strive at aiding in the research on emotion dynamics, allowing for testing the proposed theories on empirical data. The theories on emotion dynamics are growing in number, but also in complexity (Gross, 2015; Kuppens & Verduyn, 2015). Due to the expansiveness of our model, it is possible to identify relations between emotions within individuals and between individuals, but also among EDFs, such as the relation between inertia and variability, or between EDFs and external variables, such as the relation between the average inertia and the level of clinical depression. Note that to quantify such relationships, a data set with a large number of individuals is needed. This allows for a more direct way to validating theories stating relations between EDFs.

Further advantages of the model lies in its flexibility. The model, as shown in Equations 6.4 and 6.2, can be used with non-Gaussian data through the implementation of the link function or through adjustment of the distributions used. The link function also allows for data with missing values, without the need to adapt the model. Furthermore, the model can be used to implement external variables, both as active and inactive covariates. The final advantage lies in the wide range of data for which the model is applicable: one individual for which one emotion is measured, is enough for estimation. However, more individuals and emotions can be added.

Although the expansiveness of the proposed model is an advantage, it also introduces issues with estimation, and possibly with identification. While the proposed VAR-BDM (Equations 6.4 and 6.2) can be estimated with relatively little data points, the estimated credible intervals, and thus the uncertainty, is large for data sets with short time series. The estimation of the cross-lag regression carries another issue. In empirical data it may occur that the cross-lag regressions of the different variables are substantial. This introduces multicollinearity, which complicates estimation.

Further studies should focus on the estimation properties of the model. The conditions under which the model reaches convergence while estimating, are unknown, as are the conditions needed to estimate the parameters with reasonable precision. Factors which may be of influence on the estimation and convergence of the model are the number of time points, the number of emotions, the values in $\mathbf{\Phi}_n$ and the number and pattern of missing data in the data set. These same factors may influence the precision with which the parameters may be estimated.

Chapter 6

Studying the dynamics of affect through vector-autoregressive Bayesian dynamic models

Abstract

Affect research often focuses on studying various elementary features of the individual dynamics of positive affect (PA) and negative affect (NA). The features studied are typically quantified separately, hampering the study of their mutual relations. In this chapter we consider six elementary affect features to characterize the dynamics of PA and NA: within person and innovation variability, inertia, cross-lag regression, intensity and the co-occurrence of affect. To facilitate the study of these features, we propose to use a vector autoregressive Bayesian dynamic model. This model encompasses parameters that quantify the six affect features of multiple individuals at once. The model can be applied to data typically encountered in dynamic affect research, i.e., bivariate, relatively short time series of multiple individuals, even in the presence of missing time points. Furthermore, the model allows for the inclusion of external variables. We illustrate the usefulness of the model with an empirical example using relatively short time series (53 to 71 measurements) of bivariate affect data for 12 individuals. We compare three models to find whether a weekly cycle is present in the data and whether there is autoregression present in the white noise. We compare the models with regard to the white noise, the innovation noise and a log-likelihood criterion. The model estimates provide insight into the individual dynamics of PA and NA, and their interindividual differences and similarities.

Acknowedgement: We are indepted to Kim Shifren, who allowed the use of her data set collected at Syracuse University, USA.

6.1 Introduction

Emotions play an important role in our health and well-being (Tugade et al., 2004; Grühn et al., 2013; Lewis et al., 2008). Dysregulation of emotions is considered a central feature in several mental disorders (Houben et al., 2015; Wichers et al., 2015). As there is a vast range of emotions that are in continuous interplay, the emotional system is large and complex. To study such a system in detail, an extensive data set would be needed, with regard to both the number of emotions and the number of time points observed. Such a data set is extremely hard to collect and analyze, while the need for long time series increases when more emotions are considered, as more emotions add complexity to the model, requiring a larger sample size to estimate the model with reasonable precision. One way to make this system easier to study, is by decreasing the number of elements, for instance by summarizing the vast range of emotions in two affects: positive affect (PA) and negative affect (NA).

In this article we propose a model to study the affect features of PA and NA using intensive longitudinal measurements of multiple individuals. After this, we showcase the model using an empirical data set pertaining to intensive longitudinal data for PA and NA of 12 individuals, with 50 to 71 measurements per person.

While PA indicates enthusiasm, activity and alertness, NA indicates subjective distress and aversive mood states (Watson, Clark, & Tellegen, 1988). A well-known scale to measure PA and NA is the Positive and Negative Affect Schedule (PANAS) (Watson et al., 1988). The PANAS consists of ten PA (e.g., inspired, excited) and ten NA emotions (e.g., afraid, hostile), to be rated on a Likert scale of one to five. The PANAS can be used with different temporal instructions, ranging from moment ('right now, that is, at the present moment') to general ('in general, that is, on the average'). This allows for the PA and NA to be measured as both a state and a trait. From a layman's perspective PA and NA are often seen as opposites. In contrast to this view, factor analyses of scores on the PANAS items of different individuals indicated that levels of PA and NA are uncorrelated across individuals, and thus can be seen as orthogonal dimensions (Watson et al., 1988). However, within individuals – across time – levels of PA and NA are not orthogonal, and often negatively correlated. This insight is obtained from multilevel factor analyses of data on the PANAS with a short-term temporal instruction (e.g., moment or day) collected repeatedly across a substantial number of time points among multiple individuals (Merz & Roesch, 2011; Rush & Hofer, 2014).

Both PA and NA, and the correlation between these, have been the topic of a large number of studies. PA and NA have been studied on the trait level, relating PA to extraversion (Lucas & Fujita, 2000), average PA to social events (Lucas, Le, & Dyrenforth, 2008), and PA and NA to stress and depression (Dua, 1993; Erbas et al., 2014). Lately, the PANAS is more often used to examine PA and

NA as states, resulting in studies relating the intensity of affect to stress (Scott et al., 2014), the affect dynamics to age (Carstensen et al., 2000; Brose et al., 2015) and the instability of affect with mental disorders (Jahng et al., 2008). When the PANAS is used as a measure of state instead of a measure of trait, the within-person dynamics of the affects over time are of vital importance.

To study the dynamics of affect over time, intensive longitudinal time series are used. Combining the complexity of affect data with the nature of intensive longitudinal data, especially when considering multiple aspects of affect and multiple individuals, requires an expansive theoretical and modeling framework. Because affect data is strongly related to emotion data, one may use the framework as introduced by Kuppens and Verduyn (2015) and extended by Krone, Albers, and Timmerman (2016b). Per this framework the affects, their dynamics and their mutual relationships are quantified as affect variability, inertia, cross-lag and intensity. In this chapter, we add to this framework affect co-occurrence, as an expression of emotional complexity.

Affect variability, inertia, cross-lag and intensity The variability is the extent to which the intensity of an affect changes across time. To quantify the variability, one typically uses the within-person variance. It is important to note that this within-person variance only quantifies the variability over the whole period measured. This type of variability may substantially differ from variability between two consecutive time points (Jahng et al., 2008; Trull, Lane, Koval, & Ebner-Priemer, 2015). To accommodate for possible differences in within person variability and innovation variability, we quantify the within person variability as the within person variance, and the innovation variability as the innovation noise variance (see Section 6.2 for more details).

Affective inertia refers to the tendency of an affect to retain its status quo, reflecting the resistance to change (Cook et al., 1995; Suls et al., 1998; Kuppens, Allen, & Sheeber, 2010). The inertia is quantified via the autoregression of consecutive measurements of an affect. Confusingly, this has been incidentally denoted as the autocorrelation (e.g., Kuppens, Allen, and Sheeber (2010), Kuppens and Verduyn (2015)). In general, affective inertia is assumed to stay equal over time, as we do in this chapter. This is not always the case, for example, in threshold or regime switching models, the autoregression may change as the state changes (Haan-Rietdijk et al., 2014; Hamaker & Grasman, 2012).

The regulation of emotions, and thus affect, happens partially through feedback-loops (Gross, 2015; Kuppens & Verduyn, 2015). This is quantified through the cross-lag regression (Pe & Kuppens, 2012). Analogously to the term autoregression, it is sometimes mistakenly called the cross-lag correlation (e.g., Kuppens and Verduyn (2015)). When the cross-lag regression is positive, this is called augmentation: the experience of one affect increases the strength of the other affect on a later time point. A negative cross-lag relation is called blunting: the experience of one affect decreases the strength of the other affect on a later time point.

Affect intensity deals with the average strength with which an affect is experienced over time (Carstensen et al., 2000; Barrett et al., 2001). The affect intensity may differ between individuals, and also within an individual when considering various types of affect. We quantify the intensity as the mean score over time (Kashdan & Farmer, 2014; Erbas et al., 2014).

Affect co-occurrence Affect co-occurrence (Ong & Bergeman, 2004), also known as poignancy (Carstensen et al., 2000) and covariation (Grühn et al., 2013), expresses to what extent an individual may perceive high levels of PA and NA at the same time. Affect co-occurrence has been the focus of several studies and is often seen as an indicator for emotional complexity (Carstensen et al., 2000; Grühn et al., 2013). Herewith, complexity is expressed by a low co-occurrence, implying that levels of PA and NA are unrelated. Here we will denote the opposite of co-occurrence, thus experiencing PA and NA as two opposite ends of an one-dimensional scale as negative co-occurrence.

Co-occurrence is generally quantified as the correlation between PA and NA across time points (Carstensen et al., 2000; Ong & Bergeman, 2004; Hay & Diehl, 2011). A high, positive correlation indicates a high co-occurrence, a negative correlation a negative co-occurrence, and a small correlation as low co-occurrence. Quantified as such, co-occurrence is positively related with psychological resilience and negatively with psychological stress and neuroticism (Ong & Bergeman, 2004) and high positive co-occurrence is related with faster recovery from negative affect (Hay & Diehl, 2011). Across studies, co-occurrences have been found in all three categories: most studies found negative co-occurrence (Carstensen et al., 2000; Ready, Carvalho, & Weinberger, 2008; Lucas et al., 2008; Merz & Roesch, 2011), but positive co-occurrence (Ong & Bergeman, 2004) and co-occurrence close to zero (Hay & Diehl, 2011) have also been found.

Quantifying co-occurrence via the correlation of PA and NA has the advantage of using a standardized measure to express the strength of the linear relationship between PA and NA. However, a direct comparison between individuals on the basis of correlations may be hampered by low within-person variances. Low within-person variances reduce the size of the absolute correlation (Scott et al., 2014), which renders the interpretation difficult. Earlier studies shows that using correlations as opposed to covariances to quantify co-occurrence may seriously influence the results. For example, where a study using a correlation found a positive relation between age and co-occurrence (Carstensen et al., 2000), a later study using covariance found a negative relation between age and co-occurrence (Brose et al., 2015) which may be explained by the lower variance in PA and NA found in older individuals (Röcke et al., 2009). Therefore, we quantify the affect co-occurrence in terms of both the correlation and the covariance between PA and NA.

Models Intensive longitudinal PA and NA data are generally bivariate, multi-individual time series data of limited length, meaning a maximum of 50 to 75 time points per individual. In the next section we will discuss the vector autoregressive Bayesian Dynamic Model and explain why this model is very suitable for analyzing the type of data at hand. Furthermore, we discuss how to quantify the variability, inertia, cross-lag, intensity and co-occurrence of affect. We proceed by presenting three model variants for empirical PA and NA data of multiple individuals. We close with a discussion of the model and its possibilities.

6.2 Bayesian Dynamic Linear Model

The best known model to handle multivariate intensive time series is probably the vector autoregressive (VAR) model (Tiao & Box, 1981). The VAR model includes all parameters mentioned in the previous section. While the VAR model is very useful from a theoretical perspective, the estimation of the traditional VAR model using maximum likelihood estimation or exact likelihood estimation can be cumbersome. These methods are build upon strict assumptions, such as stationarity and invertibility (Kendall & Ord, 1990). Furthermore, because the extension to more than one individual is complicated computationally, one has to resort to a separate analysis for each individual. Finally, it cannot deal with missing data, implying that one needs some sort of imputation to complete the time series before a VAR analysis on observed data including missings can be applied (Liu & Molenaar, 2014).

A more flexible approach to model multivariate time series is the state space model (SSM) (cf. Durbin & Koopman, 2012). The SSM may also be used to estimate a VAR model. The essential difference between a VAR model and the SSM is that the SSM allows for the modeling of a latent trend. Furthermore, the SSM can be used with time series for which the normality assumptions do not hold, and the inclusion of external variables in the SSM is straightforward. This implies that the SSM gives a large range of opportunities. However, not all possibilities are implemented yet. For example, to the best of our knowledge, a maximum likelihood SSM for time series suffering from missing values and not meeting the normality assumptions, is lacking so far.

In Krone et al. (2016b), we have created such a model in a Bayesian context, using the vector autoregressive Bayesian Dynamic Model (VAR-BDM) (West & Harrison, 1997; Krone et al., 2016b). This model combines the VAR model with a SSM framework and Bayesian Markov-Chain Monte Carlo (MCMC) estimation. This VAR-BDM is capable of handling the typical complexities that occur in affect data. Furthermore, the Bayesian framework allows for the inclusion of prior expectations in the model in a natural way. In this chapter, we discuss the VAR-BDM for bivariate affect data and, as such, present a simplified version of the model presented in Krone et al. (2016b). This simplified model is suitable for analyzing the type of bivariate affect data that is common in emotion research.

Notation We denote the observed score $Y_{i,t,n}$ of individual n ($n = 1, 2, .., N$) at time t ($t = 1, 2, ..., T_n$) for affect i ($i = 1, 2$), where $i = 1$ correspond to PA and $i = 2$ to NA. We use bold font as notation for vectors and matrices, e.g. $\boldsymbol{Y}_{t,n} = [Y_{1,t,n} \quad Y_{2,t,n}]$ is the vector containing both the PA and NA scores for individual n at time point t. We denote the variance-covariance matrix of the time series for individual n by $\boldsymbol{\Sigma}_n$ (2×2).

Link function, observation and system equations A complete VAR-BDM contains three elements: the link function, the observation equation and the system equation. The link function is required when it is assumed that the residuals are not normally distributed, in which case it works similar to the link function in generalized linear models (Nelder & Wedderburn, 1972; McCullagh & Nelder, 1989). Furthermore, the link function can be used to handle missing data. The data set we use in this chapter does not contain missing values and we do rely on the standard assumption of normally distributed residuals. Therefore we will forgo the link function (i.e. apply the identity link) in this chapter and focus on the observation equation and the system equation. A BDM with an identity link function is also referred to as a Dynamic Linear Model (DLM; West & Harrison, 1997).

The observation equation connects the observed score vector $\boldsymbol{Y}_{t,n}$ of individual n at time t to the latent state vector $\boldsymbol{\theta}_{t,n}$:

$$\boldsymbol{Y}_{t,n} = \quad \boldsymbol{\mu}_n + \boldsymbol{\theta}_{t,n} + \boldsymbol{\varepsilon}_{t,n}, \quad \boldsymbol{\varepsilon}_{t,n} \sim N(\boldsymbol{0}, \boldsymbol{H}_n), \tag{6.1}$$

where $\boldsymbol{\mu}_n$ (2×1) denotes the mean vector of both affects, $\boldsymbol{\theta}_{t,n}$ (2×1) the latent variable vector, $\boldsymbol{\varepsilon}_{t,n}$ (2×1) the white noise vector and $\boldsymbol{H}_n$ (2×2) the covariance matrix of $\boldsymbol{\varepsilon}_{t,n}$.

The system equation models the autoregression and cross-lag regressions and the innovation over time of the latent variable vector $\boldsymbol{\theta}_{t,n}$:

$$\boldsymbol{\theta}_{t,n} = \quad \boldsymbol{\theta}_{t-1,n} \times \boldsymbol{\Phi}_n + \boldsymbol{\eta}_{t,n}, \quad \boldsymbol{\eta}_{t,n} \sim N(\boldsymbol{0}, \boldsymbol{Q}_n), \tag{6.2}$$

where $\boldsymbol{\Phi}_n$ (2×2) is the auto/cross-lag regression matrix, $\boldsymbol{\eta}_{t,n}$ (2×1) is the innovation vector and $\boldsymbol{Q}_n$ (2×2) the covariance matrix of the innovation. A graphical representation of the model expressed in Equations 6.1 and 6.2 for a single individual n can be seen in Figure 6.1.

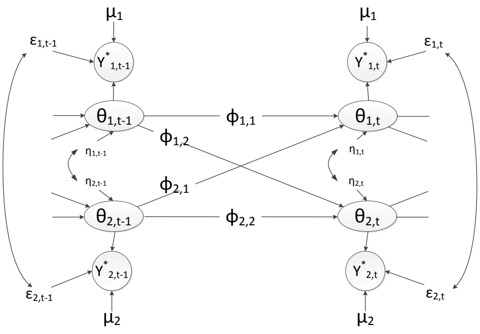


Figure 6.1: Schematic representation of the model as expressed in Equations 6.1 and 6.2 for a single individual n (subscripts left out for clarity reasons) and $i = 1, 2$ affects.

Concept	Quantification	Parameter
innovation variability	Innovation of $\boldsymbol{Y}_{i,n}$	$Q_{ii,n}$
within person variability	Variance of $\boldsymbol{Y}_{i,n}$	$\Sigma_{ii,n}$
Inertia	Autoregression	$\boldsymbol{\Phi}_{ii,n}$
Affective cross-lag	Cross-lag regression	$\Phi_{ij,n}$
Intensity	Mean estimated score	$\mu_{i,n}$
Co-occurrence	Covariance of $\boldsymbol{Y}_{ij,n}$	$\Sigma_{ij,n}$
	Correlation of $\boldsymbol{Y}_{ij,n}$	Cor($\boldsymbol{Y}_{ij,n}$)

Table 6.1: Quantification of emotion dynamics features for affect i, in relation to affect j where applicable, in the notation of Equations 6.1 and 6.2.

For large data sets, one may decide to let $\boldsymbol{H}_n$, $\boldsymbol{Q}_n$ and/or $\boldsymbol{\Phi}_n$ be time-dependent according to some pre-defined structure. However, with time series of lengths typical in affect research, the amount of data is insufficient for such complex data structures, which is why we keep these matrices constant over time.

Having defined our BDM-VAR, we can link theory and application for the affect features. In Table 6.1 we show the quantification of the different features, using the parameters of the VAR-BDM. The innovation variability is expressed via $Q_{ii,n}$, and the within person variability for individual n and affect i is expressed via $\Sigma_{ii,n}$. For individual n, the autoregression of affect i is $\Phi_{ii,n}$, and the cross-lag regression is $\Phi_{ij,n}$ for the cross-lagged effect of affect i on affect j. The intensity for individual n on affect i is the mean $\mu_{i,n}$. The co-occurrence is expressed via the covariance and the correlation of $Y_{i,n}$ and $Y_{j,n}$. In the next paragraphs, we will discuss the choice of priors and the identification in and application to empirical data.

Priors As any Bayesian model, the VAR-BDM requires priors set by the researcher. If one has strong expectations about the results to be expected, e.g. based on previous results, one can accomodate these expectations through the eliciting of the priors. If there is little information on what is to be expected, a weak informative or non-informative prior may be used (Gelman, Carlin, et al., 2013). An overview of methods for eliciting priors is provided by Garthwaite, Kadane, and O'Hagan (2005), the priors we choose in this chapter are motivated in Section 6.3.2.

Model convergence The VAR-BDM presented here is a fairly complex model. This may lead to estimation problems due to too little data in comparison to the complexity of the model, or to identification issues, where more than one solution fits the model (almost) equally well. Both of these will result in non-convergence, which implies that the estimation procedure has failed to find a single, optimal solution.

To check the convergence of a model, we can use the potential scale reduction factor, $\hat{R}$, and visual inspection of a trace plot. The $\hat{R}$ shows the ratio of the maximum expected change in a parameter when the number of iterations is doubled; an (ideal) value of 1 indicates that no change is expected (Gelman & Rubin, 1992; Stan Development Team, 2016). A trace plot shows the Bayesian Monte Carlo Markov Chain (MCMC) estimates for each parameter at each iteration. At convergence, the estimates over iterations are highly similar across chains, except at the fringe.

Model selection To compare model variants, it is advisable to use a combination of two strategies. First, the model variants can be compared using the estimated parameters or the noise, to see which parameters show the most preferable properties. For example, the white noise and innovation variance are preferred to be small, indicating a proper model fit. This may also be assessed by comparing the error between the observed and estimated scores, also known as white noise, between the different models. On the other hand, model parameters which are close to zero may be superfluous, thereby unnecessary complicating the model.

Second, global fit criteria may be used, which generally combine the log-likelihood and number of parameters of model for the used data set. The Watanabe-Akaike Information Criterion (WAIC) closely follows the Bayesian methods, as it takes into account the whole posterior distribution as opposed to just the point estimates used in the Aikake information criterion and the Bayesian information criterion (Gelman, Hwang, & Vehtari, 2013). This makes the WAIC the most suitable information criterion for comparing VAR-BDM variants. As for any information criterion, smaller values of the WAIC indicate a better fit. Two models can be compared by calculating the difference in WAIC and the corresponding standard error (Gelman, Hwang, & Vehtari, 2013).

6.2.1 Model estimation

The models are estimated using Bayesian MCMC estimation. For the MCMC estimates we use Hamiltonian Monte Carlo (HMC), a generalization of the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) that allows for an efficient estimation of the parameters (Gelman, Carlin, et al., 2013). This is incorporated in the software RStan (Stan Development Team, 2015; R Core Team, 2015). For the calculation of the WAIC we use the package 'loo' for the program R (Vehtari et al., 2015).

6.3 Empirical study

To illustrate the bivariate VAR-BDM, we reanalyze data from twelve individuals who completed the Positive and Negative Affect Schedule (PANAS; Watson et al., 1988) questionnaire on a daily basis for 53 to 70 consecutive days (Shifren, Hooker, Wood, & Nesselroade, 1997). These individuals were all diagnosed with Parkinson's disease and between 59 and 81 years of age (Mean $= 68.75, SD = 7.24$) at the time. The PANAS consists of 20 items, which make up two scales: positive affect (PA) and negative affect (NA). Each item is rated on a Likert scale from 1 to 5, giving a range of 10 to 50 for both PA and NA. We use three variants of the bivariate VAR-BDM to analyze the PA and NA, after which we compare the models and interpret the results.

6.3.1 Models

The three models used in this chapter are defined through the observation and system equation of the VAR-BDM, as discussed in Section 6.2.

Model 1 Model 1 is the vector-autoregressive lag 1 model as expressed in Equations 6.1 and 6.2. Inclusion of the white noise next to the innovation error improves the estimation of the autoregression (Schuurman et al., 2015). However, it may also introduce estimation issues due to the extra complexity introduced in the model. To allow for the inclusion of white noise without raising estimation issues, we set $\boldsymbol{H}_n$, the variance-covariance matrix of the white noise, equal to the identity matrix for all measured time points.

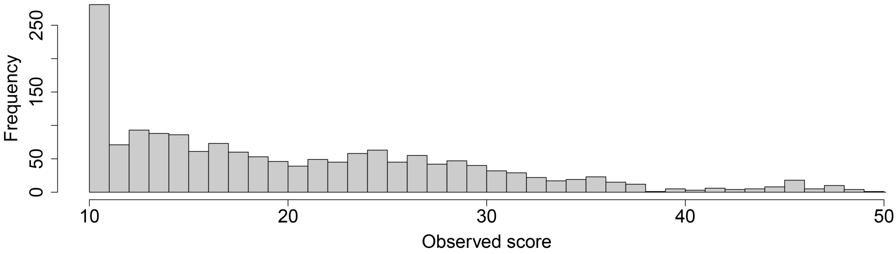


Figure 6.2: The observed scores on PA and NA for all individuals

Model 2 In Model 2 we add a weekly cycle to see whether we find an effect of day of the week in our sample. Earlier studies found evidence for a weekly cycle in mood data (e.g., Larsen & Kasimatis, 1990; Egloff, Tausch, Kohlmann, & Krohne, 1995; Harvey et al., 2015). However it is questionable whether this is indeed experienced or due to bias in expectancy and remembering (Areni, 2008; Croft & Walker, 2001). For Model 2, the observation equation remains equal to the one in Model 1, but an autoregressive element with lag 7 is added to the system equation:

$$\boldsymbol{\theta}_{t,n} = \quad \boldsymbol{\theta}_{t-1,n} \times \boldsymbol{\Phi}_{1,n} + \boldsymbol{\theta}_{t-7,n} \times \boldsymbol{\Phi}_{2,n} + \boldsymbol{\eta}_{t,n}, \quad \boldsymbol{\eta}_{t,n} \sim N(\mathbf{0}, \boldsymbol{Q}_n), \tag{6.3}$$

where $\boldsymbol{\Phi}_{1,n}$ (2×2) is the auto/cross-lag regression matrix for lag 1 and $\boldsymbol{\Phi}_{2,n}$ (2×2) is the auto/cross-lag regression matrix for lag 7.

Model 3 For Model 3 we add a moving average parameter to the white noise. Earlier studies showed that the noise in time series data is often not independent between moments (Goldstein et al., 1994; Goldstein, 2011). Therefore, adding a moving average parameter, i.e., an autoregression of the noise, may improve the model fit. The system equation is equal to the one of Model 1. The moving average element is added to the observation equation:

$$\boldsymbol{y}_{t,n} = \quad \boldsymbol{\mu}_n + \boldsymbol{\theta}_{t,n} + \boldsymbol{\varepsilon}_{t-1,n} \boldsymbol{\Psi}_n + \boldsymbol{\varepsilon}_{t,n}, \quad \boldsymbol{\varepsilon}_{t,n} \sim N(\mathbf{0}, \boldsymbol{I}), \tag{6.4}$$

where $\boldsymbol{\Psi}_n (2 \times 2)$ is a diagonal matrix holding the moving average parameters.

6.3.2 Priors

We have to set the priors for $\boldsymbol{\mu}_n$, $\boldsymbol{Q}_n$, and for the elements of $\boldsymbol{\Phi}_{\cdot,n}$ and $\boldsymbol{\Psi}_n$. For the parameter $\boldsymbol{\mu}_n$, we set a normal prior with mean 25 and standard deviation 15, allowing for posterior means in the whole range of the possible values of PA and NA

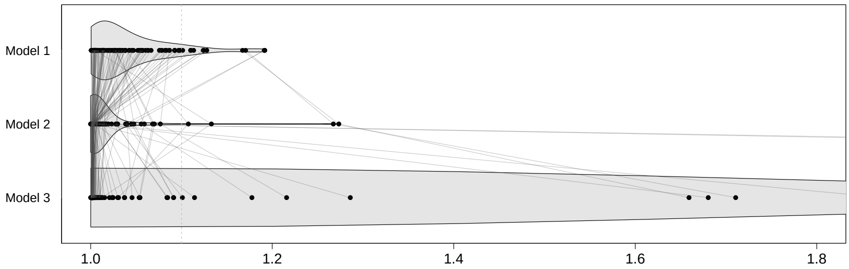


Figure 6.3: $\hat{R}$ scores for all parameters per model.

while still allowing for an emphasis on the lower scores, as found in a histogram of all observed values in $Y_{\cdot,\cdot,\cdot}$ (see Figure 6.2). The prior of the autoregression parameters $\boldsymbol{\Phi}_{\cdot,n}$ and $\boldsymbol{\Psi}_n$ is set to a symmetrized reference prior (Berger & Yang, 1994), which needs no hyperparameters. For the covariance matrix we used a LKJ-correlation prior for the correlation matrix, combined with a Cauchy(0,2.5) prior for the scale parameter.

6.4 Results

As a visual aid when discussing our results, we use violin plots, which display the individual values through dots and a corresponding smoothed density through the colored area. Where possible, the variables are connected by lines for each individual, showing the relations of the estimations between different models or between different variables.

To express the relations between parameters across individuals, we use Spearman's rank correlation coefficient. The computed correlations will be accompanied by Bayes Factors (BFs) as calculated with the method discussed by Wetzels and Wagenmakers (2012). Throughout, we will report only the BF_{01} values, simply denoted by BF. The BF_{10} values can be found by taking the reciprocal. For each of these correlations, a BF below 1 is relative evidence for the null hypothesis of zero correlation versus the two-sided alternative. When a BF is between 0.33 and 3.00, the evidence for either H_0 or H_1 is considered merely anecdotal.

6.4.1 Convergence

The model convergence is assessed by the $\hat{R}$ and the trace plot of the parameters. As can be seen in Figure 6.3, the convergence differs strongly between the three models. Where Models 1 and 2 show a maximum $\hat{R}$ of 1.19 and 1.27 respectively, Model 3 shows a maximum $\hat{R}$ of 21. It is remarkable that the nine highest $\hat{R}$ values found for Model 3 are from parameters of Individual 5, as are the two highest $\hat{R}$

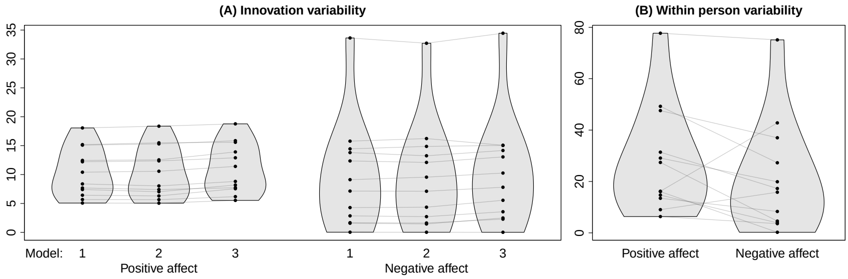


Figure 6.4: Individual scores for (A) innovation variability per model and (B) within person variability for PA and NA.

	Model 1	Model 2	Model 3
Innovation variability	0.51(0.91)	0.51(0.91)	0.53(1.05)
Inertia	-0.25(0.29)	-0.09(0.22)	-0.16(0.24)
Cross-lag	0.52(0.96)	0.48(0.77)	0.67(3.72)
Intensity	-0.20(0.26)	-0.54(1.11)	-0.26(0.30)

Table 6.2: Spearman's correlation between estimated parameters for positive and negative affect with the Bayes factor in parentheses

values for Model 2. The highest $\hat{R}$ for any other individual in Model 3 is 1.22. This suggest that for Individual 5 Model 3 is not identified, and Model 2 is only just identified. The trace plots confirm these results. As for all models, at least 90% of the $\hat{R}$ is below 1.10, and since we are working with complicated data and models we consider the convergence of the models acceptable.

6.4.2 Affect measures

Affect variability The innovation variability, as denoted by the innovation variance $Q_{ii,n}$, is shown in Figure 6.4, panel (A). When following the lines between the individuals $Q_{ii,n}$ per model, it is clear that the different models estimate similar values of $Q_{ii,n}$. Further, PA shows a much smaller range of innovation variance than NA does. In Table 6.2, the innovation variability correlation (i.e., the correlation between the innovation variances of PA and NA) is presented for the three models. As can be seen, the innovation variability correlation shows a trend towards a positive correlation, although the BF indicates anecdotal for evidence for both H_0 (Model 1 and 2) and H_1 (Model 3) .

The within person variability is the variance of the time series, and hence does not depend on the specific model considered. As can be seen in panel (B) of Figure 6.4, the variances of PA and NA are about similarly distributed. The correlation between the variances of the two affects across individuals is 0.72 (BF = 6.93),

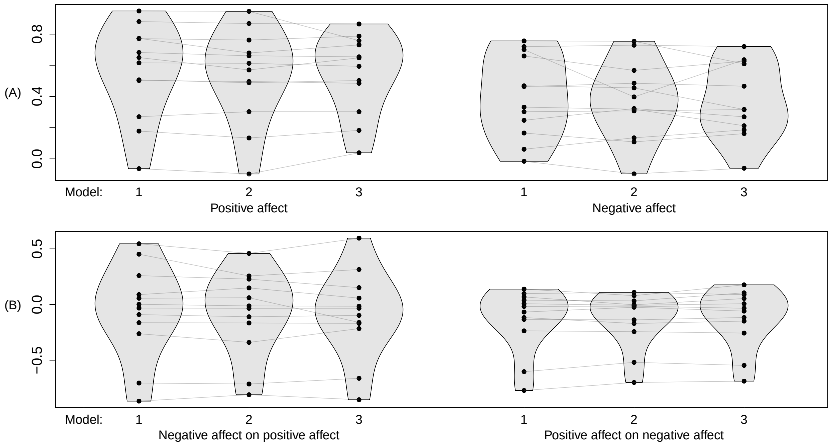


Figure 6.5: Individual scores for affect inertia and affective cross-lag for PA and NA.

giving substantial evidence for a positive correlation. It is noteworthy that the outlier towards higher values for the variances is the same individual as the outlier towards higher values for innovation variances for PA and NA.

Affect inertia The inertia, or the tendency to retain status quo, is expressed through the autoregression and displayed in panel (A) of Figure 6.5. In general, the autoregression is slightly higher for PA than for NA in all three models. The Figure clearly demonstrates large differences between individuals: nearly the whole positive range $[0, 1]$ for these correlations is covered. Further, the individual autoregressions for PA are focused in the upper range, while the individual autoregressions of NA are more evenly distributed. In Table 6.3 it can be seen that for most individuals the 95% credible interval (CrI) of the autoregression does not include zero, giving evidence that an effect is indeed present for most individuals. In Table 6.2, it can be seen that the correlation between the autoregressions of PA and NA across individuals show a slight trend towards a negative correlation. The associated BF for all models indicate substantial evidence for a zero correlation. For both PA and NA, several individuals show different estimated autoregressions in the different models, indicating that this parameter is influenced by the inclusion of a seasonal or moving average effect.

Affect cross-lag The cross-lag regression is shown in panel (B) of Figure 6.5, where positive values express augmentation and negative values express blunting effects over time. The cross-lag effect for NA on PA differs greatly across individuals, with values between -0.87 and 0.59 across the different models. As can be seen, the cross-lag regression for the lagged effect of PA on NA is mostly centered

	Negative affect	Positive affect	Negative affect on positive affect	Positive affect on negative affect
		Elements of $\mathbf{\Phi}_{1,n}$		
Model 1	9	10	3	3
Model 2	8	10	3	2
Model 3	8	9	3	3
		Elements of $\mathbf{\Phi}_{2,n}$		
Model 2	1	0	0	0
		Elements of $\mathbf{\Psi}_n$		
Model 3	0	0		

Table 6.3: Number of individuals for whom the 95% CrI of the estimated parameter excludes zero.

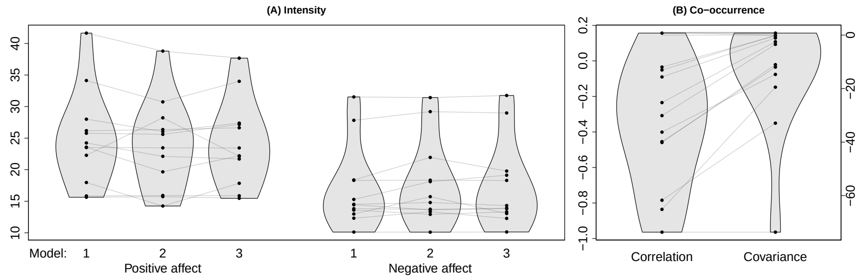


Figure 6.6: Individual scores for intensity for PA and NA.

around zero, with two negative outliers. For both cross-lag effects, the 95% CrI includes zero for at least nine out of twelve individuals, as shown in Table 6.3. As can be seen in Table 6.2, the correlations between the two cross-lag effects over individuals are positive, with for Model 1 and 2 anecdotal evidence in favor of a zero correlation between the two cross-lag effects, and for Model 3 substantial evidence in favor of a positive correlation. It is noteworthy that the individual with the highest augmenting cross-lag regression of NA on PA shows also the highest augmenting cross-lag regression for PA on NA.

Affect Intensity The intensity, expressed as the posterior estimated mean, is shown in Figure 6.6. In general, a higher mean is found for PA than for NA. The posterior estimated mean shows some differences between the models, where Model 1 and 3 are more in concurrence with each other than either is with Model 2. As can be seen in Table 6.2, the correlation between the means of PA and NA is slightly negative, with substantial evidence in favor of a zero correlation for Model 1 and 3, and anecdotal evidence for a negative correlation in Model 2.

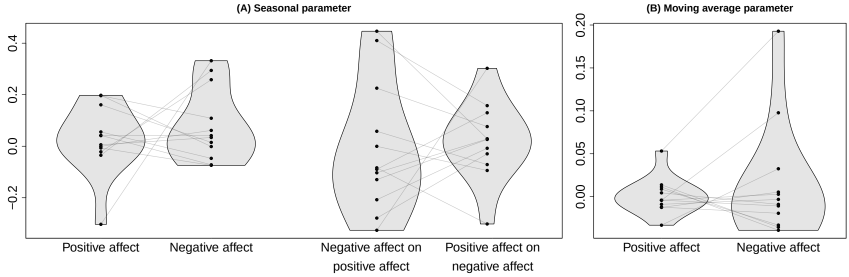


Figure 6.7: Individual scores for seasonality and moving average for PA and NA.

Affect co-occurrence The co-occurrence is expressed as the correlation or covariance between the time series. These measures are independent of the model considered, as with the within person variability. As can be seen in panel (B) of Figure 6.6 with the left axis for correlation and the right for covariation, the co-occurrence is negative for almost all individuals. The correlation between two the co-occurrence measures is rather high as expected, at 0.95 with a BF of 22913.63, giving decisive evidence for the presence of a non-zero correlation.

Seasonality In Model 2, we included a seasonality parameter $\mathbf{\Phi}_{2,n}$, indicating a weekly cycle. In this parameter we allowed both an autoregression, indicating inertia, and a cross-lag regression, indicating augmenting and blunting. As can be seen in panel (A) of Figure **??**, the autoregression is around zero for most individuals. The correlation between the individuals' autoregression for NA and PA is -0.52 (BF $= 0.96$). The cross-lag regression is more spread out, especially for the lagged effect of NA on PA. For almost all individuals, the estimated 95% CrI of all elements of $\mathbf{\Phi}_{2,n}$ included zero, as shown in Table 6.3. The correlation between the cross-lag of NA on PA and of PA on NA for lag 7 is 0.03 (BF $= 0.21$), giving substantial evidence in favor of H_0.

Moving average In Model 3 we included a moving average parameter, to account for a possible correlation in white noise. As can be seen in panel (B) of Figure **??**, the posterior mean for this moving average parameter is generally small, with only three values larger than $|0.05|$. For all twelve individuals, the 95% CrI of the moving average parameter for both PA and NA includes zero, as can be seen in Table 6.3, indicating that it is questionable whether this effect indeed differs from zero for most individuals. The correlation between the individuals' moving average parameter for NA and PA is 0.49 (BF $= 0.80$).

6.4.3 Model comparison

We used three different models for the analysis, which we compare using the absolute error, innovation noise and the WAIC.

Absolute error As we set the white noise variance to one for each model, we cannot compare these across models. As an alternative, we use the absolute error. The absolute error is calculated as the absolute value of the $\boldsymbol{\varepsilon}_{t,n}$ as defined in Equation 6.1. Table 6.4 lists the mean absolute errors (MAE) for the three models. Whereas Model 1 and Model 2 have quite similar MAE, Model 3 shows a lower MAE, indicating a better model fit. A paired t-test between Model 1 and Model 2 on the absolute error, gives a BF of 0.16, substantial evidence in favor of H_0, indicating no difference between the two variables. However, when performing a paired t-test on the absolute error between Model 1 and Model 3, a BF of $62,955.55$ is given, giving decisive evidence for a difference between the two variables, for Model 2 and Model 3 a BF of 101.06 is obtained, also indicative of decisive evidence in favor of a difference.

Innovation variance Where the absolute error indicates the noise in the observation equation, the innovation variance is indicative of the noise in the system equation. When comparing the innovation noise of the three models, as done in Section 6.4.2, we see small but consistent differences between the models. In general the innovation variance for Model 3 is slightly higher than for Model 1 and 2, both for PA and for NA. This may be an indication that Model 3 shows a better estimation of the parameters in the observation equation, being $\boldsymbol{\mu}_n$ and $\boldsymbol{\Theta}_n$, but not necessarily a better estimation of the parameters in the system equation, being $\boldsymbol{\Phi}_n$.

WAIC For each of the models, the WAIC is reported in Table 6.4. The WAIC is based on the loglikelihood of the full model, which is the sum of the loglikelihood for the observation and system equation in the VAR-BDM. Models 1 and 2 are comparable with regard to fit, with a difference in WAIC of 28.5 (standard error(SE) = 21.4). Model 3 shows the best fit of the three models, with a difference in WAIC with Model 1 of 1016.2 (SE = 144.0) and Model 2 of 1045.4 (SE = 144.0). As the model comparison of both the mean absolute error and the WAIC point to Model 3 as the best fitting model, we use this model to assess the relations between the parameters.

6.4.4 Relations between parameters

Using the parameter estimates of Model 3, we can consider the relations between the estimated parameters. As can be seen in Table 6.5, four out of the 12 corre-

Model	MAE (SE)	WAIC (SE)
Model 1	0.44 (0.03)	13,179.6 (121.6)
Model 2	0.41 (0.03)	13,208.6 (122.8)
Model 3	0.32 (0.04)	12,163.3 (194.7)

Table 6.4: Mean absolute error (MAE) and WAIC per model.

Negative affect Positive affect	Innovation	Within person	Inertia	Intensity
Innovation variability		0.91 (766.15)	0.15 (0.24)	0.68 (4.04)
Within person variability	0.85 (97.53)		0.32 (0.36)	0.74 (9.43)
Inertia	0.17 (0.25)	0.35 (0.40)		0.43 (0.58)
Intensity	-0.12 (0.23)	0.12 (0.23)	0.39 (0.48)	

Table 6.5: Spearman's correlation between parameters for negative affect (upper triangle) and positive affect (upper triangle), with the accompanying BF in parentheses.

lations for both PA and NA between the measures show only anecdotal evidence towards either H_0 or H_1. Four show substantial evidence towards H_0, three for PA and one for NA. That is, for PA, the correlations between innovation variability on the one hand and inertia and intensity on the other hand, and the correlation between within person variability and the intensity, show substantial evidence towards H_0. For NA, the correlation between inertia and innovation variability shows substantial evidence for H_0.

Substantial evidence in favor of H_1 is found for the correlations between the intensity for NA on the one hand, and the innovation and within person variability for NA on the other. The correlation between intensity and variability for NA may be related to the fact that the intensity is generally lower for NA, generating a floor effect which keeps the variability low for those who score low on NA. For PA, very strong evidence is found for a correlation between short and within person variability. For NA, decisive evidence is found for a correlation between innovation and within person variability. The correlations between within person and innovation variability for both PA and NA confirm that these are two related aspects.

A correlation between co-occurrence and cross-lag regression may be present, as both deal with the relation between the two affects. Since the covariance is strongly influenced by the variance, we use the correlation as measure for co-occurrence here. The correlation between the contemporary correlation of PA and NA, and the cross-lag effect of NA on PA is 0.67 (BF = 3.72), for the cross-lag effect of PA on NA the correlation is 0.85 (BF = 80.10). This implies that the contemporary and lagged relations between PA and NA may not be independent of each other.

6.4.5 Conclusion

The results indicate that the best fitting model is Model 3, which includes the moving average parameter. This is in line with the assumption that the white noise of a time series is often autocorrelated. However, Model 3 may not be identified for all time series, as shown for Individual 5. We found that the difference in variability between PA and NA is small, with a lower variability for NA than for PA. The inertia is, on average, slightly higher for PA than for NA and the cross-lag effect of PA on NA is smaller than the effect of NA on PA. The intensity is higher for PA than for NA and the co-occurrence is negative for almost all individuals. The seasonality shows small auto- and cross-lag regressions in most cases, and does not improve model fit. Adding a moving average parameter improves the model strongly, even though only some individuals show a moving average parameter larger than 0.05.

6.5 Discussion

In this chapter we discuss the application of the vector autoregressive Bayesian dynamic model (VAR-BDM) to positive affect (PA) and negative affect (NA) data. Using this VAR-BDM on the bivariate affect data, we can estimate innovation variability, within person variability, inertia, cross-lag, intensity and co-occurrence of PA and NA. In the empirical application, we compared three model variants for an empirical intensive longitudinal data set of twelve subjects on positive and negative affect. Model 1 was an AR(1) model, for Model 2 we added a weekly cycle to the AR(1) model, and for Model 3 we included a moving average parameter for the white noise.

The PANAS has been studied thoroughly (e.g., Hay & Diehl, 2011; Ong & Bergeman, 2004; Merz & Roesch, 2011), but not yet with a model as complete as done in this chapter. The multisubject, bivariate analysis encompassing several features of the PANAS allows for a complete study of the dynamics of PA and NA. In this way, we contribute to the analysis of emotional complexity, of which co-occurrence of PA and NA is often deemed to be an important part (e.g., Barrett et al., 2001; Grühn et al., 2013; Scott et al., 2014).

We compared three models differing in complexity. When comparing the models, the effect of the weekly cycle as used in Model 2, does not improve the model fit. This is in concurrence with the theory that there is no observable weekday effect (Areni, 2008; Croft & Walker, 2001), even though this may be perceived in hindsight, or anticipated. The moving average parameter added to the AR(1) model in Model 3, does significantly improve model fit. This is in concurrence with the theory that the white noise in intensive longitudinal data is not independent over time, but is often autocorrelated (Goldstein et al., 1994).

As earlier studies focus on the relation of PA and NA with other factors, such as age (e.g., Carstensen et al., 2000; Hay & Diehl, 2011; Ong & Bergeman, 2004), there is very little material to compare our results to. One aspect that has been studied, is the co-occurrence. However, the results with regard to this differ strongly across studies. In earlier studies, negative correlations, no correlation or even a positive correlation was found for the combination of PA and NA (Carstensen et al., 2000; Hay & Diehl, 2011; Ong & Bergeman, 2004). We found negative co-occurrence for all but one individual.

With our model, we strive to add to the research on affect dynamics, and to aid further research. The model allows for the analysis of large data sets containing bivariate PA and NA measurements, but also for the analysis of small data sets. When the PANAS is used, results from different studies using this model are comparable for all features mentioned, i.e., the variability, inertia, cross-lag, intensity and co-occurrence. Further more, the flexibility of the model allows for inclusion of external variables on any of the used parameters (e.g., Krone et al., 2016c).

In this chapter we aim at a complete study of the structure of the PANAS, using a flexible and complete model. One issue that arises, is that our chosen model is not identifiable for all possible data sets, as shown by the convergence issues pertaining Individual 5. This may be an indication that it is important to find a fitting model for each individual, or accept a less well fitting model that is identifiable for all. The latter would be, in this case, the VAR-BDM with only an AR(1) parameter we used as Model 1. A future study may focus on the difference in model fit between different individuals. This may be linked to external variables, such as age, gender or stress levels.

The VAR-BDM we propose is time independent, meaning that all variables are assumed to be equal over time. As former studies established that the dynamics may be dependent on stress (Scott et al., 2014), age (Carstensen et al., 2000) and other external factors, it may be beneficial to allow the model to change over time. This can be done by making parameters dependent on external variables, or by applying treshold or regime switching models (Haan-Rietdijk et al., 2014; Hamaker & Grasman, 2012). These models allow for different latent state, each with their own parameters. However, it should be taken into account that for each latent state a number of observations is needed to estimate the added parameters.

Chapter 7

Conclusion

The aim of this thesis was to study time series analysis in psychology. To narrow the focus in such an extensive field, the focus was lain on AR(1) models and the challenges of empirical data analysis. In this conclusive chapter, I will discuss the issues raised in the first chapter. After this, I will close this thesis with some recommendations for further studies.

7.1 Estimation of the autoregressive model

7.1.1 The difference between the estimators

The first question posed was: which estimator is preferred for AR(1) data? To answer this question, I compared several estimators for univariate, single individual data using a simulation study. I analyzed the data using six estimators: the r_1 estimator (Yule, 1927; Walker, 1931; Box & Jenkins, 1976), C-statistic (Young, 1941), ordinary least squares estimator (OLS), maximum likelihood estimator (MLE) and Bayesian Markov Chain Monte Carlo (MCMC) estimator using either a flat prior (B_f), or a symmetrized reference prior (B_{sr}).

I compared the estimators under several conditions, varying the length of the time series between 10 and 100 time points, and varying the true autocorrelation between -0.90 and 0.90. This showed that the distinction is not between Bayesian and frequentistic methods, but between iterative and closed form methods. The iterative methods, i.e., MLE, B_f and B_{sr}, showed better results with regard to the bias of the estimated autocorrelation. The closed form estimators showed a smaller variability for the estimators. However, the variability of the iterative estimators decreased strongly as the number of time points increased, decreasing the difference between iterative and closed form estimators for longer time series. Comparing the two Bayesian estimators showed that B_{sr} is to be preferred over

B_f. This led to the conclusion that the MLE and B_{sr} merit further study.

In two subsequent studies I compared the MLE and the B_{sr}. The first study was aimed at the robustness of the estimators, examining the effect of underspecification by using an AR(1) to analyze ARMA(1,1) data. Again, I varied several parameters, being the number of time points (25 or 50) and the size of both the AR(1) and MA(1) parameters (between -0.90 and 0.90). The differences between the two estimators were small and inconsistent over the conditions. In general, the B_{sr} showed a slightly larger variability and bias then the MLE. However, the difference is too small to draw strong conclusions as to the difference in robustness of the estimators.

The second study compared the MLE and B_{sr} for multiple individuals using a simulation study. Here, I compared a total of four estimators, analyzing the data with both estimation methods under a random coefficients model and a fixed effects model. As with single individual data, the multiple individual data was simulated under several conditions, varying the number of individuals (10 or 25), the length of the time series (also 10 or 25), and the mean (-0.60 to 0.60) and standard deviation (0.25 or 0.40) of the true autocorrelation distribution. The differences between the estimation methods were small to negligible. However, the differences between the random coefficients model and the fixed model were substantial. The random coefficients model shows better results for five out of the six measurements, making it the preferred model over the fixed model where possible, be it estimated with MLE or B_{sr}.

In conclusion, both the MLE and the B_{sr} show promising results in estimating the autocorrelation of AR(1) data. The difference between the two is small, where the B_{sr} shows a smaller bias and the MLE a smaller variability. When analyzing multiple individuals, it is advised to use the random coefficients model instead of the fixed coefficients model where possible.

7.1.2 The effect of data properties on the estimation of the AR(1) model

The second question I sought to answer is: what is the influence of data properties on the estimation of the AR(1) model? In the first two studies, the estimators were compared using data simulated under different conditions. This allows for a comparison of the bias, variability and power with regard to the length of the time series, the number of individuals in a dataset, and the mean and standard deviation of the true autocorrelation.

The most prominent effect is found for the length of the time series, especially for single individual data. As expected, the bias and the variability become smaller as the time series becomes longer. An often asked question is how long the time series must be to get a fairly trustworthy estimation. For an univariate, single individual dataset, a length of 50 time points is generally advised for an AR(1) model (Box & Jenkins, 1976). While the B_{sr} shows a small bias for time series as short as 25 time points for most autocorrelations, 40 and preferably 50 still is the advised number of time points. This is due to the variability, which is still rather high at 25 time points.

For the random model used on multilevel data sets, the length of the time series is about as important as the number of individuals included. As either the number of individuals or the number of time points increases, the bias and the variability decrease. The required sample size depends on the size of the standard deviation of the autocorrelation. In my studies I found that for samples with little variability in the autocorrelation, i.e., a standard deviation of 0.25 or less, the random model may produce results with an acceptable size of bias when either the number of time points or the number of individuals is at least 25, while the other one of the two is at least 10. For samples with more variability in the autocorrelation autocorrelation, i.e., a standard deviation of 0.40 or less, the random model may produce results with an acceptable size of bias when both the number of time points and the number of individuals is at least 25.

For the fixed model, the effect of the number of individuals is small, while the effect of the number of time points is similar to that for the single individual design. The results of the fixed model are comparable with those of the univariate, single individual data. Taking this into account, it is advisable to follow the guideline of 50 time points for the fixed model, regardless of the number of individuals included.

The mean and standard deviation across individuals for the true autocorrelation influence the estimation of the model, but are outside the influence of the researcher. For the single individual study we varied the true autocorrelation of the series between -0.90 and 0.90, with a fixed standard deviation of zero within the population. For the multilevel study we varied the autocorrelation between -0.60 and 0.60, with a standard deviation of either 0.25 or 0.40. Over both studies, the bias decreases and the variability increases when the autocorrelation becomes closer to zero. The power increases as the autocorrelation is further from zero. For the standard deviation of the autocorrelation, as compared in the multilevel study, the variability and the bias increase as the standard deviation becomes higher. The power decreases for a higher standard deviation of the autocorrelation.

Finally, a short robustness study was done on univariate, single individual data, to see the effect of underspecifying a model. The results show that the effect of underspecification increases with the size of the unmodeled parameter, and affects the bias, variability, rejection rate and power when compared to the results for the justly specified AR(1) model. The estimated variability of the autocorrelation in the data used for this study became smaller for a longer time series. The effect of a longer time series on the bias was small and dependent on the estimation method and the size of the autocorrelation. With regard to the data properties, detecting misspecification is more important than the length of the time series.

In conclusion, the properties of the data strongly influence the estimation of the autocorrelation. As the number of time points and individuals are the only factors which can be influenced by the researcher, we focus our answer on these. First, for univariate, single subject studies as for multilevel studies following the fixed model, at least 50 time points are advised. However, this may still provide results with a low power and high variability if the true autocorrelation is small. For univariate, multiple individual data sets which are analyzed using the random coefficients model, it is advisable to strive for 25 individuals with 25 time points each. Especially when the expected mean of the autocorrelation is close to zero, and the standard deviation of the autocorrelation in the population is expected to be high.

7.2 Empirical data analysis using the BDM

7.2.1 Practical issues in time series analysis

Studying empirical time series data poses certain challenges, such as properly including in the model trends, dynamics and external variables and dealing with missing data and non-normally distributed residuals. While the other possible models were unable to deal with some or all of these issues, the BDM can implement all elements needed to handle this in one model, as shown in Chapter 4. In the following section, each of the issues found is posed and then I show how this may be handled using the BDM.

As discussed in the introduction, the trend and the dynamics of the data must be studied in order to create a fitting model. The BDM has shown to be capable of both. The trend may be studied by inclusion of a slope, as shown in Chapter 4 pertaining to the panic attack data. The model dynamics, such as the autregression for the score and the white noise, may be implemented as a model element. The BDM also allows for a combination of the trend and the dynamics, which is shown in Chapter 4. Here we complement the linear model with an autregression element for the noise. It is important to keep in mind that adjusting the model may influence the identifiability of the model.

An important issue for empirical data is missing data, especially in time series analysis. Commonly, the amount of incomplete data substantially increases with larger numbers of scheduled time points. The BDM can handle both incidental missing data and drop outs. The incidental missing data is handled through application of the link function. However, large amounts of missing data introduce uncertainty in the parameter estimates. When a drop-out occurs, the analysis is stopped for that series.

The BDM can deal with non-normally distributed residuals. For an observed variable, the link function can be implemented in the same manner as used in generalized linear models (Nelder & Wedderburn, 1972; McCullagh & Nelder, 1989). It must be noted that using a link function may increase the amount of data needed to obtain a model that can be estimated.

Furthermore, the BDM allows for the inclusion of external variables in two ways. First, the external variable can be included as an active covariate. As an active covariate, it may influence the estimation of the model. This is done in Chapter 4, where the slope for each individual is dependent on the treatment and the presence of agoraphobia. Second, the external variable can be implemented post-hoc. Using a variable post-hoc does not influence the estimation, but allows to compare the results with regard to an external variable. Again, this was shown in Chapter 4. The external variable for both active and inactive covariates may be time dependent or independent.

Finally, the BDM allows for easy comparison of different models. Often a researcher has a general idea whether to look for a trend, or which dynamic is most likely present in the data. However, this still allows for some fine-tuning of the model, such as adding an autoregression for the white noise or including an external variable. To study the result of these modifications, one can compare different models. To compare the likelihood, the Watanabe-Aikake information criterion (WAIC) (Watanabe, 2010) may be used. The WAIC is very well adapted for comparing the likelihood of Bayesian estimated models, as it uses the whole distribution instead of only the point estimation of the parameter. Further, as we like to have as little noise as possible, the white noise and the innovation noise may be compared between models. Finally, the distribution of the model parameters may be studied to exclude superfluous variables.

In conclusion, several challenges may be found in empirical data analysis. The challenges I found were defining both trend and dynamics, the handling of missing data, the presence of non-normally distributed residuals, the inclusion of external variables and the model selection process. As shown, the BDM can handle all of these challenges. However, the presence of these challenges may result in convergence problems due to lack of data or identifiability issues.

7.2.2 Integrating psychological hypotheses into the BDM

In psychological research, models are used to quantify hypotheses. In Chapters 4, 5 and 6 I explored the possibilities of the BDM with regard to the quantification of hypotheses, and the integration between a psychological hypotheses and frameworks, and the statistical model.

In Chapter 4, I created a model which would fit the hypothesized effects in the data. Here it was hypothesized that the slope was influenced by the presence of agoraphobia and the treatment the individual received. By comparing two models, I was able to study the effect of the pre-treatment symptoms on the intercept of an individual. Using the BDM, I found that there is a difference between treatments with regard to the decrease in symptoms over time. Individuals who received only behavioral therapy showed a slower decrease in symptoms than individuals receiving medication or both medication and behavioral therapy. The presence of agoraphobia had no clear effect on the decrease in symptoms over the different treatments.

In Chapter 5, I put the emotion dynamic features framework as discussed by Kuppens and Verduyn (2015) into one model. To this end I quantified within person and innovation variability, granularity, inertia, cross-lag correlation and the intensity of the emotions in one model. To this end, I used the VAR-BDM, which can handle the multi-individual, multivariate dataset with missing data. Thus the link was lain between a psychological framework, represented by the six emotion dynamic features, and the statistical model used to analyze the data. All of the features were quantified in one model, allowing for an easy comparison of the six features between the different individuals and emotions included in the dataset.

In Chapter 6, I adapted the framework of Kuppens and Verduyn (2015) for affect data. Again, I used the VAR-BDM to analze the multi-individual, bivariate data. Apart from quantifying the six affect features defined, I aimed to see whether there was a weekday effect in the data. This weekday effect is hypothesized on the basis of earlier studies, where there is no clear consensus whether it exists or not. Furthermore, I checked the presence of a moving average effect, i.e., an autoregression in the error. While this is not firmly based in affect theory, it is known that in time series the error is often autocorrelated. The weekday effect was not found in this data, but including the moving average effect improved the model. This combined into a model which included both the psychological framework on affect, given by the interpretation of the parameter of the VAR-BDM, and the statistical theory on modeling time series data, given by the inclusion of the autoregression in the white noise.

7.3 Important points when studying time series

The aim of this thesis was to discuss the limitations and explore the possibilities of psychological time series analysis. However, through the writing of this thesis several points sprang out that beg for attention but I could not elaborate on. These points are, in general, applicable to the analysis of time series data with any model.

The advised data points needed to analyze a data set with a certain model are specific for that model. In this thesis I only structurally studied a simple AR(1) model, with univariate data. The demands for a more complex model may be very much larger. However, little is known about this, apart from the notion that an increase in parameters demands an increase in data points. As such, it is important to check for convergence when analyzing any time series data set and to take into account the uncertainty of the estimated parameters. This should also be taken into account when designing and planning an empirical study. A wise strategy may be to perform a simulation study to find the minimal sample size needed to estimate the preferred model, before deciding on the preferred sample size.

Further, the model comparison is heavily dependent on the choices made by the researcher. While there are several ways to compare models, it is impossible to say whether the optimal model has been found. When all models misspecify the data, the best fitting model will just be the least bad model out of several bad models. It is thus advisable to check for the presence of important trend and dynamic elements before modifying the lesser elements, such as the dependent on external variables.

I found that a model that fits most of the data, may not fit all of the data. For example, in Chapter 6, a general fit comparison shows that Model 3 fits the data better than Model 1 and 2. However, one individual out of the 12 individuals shows strong convergence problems for this model. For this individual, the estimation of Model 3 does not converge, indicating that the data of this individual may better fit Model 1 or 2. Taking this into account, it is important to not only compare a model with a general fit measure, but to study the fit for different individuals. A wise approach may be to compare models per individual.

Finally, the VAR-BDM may be implemented in another modeling framework, such as a network model. A network model aims to charter the relations between several observed variables. In a longitudinal setting, a VAR network model may be used (e.g., Bringmann et al., 2013; Bulteel, Tuerlinckx, Brose, & Ceulemans, 2016). In a VAR network model the autoregressive and cross-regressive relations are represented as the edges between the nodes, which represent the observed variables. As such, it offers an appealing visualization of the relations between the variables studies. A network model generally encompasses a large number of variables, which complicates estimation. As a VAR network is calculated using a VAR model on all dependent variables, one may see the possibilities of using the VAR-BDM to estimate a VAR network model.

Appendices

Appendix A

R code

```
require(rstan)
##take note: need to install Rtools first(see http://mc-stan.org/)
rstan_options(auto_write=TRUE)
options(mc.cores = 5)

#create data
data <- list()
data$N <- 1 #subjects
data$T <- 20 #maximum time points
data$I <- 2 #number of emotions
data$T_N <- array(data=20, dim=1) #time points per individual
#Missing value indicator
data$M<-array(data=c(1,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,
                   1,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0),dim=c(N,T,I))
#Observed scores
data$Y<-array(data=c(1,3,2,5,3,5,5,4,3,3,1,1,2,6,6,5,5,4,1,6,1,1,1,1,1,2,3,3,2,1,1,1,5,6,3,5,2
,4,1,5),
           dim=c(N,T,I))


#call model file from working directory
Mod <- stan_model(file="Stancode_full_model.stan")

#sample model
fit <- sampling(
Mod
, data = data #The data list
, iter = 300 #This number should be large enough to reach convergence (i.e., 20.000 or such)
, chains = 3 #number of MCMC chains
, verbose = F
, refresh = 300 #update on every 300 iterations
, seed = 2016 #seed to replicate
)
```

Appendix B

Stan code

```
data{
  int <lower=2> T;      //maximum length of series across all individuals
  int <lower=1> I;      //number of emotions
  int <lower=1> N;      //number of individuals
  matrix [T,I] Y [N];   //observed scores in N matrices of T*I
  int M[N,T,I];         //m==1 if missing, m==0 is non-missing
  int T_N[N];           //length of individual time series
}

parameters{
  matrix <lower=-1, upper=1> [I,I] phi [N]; //cross-lag matrix, values restricted to (-1,1)
  corr_matrix[I] omega [N];                 //correlation matrix Q
  vector<lower=0>[I] tau [N];               //scale matrix Q
  row_vector [I] mu [N];                    //mean vector
  matrix [T, I] Z [N];                      //latent scores y*
  matrix <upper=0> [I,I] X [N];             //constant for prior of phi
  vector<lower=0>[I] lambda [N];            //vector matrix Q
}

transformed parameters{
  matrix [I,I] H[N];                        //covariance matrix of white noise
  matrix[I,I] Q[N];                         //covariance matrix of innovation
  for (n in 1:N){
  //quad_form=diag(tau)*omega*diag(tau)', diag(tau) is a diagonal matrix of vector tau
  H[n] <- quad_form_diag(omega[n],tau[n]);
  Q[n] <- diag_matrix(lambda[n]);
  }
}

model{
  for (n in 1:N){
    for (t in 2:T_N[n]){
    //system equation: estimate latent score theta
    Z[n,t] ~ multi_normal(Z[n,t-1]*phi[n], Q[n]);
        //link to vector Y_t of individual n if Y is observed for all emotions:
        if (sum(M[n,t])==0){
        //observation equation: estimate y (identity link function omitted)
        Y[n,t] ~ multi_normal(mu[n]+Z[n,t], H[n]);
        }
    }
  }
#prior
  for (n in 1:N){
  mu[n] ~ normal(0,2);
  tau[n] ~ cauchy(0,2.5);
  omega[n] ~ lkj_corr(2);
  lambda[n]~gamma(3,3);
    for (i in 1:I){
        for (j in 1:I){
        //symmetrized reference prior for phi
        increment_log_prob(X[n,i,j] - log(1-(phi[n,i,j]*phi[n,i,j]))/2);
        }
    }
 }
}
```

Summary

In this thesis, I focus on the use of time series analysis in psychological research. I focus on two main issues. The first one pertains to the effect of different estimators and data properties on the estimation of the autocorrelation. Further, I study the challenges of empirical data and how to link the psychological theory with the statistical models. For the last part I use the Bayesian dynamic model, which is able to handle the challenges present in the empirical data sets I use.

In **Chapter 2** I discuss various estimators of the autoregressive model for univariate data. I compare their performance in estimating the autocorrelation in short time series. In Study 1, under correct model specification, I compare the frequentist r_1 estimator, C-statistic, ordinary least squares estimator (OLS) and maximum likelihood estimator (MLE), and a Bayesian method, considering flat (B_f) and symmetrized reference (B_{sr}) priors. In a completely crossed experimental design I vary lengths of time series (i.e., T = 10, 25, 40, 50 and 100) and autocorrelation (from -0.90 to 0.90 with steps of 0.10). The results show the lowest bias for the B_{sr}, and the lowest variability for r_1. The power in different conditions is the highest for B_{sr} and OLS. For $T = 10$, the absolute performance of all measurements is poor, as expected. In Study 2, I study robustness of the methods through misspecification by generating the data according to an ARMA(1,1) model, but still analysing the data with an AR(1) model. I use the two methods with the lowest bias for this study, i.e., B_{sr} and MLE. The bias gets larger when the non-modelled moving average parameter becomes larger. Both the variability and power show a dependency on the non-modelled parameter. The differences between the two estimation methods are negligible for all measurements.

In **Chapter 3** I estimate a time series model for multiple individuals using multilevel models. I compare two estimation methods for the autocorrelation in Multilevel AR(1) models, namely MLE and Bayesian Markov Chain Monte Carlo, earlier denoted as B_{sr}. Furthermore, I examine the difference between modeling fixed and random individual parameters. To this end, I perform a simulation study with a fully crossed design, in which I vary the length of the time series (10 or 25), the number of individuals per sample (10 or 25), the mean of the autocorrelation (-0.6 to 0.6 inclusive, in steps of 0.3) and the standard deviation of the autocor-

relation (0.25 or 0.40). I found that the random estimators of the population autocorrelation show less bias and higher power, compared to the fixed estimators. As expected, the random estimators profit strongly from a higher number of individuals, while this effect is small for the fixed estimators. The fixed estimators profit slightly more from a higher number of time points than the random estimators. When possible, random estimation is preferred to fixed estimation. The difference between MLE and Bayesian estimation is nearly negligible. The Bayesian estimation shows a smaller bias, but MLE shows a smaller variability (i.e., standard deviation of the parameter estimates). Finally, better results are found for a higher number of individuals and time points, and for a lower individual variability of the autocorrelation. The effect of the size of the autocorrelation differs between outcome measures.

In **Chapter 4** I use the Bayesian Dynamic Model (BDM) to compare different models for a univariate, multi-individual dataset posing several challenges, such as missing data and non-normally distributed observed data. I represent the complex structure of intensive longitudinal data of multiple individuals with a hierarchical BDM. This BDM is a generalized linear hierarchical model where the individual parameters do not necessarily follow a normal distribution. The model parameters can be estimated on the basis of relatively small sample sizes and in the presence of missing time points. I present the BDM and discuss the model identification, convergence and selection. The use of the BDM model is illustrated using data from a randomized clinical trial to study the differential effects of three treatments for panic disorder. The data involves the number of panic attacks experienced weekly (73 individuals, 10 to 52 time points) during treatment. Presuming that the counts are Poisson distributed, the BDM considered involves a linear trend model with an exponential link function. The final model included a moving average parameter, and an external variable (duration of symptoms pre-treatment). Our results show that cognitive behavioral therapy is less effective on the reduction of panic attacks than serotonin selective re-uptake inhibitors or a combination of both. Post-hoc analyses revealed that males show a slightly higher number of panic attacks at the onset of treatment than females.

In **Chapter 5** I use the BDM for the analysis of multivariate, multi-individual emotional data with missing data points. In emotion dynamic research one distinguishes various elementary emotion dynamic features, which are studied using intensive longitudinal data. Typically, each emotion dynamic feature is quantified separately, which hampers the study of relationships between various features. Further, the length of the observed time series in emotion research is limited, and often suffers from a high percentage of missing values. In this chapter I propose a vector autoregressive Bayesian dynamic model, that is useful for emotion dynamic research. The model encompasses six elementary properties of emotions, and can be applied with relatively short time series, including missing data. The individual elementary properties covered are: within person and innovation variability, inertia, granularity, cross-lag correlation and the intensity. The model can be applied to both univariate and multivariate time series, allowing to model the relationships between emotions. Further, it may model multiple individuals jointly. One may include external variables and non-Gaussian observed data. We illustrate the usefulness of the model with an empirical example of three emotions of three individuals (47 to 70 measurements), with missing time points within the series.

Finally, in **Chapter 6**, I apply the vector autoregressive BDM to bivariate affect data where I compare several models to find the best fitting one. Affect research often focuses on studying various elementary features of the individual dynamics of positive affect (PA) and negative affect (NA). The features studied are typically quantified separately, hampering the study of their mutual relations. In this chapter I consider six elementary affect features to characterize the dynamics of PA and NA: within person and innovation variability, inertia, cross-lag regression, intensity and the co-occurrence of affect. To facilitate the study of these features, I propose to use a vector autoregressive Bayesian dynamic model. This model encompasses parameters that quantify the six affect features of multiple individuals at once. The model can be applied to data typically encountered in dynamic affect research, i.e., bivariate, relatively short time series of multiple individuals, even in the presence of missing time points. Furthermore, the model allows for the inclusion of external variables. I illustrate the usefulness of the model with an empirical example using relatively short time series (53 to 71 measurements) of bivariate affect data for 12 individuals. I compare three models to find whether a weekly cycle is present in the data and whether there is autoregression present in the white noise. I compare the models with regard to the white noise, the innovation noise and a log-likelihood criterion. The model estimates provide insight into the individual dynamics of PA and NA, and their interindividual differences and similarities.

In this thesis I studied the effect of different estimators and data properties on the estimation of the autocorrelation. I concluded that the MLE and the B_{sr} are the best estimators of those used in this thesis. I showed that the estimation is improved when the time series are longer, more individuals are involved, the absolute autocorrelation is larger and the standard deviation of the autocorrelation is smaller.

Furthermore, I studied how the BDM handles practical issues in empirical data and how it can link psychological hypotheses with statistical models. The BDM was capable of handling all practical issues encountered in one model, i.e., missing values, the inclusion of external variables and the use of observed scores with non-normally distributed residuals. Finally, the BDM is capable of handling the psychological research questions posed for the data sets analyzed in this thesis, for example by comparing models to see whether the hypothesized effects are indeed found in the data.

Samenvatting

In mijn proefschrift richt ik mij op het gebruik van tijdsreeksanalyse in psychologisch onderzoek. Ik focus hierbij op twee aspecten. Ten eerste onderzoek ik het effect van verschillende schatters en data-eigenschappen op het schatten van de autocorrelatie. Dit gebeurt in Hoofdstuk 2 en 3. Ten tweede bestudeer ik de uitdagingen van empirische data-analyse en hoe de link gelegd kan worden tussen psychologische theorieën en de statistische modellen. Dit gebeurt in Hoofdstuk 4, 5 en 6. Voor het tweede aspect gebruik ik het Bayesiaanse dynamische model, dat alle uitdagingen aankon die ik vond in de empirische datasets die ik heb geanalyseerd.

In **Hoofdstuk 2** bespreek ik verschillende schatters van het autoregressieve model voor univariate data. Ik vergelijk hun prestaties bij het schatten van de autocorrelatie in korte tijdsreeksen. In Studie 1, onder correcte modelspecificatie, vergelijk ik de frequentistische r_1, C-schatter, de ordinary least squares (OLS) schatter en de maximum likelihood estimator (MLE), alsmede een Bayesiaanse schattingsmethode met een platte prior (B_f), dan wel de zogenaamde symmetrized reference prior (B_{sr}). In een compleet gekruist onderzoeksontwerp varieer ik de lengte van de tijdsreeks (d.w.z., $T = 10, 25, 40, 50$ en 100) en de autocorrelatie (van -0.90 tot 0.90 met stappen van 0.10). De resultaten laten zien dat de kleinste bias (vertekening) wordt gevonden voor B_{sr} en de laagste variabiliteit voor r_1. In de verschillende condities is de power (het onderscheidingsvermogen) het grootst voor B_{sr} en OLS. Voor $T = 10$ presteren, zoals verwacht, alle schatters slecht In Studie 2 bestudeer ik de robuustheid van twee van de schatters aan de hand van een misspecificatie. In deze studie wordt de data gegenereerd met behulp van een ARMA(1,1) model, maar geanalyseerd met een AR(1) model. Hiervoor gebruik ik de twee methoden met de laagste bias in de eerste studie, d.w.z., B_{sr} en MLE. De bias wordt groter wanneer de ongemodelleerde autocorrelatieparameter van de ruis groter wordt. Zowel de variabiliteit als de power zijn afhankelijk van deze ongemodelleerde parameter. De verschillen tussen de MLE en B_{sr} zijn te verwaarlozen voor alle gebruikte uitkomstmaten.

In **Hoofdstuk 3** schat ik tijdsreeksmodellen voor meerdere individuen met behulp van multilevel modellen. Ik vergelijk hierbij twee schattingsmethoden voor de

autocorrelatie in multilevel AR(1) modellen, namelijk MLE en Bayesiaanse Markov Chain Monte Carlo, eerder B_{sr} genoemd. Verder bestudeer ik de verschillen tussen het modelleren van zogenaamde random en fixed individuele parameters. Om dit te bereiken voer ik een simulatiestudie uit met een volledig gekruist onderzoeksontwerp, waarbij ik de lengte van de tijdsreeksen (10 of 25) en de hoeveelheid individuen in de datasets (10 of 25) varieer, alsmede het gemiddelde (-0.60 tot 0.60 in stappen van 0.30) en de standaarddeviatie (0.25 of 0.40) van de verdeling waaronder ik de autocorrelatie genereer. De random schatters van de gemiddelde autocorrelatie in de populatie laten een kleinere bias en hogere power zien dan de fixed schatters. Zoals verwacht profiteren de random schatters sterk van een toegenomen aantal individuen, maar is dit effect klein voor de fixed schatters. De fixed schatters hebben iets meer voordeel van een toegenomen aantal tijdspunten dan de random schatters. Gelet op de hogere power en kleinere bias, wordt waar mogelijk het random schatten van de parameters geprefereerd boven het fixed schatten. Het verschil tussen de MLE en de B_{sr} is bijna verwaarloosbaar. De B_{sr} heeft een kleinere bias, maar MLE heeft een kleinere variabiliteit (i.e., de standaarddeviatie van de parameter schattingen). Er worden betere resultaten gevonden wanneer er meer individuen en tijdspunten zijn, evenals voor een kleine standaarddeviatie van de autocorrelatie. Het effect van het gemiddelde van de autocorrelatie is afhankelijk van de gebruikte uitkomstmaat.

In **Hoofdstuk 4** gebruik ik het Bayesiaanse dynamische model (BDM) om verschillende modelvarianten te vergelijken voor een univariate dataset bestaande uit metingen van meerdere individuen. Deze dataset vertoont verschillende uitdagingen, zoals ontbrekende scores en niet-normaal verdeelde residuen. Deze complexe structuur geef ik weer met een hiërarchisch BDM. Het BDM is een gegeneraliseerd lineair hiërarchisch model waar de individuele parameters niet per se een normale verdeling volgen. Het model kan geschat worden op basis van een relatief kleine dataset met ontbrekende scores. Ik presenteer het BDM en bespreek de modelidentificatie, convergentie en selectie. Het gebruik van het BDM illustreer ik aan de hand van data van een gerandomiseerde klinische proef, opgezet om het verschil in effect van drie behandelingen tegen paniekstoornissen te bestuderen. De data bestaan uit de hoeveelheid paniekaanvallen die wekelijks ervaren worden (73 individuen, 10 tot 52 datapunten per individu) tijdens de behandeling. Onder de aanname dat de scores een Poissonverdeling volgen, wordt er een model gebruikt met een lineaire trend en een exponentiele link functie. Het uiteindelijk model bevat ook een autocorrelatiecoëfficiënt voor de ruis en een externe variabele (tijdsduur van de symptomen voor de behandeling). Onze resultaten laten zien dat cognitieve gedragstherapie minder effectief is dan het gebruik van selectieve serotonine-heropnameremmers of een combinatie van beide in het verminderen van het aantal ervaren paniekaanvallen per week. Post-hoc analyses laten zien dat mannen een licht hoger aantal paniekaanvallen hebben aan het begin van de behandeling vergeleken met vrouwen.

In **Hoofdstuk 5** gebruik ik het BDM om een multivariate dataset met emotiedata van meerdere individuen met ontbrekende scores te analyseren. Onderzoek naar emotiedynamiek richt zich over het algemeen op het onthullen van informatie over het functioneren en reguleren van emoties en affect. Hierbij worden verschillende elementaire eigenschappen van de emotiedynamiek onderscheiden, die worden bestudeerd met behulp van intensieve longitudinale data. Over het algemeen wordt elk van deze eigenschappen apart gekwantificeerd, wat het bestuderen van de onderlinge relaties tussen deze eigenschappen bemoeilijkt. Dit, op zijn beurt, bemoeilijkt de validatie van de theorieën over deze onderlinge relaties. In emotieonderzoek is de lengte van de geobserveerde tijdsreeksen beperkt en is er vaak sprake van veel ontbrekende scores. In dit hoofdstuk gebruik ik een Bayesiaans vector autoregressief model, een variant van het BDM, dat gebruikt kan worden in het onderzoek naar de dynamiek van emoties. Het model omvat zes centrale eigenschappen van de emotiedynamiek in één keer en kan worden toegepast op relatief korte tijdsreeksen met ontbrekende scores. De gebruikte eigenschappen van emotiedynamiek zijn: de variabiliteit op de lange en korte termijn, de granulariteit, de inertia, de cross-lag regressie en de intensiteit. Het model kan worden toegepast op zowel univariate als multivariate tijdsreeksen, waardoor het de relaties tussen emoties ook kan modelleren. Verder kan het meerdere individuen tegelijk modelleren en indien gewenst externe variabelen meenemen. Het model kan gespecificeerd worden voor niet-normaal verdeelde geobserveerde data en omgaan met ontbrekende scores met behulp van een link functie. Ik laat het nut van het model zien aan de hand van een empirisch voorbeeld met relatief korte tijdsreeksen (47 tot 70 metingen) van drie emoties, met ontbrekende scores in de tijdsreeksen, gemeten voor drie individuen. Ik bespreek hoe het model kan worden uitgebreid en de beperkingen waar men tegenaan kan lopen wanneer het model gebruikt wordt voor de analyse van empirische data.

In **Hoofdstuk 6** pas ik het VAR-BDM toe op bivariate affect-data en vergelijk modellen om het best passende model te vinden. Affectonderzoek richt zich vaak op het bestuderen van verschillende elementaire eigenschappen van de individuele dynamiek van Positief Affect (PA) en Negatief Affect (NA). Deze eigenschappen worden meestal separaat gekwantificeerd, wat het bestuderen van de onderlinge relaties bemoeilijkt. Hier bespreek ik zes elementaire eigenschappen van affect die de dynamiek van PA en NA karakteriseren: de variabiliteit op de lange en korte termijn, de inertia, de cross-lag regressie, de intensiteit en de co-occurrence van affect. Om het bestuderen van deze eigenschappen te faciliteren gebruik ik een vector autoregressief BDM. Dit model bevat parameters die alle zes genoemde eigenschappen van de affect dynamiek kunnen kwantificeren. Het model kan worden toegepast op het soort data dat typisch is voor onderzoek naar affectdynamiek, i.e., bivariate, relatief korte tijd series van verschillende individuen, zo nodig met ontbrekende scores. Verder kan het model externe variabelen meenemen indien gewenst. Het nut van dit model laat ik zien aan de hand van een empirisch voorbeeld, waarbij ik een dataset gebruik van relatief korte (53 tot 71 metingen), bivariate affectdata van 12 individuen. Om de aanwezigheid van een weekdageffect dan wel een autoregressie in de afwijkingen vast te stellen vergelijk ik drie verschillende modellen. Ik vergelijk de modellen met betrekking tot de ruis, de innovatieruis en een log-likelihood criterium. De modelschattingen geven inzicht in de individuele dynamiek van PA en NA, en de bijbehorende interindividuele verschillen en overeenkomsten.

In dit proefschrift heb ik het effect onderzocht van verschillende schatters en data-eigenschappen op het schatten van de autocorrelatie. Uit mijn onderzoek blijkt dat, van de gebruikte schatters in dit proefschrift, de MLE en de B_{sr} de beste schatters zijn voor een autoregressief model. Verder blijkt dat het schatten van de autocorrelatie accurater wordt wanneer de tijdreeksen langer zijn, er meer mensen mee doen, de gemiddelde autocorrelatie in een serie groter is en de standaard deviatie van deze autocorrelatie kleiner is.

Daarnaast heb ik gekeken hoe het BDM omgaat met de problemen in empirische data en hoe het psychologische theorieën met statistische modellen kan verbinden. Het BDM kan alle problemen die aanwezig waren in de geanalyseerde datasets aan, d.w.z, missende waardes, de inclusie van externe variabelen en geobserveerde scores met niet-normaal verdeelde residuen. Ook kan het BDM de psychologische onderzoekvragen beantwoorden die van belang zijn voor de gebruikte datasets, bijvoorbeeld door het vergelijken van modellen om te kijken of een verwacht effect ook werkelijk te vinden is in de data.

Bibliography

Altamura, A., Santini, A., Salvadori, D., & Mundo, E. (2005). Duration of untreated illness in panic disorder: a poor outcome risk factor? *Neuropsychiatric disease and treatment*, *1*(4), 345–347. doi:DOI:10.1016/S0924-977X(03)92195-X

Areni, C. S. (2008). (Tell me why) I don't like Mondays: Does an overvaluation of future discretionary time underlie reported weekly mood cycles? *Cognition & Emotion*, *22*(7), 1228–1252. doi:10.1080/02699930701686107

Arnau, J. & Bono, R. (2001). Autocorrelation and Bias in Short Time Series: An Alternative Estimator. *Quality and Quantity*, *35*(4), 365–387. doi:10.1023/A:1012223430234

Bandelow, B., Behnke, K., Lenoir, S., Hendriks, G. J., Alkin, T., Goebel, C., & Clary, C. M. (2004). Sertraline Versus Paroxetine in the Treatment of Panic Disorder. *Journal of Clinical Psychiatry*, *65*(3), 405–413. doi:10.4088/jcp.v65n0317

Barrett, L. F. & Barrett, D. (2000). The Experience Sampling Program. Retrieved from http://www.experience-sampling.org

Barrett, L. F. & Gross, J. J. (2001). Emotional intelligence: A process model of emotion representation and regulation. In T. J. M. G. A. Bonanno (Ed.), *Emotions: currrent issues and future directions* (pp. 286–310). Emotions and social behavior. New York: The Guilford Press.

Barrett, L. F., Gross, J., Christensen, T. C., & Benvenuto, M. (2001). Knowing what you're feeling and knowing what to do about it: Mapping the relation between emotion differentiation and emotion regulation. *Cognition and Emotion*, *15*(6), 713–724. doi:10.1080/02699930143000239

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

Berger, J. O. & Yang, R.-Y. (1994). Noninformative priors and Bayesian testing for the AR(1) model. *Econometric Theory*, *10*(3-4), 461–482. doi:10.1017/S026646660000863X

Bolger, N. & Laurenceau, J.-P. (2013). *Methodology In The Social Sciences: Intensive Longitudinal Methods: An Introduction to Diary and Experience Sampling Research*. New York: The Guildford Press.

Bos, F. M., Schoevers, R. A., & Aan het Rot, M. (2015). Experience sampling and ecological momentary assessment studies in psychopharmacology: A systematic review. *European neuropsychopharmacology*, *25*(18), 1853–1864. doi:10.1016/j.euroneuro.2015.08.008

Box, G. E. P. & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control*. San Francisco, CA: Holden-Day.

Brans, K., Koval, P., Verduyn, P., Lim, Y. L., & Kuppens, P. (2013). The regulation of negative and positive affect in daily life. *Emotion*, *13*(5), 926–939. doi:10.1037/a0032400

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., ... Tuerlinckx, F. (2013). A network approach to psychopathology: new insights into clinical longitudinal data. *PloS one*, *8*(4), e60188. doi:10.1371/journal.pone.0060188

Brose, A., de Roover, K., Ceulemans, E., & Kuppens, P. (2015). Older adults' affective experiences across 100 days are less variable and less complex than younger adults'. *Psychology and Aging*, *30*(1), 194–208. doi:10.1037/a0038690

Bryk, A. S. & Raudenbush, S. W. (1992). *Hierarchical linear models: applications and data analysis methods*. Newbury Park, CA: SAGE publications.

Bulteel, K., Tuerlinckx, F., Brose, A., & Ceulemans, E. (2016). Using Raw VAR Regression Coefficients to Build Networks can be Misleading. *Multivariate Behavioral Research*, 1–15. doi:10.1080/00273171.2016.1150151

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208. doi:10.1137/0916069

Carstensen, L. L., Pasupathi, M., Mayr, U., & Nesselroade, J. R. (2000). Emotional experience in everyday life across the adult life span. *Journal of Personality and Social Psychology*, *79*(4), 644–655. doi:10.1037/0022-3514.79.4.644

Chaubert, F., Mortier, F., & Saint André, L. (2008). Multivariate dynamic model for ordinal outcomes. *Journal of Multivariate Analysis*, *99*(8), 1717–1732. doi:http://dx.doi.org/10.1016/j.jmva.2008.01.011

Cook, J., Tyson, R., White, J., Rushe, R., Gottman, J., & Murray, J. (1995). Mathematics of marital conflict: Qualitative dynamic mathematical modeling of marital interaction. *Journal of Family Psychology*. Methodological Advances in Family Psychology, *9*(2), 110–130. doi:10.1037/0893-3200.9.2.110

Cox, D. D. & Llatas, I. (1991). Maximum Likelihood Type Estimation for Nearly Nonstationary Autoregressive Time Series. *Annals of Statistics*, *19*(3), 1109–1128. doi:10.1214/aos/1176348240

Croft, G. P. & Walker, A. E. (2001). Are the Monday blues ail in the mind? The role of expectancy in the subjective experience of mood. *JASP Journal of Applied Social Psychology*, *31*(6), 1133–1145. doi:10.1111/j.1559-1816.2001.tb02666.x

DeCarlo, L. T. & Tryon, W. W. (1993). Estimating and testing autocorrelation with small samples: a comparison of the C-statistic to a modified estimator. *Behaviour Research And Therapy*, *31*(8), 781–788. doi:10.1016/0005-7967(93)90009-J

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–38. Retrieved from http://www.jstor.org/stable/2984875

Dua, J. K. (1993). The role of negative affect and positive affect in stress, depression, self-esteem, assertiveness, Type A behaviors, psychological health, and physical health. *Genetic, Social & General Psychology Monographs*, *119*(4), 517.

Durbin, J. & Koopman, S. J. (2012). *Time series analysis by state space models*. Oxford, UK: Oxford University Press.

Egloff, B., Tausch, A., Kohlmann, C.-W., & Krohne, H. W. (1995). Relationships between time of day, day of the week, and positive mood: Exploring the role of the mood measure. *Motivation and Emotion*, *19*(2), 99–110. doi:10.1007/BF02250565

Elkins, S. R. & Moore, T. M. (2011). A Time-Series Study of the Treatment of Panic Disorder. *Clinical Case Studies*, *10*(1), 3–22. doi:10.1177/1534650110391901

Erbas, Y., Ceulemans, E., Pe, M. L., Koval, P., & Kuppens, P. (2014). Negative emotion differentiation: Its personality and well-being correlates and a comparison of different assessment methods. *Cognition and Emotion*, *28*(7), 1196–1213. doi:10.1080/02699931.2013.875890

Federici, C. & Tommasini, N. (1992). The assessment and management of panic disorder. *The Nurse practitioner*, *17*(3), 20–22. doi:10.1097/00006205-199203000-00007

Feldman, L. A. (1995). Valence focus and arousal focus: Individual differences in the structure of affective experience. *Journal of Personality and Social Psychology*, *69*(1), 153–166. doi:10.1037/0022-3514.69.1.153

Frijda, N. H. (2007). *The laws of emotion*. Mahwah, NJ: Erlbaum.

Garcia-Hiernaux, A., Casals, J., & Jerez, M. (2009). Fast estimation methods for time-series models in state-space form. *Journal of Statistical Computation and Simulation*, *79*(2), 121–134. doi:10.1080/00949650701617249

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, *100*(470), 680–701. doi:10.1198/016214505000000105

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd). Roca Baton, FL: Chapman and Hall/CRC.

Gelman, A., Hwang, J., & Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. doi:10.1007/s11222-013-9416-2

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

Goldstein, H. (2011). *Multilevel statistical models* (4th). Hoboken, N.J.: Wiley. doi:10.1002/9780470973394

Goldstein, H., Healy, M. J. R., & Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine*, *13*(16), 1643–1655. doi:10.1002/sim.4780131605

Gross, J. J. (2015). The extended Process model of emotion regulation: elaborations, applications, and future directions. *Psychological Inquiry*, *26*(1), 130–137. doi:10.1080/1047840X.2015.989751

Grühn, D., Lumley, M. A., Diehl, M., & Labouvie-Vief, G. (2013). Time-based indicators of emotional complexity: Interrelations and correlates. American Psychological Association. doi:10.1037/a0030363

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*. doi:10.1007/s11336-014-9417-x

Hamaker, E. L., Ceulemans, E., Grasman, R. P. P. P., & Tuerlinckx, F. (2015). Modeling Affect Dynamics: State of the Art and Future Challenges. *Emotion Review*, *7*(4), 316–322. doi:10.1177/1754073915590619

Hamaker, E. L. & Grasman, R. P. P. (2012). Regime switching state-space models applied to psychological processes: Handling missing data and making inferences. *Psychometrika*, *77*(2), 400–422. doi:10.1007/s11336-012-9254-8

Hamilton, J. D. (1994). *Time series analysis*. Princeton, N.J., Princeton University Press.

Harvey, A. C. (1990). Forecasting, structural time series models, and the Kalman filter. Cambridge, UK: Cambridge University Press.

Harvey, B., Milyavskaya, M., Hope, N., Powers, T. A., Saffran, M., & Koestner, R. (2015). Affect variation across days of the week: influences of perfectionism and academic motivation. *Motiv Emot Motivation and Emotion*, *39*(4), 521–530. doi:10.1007/s11031-015-9480-3

Harville, D. A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, *72*(358), 320–338. Retrieved from http://www.jstor.org/stable/2286796

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109. doi:10.1093/biomet/57.1.97

Hay, E. L. & Diehl, M. (2011). Emotion complexity and emotion regulation across adulthood. *European journal of ageing*, *8*(3), 157–168. doi:10.1007/s10433-011-0191-7

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*. doi:10.1037/a0038822

Hox, J. J. (2010). *Multilevel Analysis* (2nd). Quantitative methodology. Routledge.

Huitema, B. E. & McKean, J. W. (1991). Autocorrelation estimation and inference with small samples. *Psychological Bulletin*, *110*(2), 291–304. doi:10.1037/0033-2909.110.2.291

Huitema, B. E. & McKean, J. W. (1994). Reduced Bias Autocorrelation Estimation: Three Jackknife Methods. *Educational and Psychological Measurement*, *54*(3), 654. Retrieved from http://www.editlib.org/p/78923

Jahng, S., Wood, P., & 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. doi:10.1037/a0014173

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. doi:10.1080/00273171.2014.1003772

Kashdan, T. B. & Farmer, A. S. (2014). Differentiating emotions across contexts: Comparing adults with and without social anxiety disorder using random, social interaction, and daily experience sampling. *Emotion*, *14*(3), 629–638. doi:10.1037/a0035796

Kashdan, T. B. & Rottenberg, J. (2010). Psychological flexibility as a fundamental aspect of health. *Clinical Psychology Review Clinical Psychology Review*, *30*(7), 865–878. doi:10.1016/j.cpr.2010.03.001

Kellett, S. (2007). A time series evaluation of the treatment of histrionic personality disorder with cognitive analytic therapy. *Psychology and Psychotherapy: Theory, Research and Practice*, *80*(3), 389–405. doi:10.1348/147608306X161421

Kendall, S. M. & Ord, J. K. (1990). *Time series*. London, Great Britain: Edward Arnold.

Khlat, M., Legleye, S., & Sermet, C. (2013). Factors Influencing Report of Common Mental Health Problems Among Psychologically Distressed Adults. *Community Mental Health Journal*, 1–7. doi:10.1007/s10597-013-9680-9

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), 625–636. doi:10.1037/emo0000059

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. doi: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), 1132–1141. doi:10.1037/a0033579

Krone, T., Albers, C. J., & Timmerman, M. E. (2016a). A comparative simulation study of AR(1) estimators in short time series. *Quality & Quantity (in press)*. doi:10.1007/s11135-015-0290-1

Krone, T., Albers, C. J., & Timmerman, M. E. (2016b). A multivariate model for emotion dynamics. *Submitted*.

Krone, T., Albers, C. J., & Timmerman, M. E. (2016c). Bayesian dynamic modeling to assess differential treatment effects on panic attack frequencies. *Statistical Modelling (in press)*. doi:10.1177/1471082X16650777

Krone, T., Albers, C. J., & Timmerman, M. E. (2016d). Comparison of estimation procedures for multilevel AR(1) models. *Frontiers in Psychology*, *7*(486). doi:10.3389/fpsyg.2016.00486

Kunitomo, N. & Yamamoto, T. (1985). Properties of Predictors in Misspecified Autoregressive Time Series Models. *Journal of the American Statistical Association*, *80*(392), 941. doi:10.2307/2288558

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

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. doi:10.1037/a0020962

Kuppens, P. & Verduyn, P. (2015). Looking at emotion regulation through the window of emotion dynamics. *Psychological Inquiry*, *26*(1), 72–79. doi:10.1080/1047840X.2015.960505

Larsen, R. J. (2000). Toward a Science of Mood Regulation. *Psychological Inquiry*, *11*(3), 129–141. doi:10.1207/S15327965PLI1103_01

Larsen, R. J. & Kasimatis, M. (1990). Individual differences in entrainment of mood to the weekly calendar. *Journal of personality and social psychology*, *58*(1), 164–171. doi:10.1037//0022-3514.58.1.164

Larson, R. & Csikszentmihalyi, M. (1983). The experience sampling method. In *New directions for methodology of social and behavioral sciences* (Vol. 15, pp. 41–56).

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. doi:10.1016/j.jmva.2009.04.008

Lewis, M., Haviland-Jones, J. M., & Feldman Barret, L. (Eds.). (2008). *Handbook of emotions* (3rd). New York: The Guildford Press.

Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, *73*(1), 13–22. doi:10.1093/biomet/73.1.13

Liu, S. & Molenaar, P. C. M. (2014). iVAR: A program for imputing missing data in multivariate time series using vector autoregressive models. *Behavior Research Methods*. doi:10.3758/s13428-014-0444-4

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. doi:10.1016/j.jmp.2010.08.004

Lucas, R. E. & Fujita, F. (2000). Factors influencing the relation between extraversion and pleasant affect. *Journal of Personality and Social Psychology*, *79*(6), 1039–1056. doi:10.1037/0022-3514.79.6.1039

Lucas, R. E., Le, K., & Dyrenforth, P. S. (2008). Explaining the extraversion/positive affect relation: Sociability cannot account for extraverts' greater happiness. *Journal of Personality*, *76*(3), 385–414. doi:10.1111/j.1467-6494.2008.00490.x

Lütkepohl, H. (1991). *Introduction to multiple time series analysis*. New York; Berlin; London and Tokyo:

McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models*. London; New York: Chapman and Hall.

Merz, E. L. & Roesch, S. C. (2011). Modeling trait and state variation using multilevel factor analysis with PANAS daily diary data. *Journal of Research in Personality*, *45*, 2–9. doi:10.1016/j.jrp.201.11.003

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087–1092. doi:10.1063/1.1699114

Morey, R. D. & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*(4), 406–419. doi:10.1037/a0024377

Morey, R. D., Rouder, J. N., & Jamil, T. (2015). BayesFactor: Computation of Bayes factors for common designs. Retrieved from http://bayesfactorpcl.r-forge.r-project.org/

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, *135*(3), 370–384. doi:10.2307/2344614

Ong, A. D. & Bergeman, C. S. (2004). The complexity of emotions in later life. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *59*(3), P117–P122. doi:10.1093/geronb/59.3.P117

Oravecz, Z., Tuerlinckx, F., & Vandekerckhove, J. (2011). A hierarchical latent stochastic differential equation model for affective dynamics. Oravecz, Zita:

Department of Psychology, University of Leuven, Tiensestraat 102 Box 3713, Leuven, Belgium, B-3000, zita.oravecz@psy.kuleuven.be: American Psychological Association. doi:10.1037/a0024375

Pantula, S. G. & Fuller, W. A. (1985). Mean estimation bias in least squares estimation of autoregressive processes. *Journal of Econometrics*, *77*(1), 99–121. doi:10.1016/0304-4076(85)90046-6

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 Clinical Psychological Science*, *3*(2), 292–300. doi:10.1177/2167702614540645

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), 1320–1328. doi:10.1037/a0028262

Petris, G., Petrone, S., & Campagnoli, P. (2009). *Dynamic Linear Models with R.* New York, NY: Springer.

Pole, A., West, M., & Harrison, J. (1994). *Applied Bayesian Forecasting and Time Series Analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Powell, M. J. D. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. Technical Report DAMTP 2009/NA06. Cambridge, UK: University of Cambridge.

Price, L. R. (2012). Small Sample Properties of Bayesian Multivariate Autoregressive Time Series Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(1), 51–64. doi:10.1080/10705511.2012.634712

R Core Team. (2015). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from http://www.r-project.org/

Ready, R. E., Carvalho, J. O., & Weinberger, M. I. (2008). Emotional Complexity in Younger, Midlife, and Older Adults. *Psychology and aging*, *23*(4), 928–933. doi:10.1037/a0014003

Röcke, C., Li, S.-C., & Smith, J. (2009). Intraindividual variability in positive and negative affect over 45 days: Do older adults fluctuate less than young adults? *Psychology and Aging*, *24*(4), 863–878. doi:10.1037/a0016276

Ross, D., Klein, D., & Uhlenhuth, E. (2010). Improved statistical analysis of moclobemide dose effects on panic disorder treatment. *European archives of psychiatry and clinical neuroscience*, *260*(3), 243–248. doi:10.1007/s00406-009-0062-9

Rush, J. & Hofer, S. M. (2014). Differences in within- and between-person factor structure of positive and negative affect: Analysis of two intensive measurement studies using multilevel SEM. *Psychological assessment*, *26*(2), 462–473. doi:10.1037/a0035666

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion*, *23*(7), 1307–1351. doi:10.1080/02699930902928969

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*. doi:10.1037/met0000062

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

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. doi:10.1214/aos/1176344136

Scott, S. B., Sliwinski, M. J., Mogle, J. A., & Almeida, D. M. (2014). Age, stress, and emotional complexity: Results from two studies of daily experiences. *Psychology and Aging*, *29*(3), 577–587. doi:10.1037/a0037282

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual review of clinical psychology*, *4*, 1–32. doi:10.1146/annurev.clinpsy.3.022806.091415

Shifren, K., Hooker, K., Wood, P., & Nesselroade, J. R. (1997). Structure and variation of mood in individuals with Parkinson's disease: A dynamic factor analysis. *Psychology and aging*, *12*(2), 328–339. doi:10.1037/0882-7974.12.2.328

Snijders, T. A. B. & Bosker, R. (1999). *Multilevel Analysis*. London: SAGE publications.

Solanas, A., Manolov, R., & Sierra, V. (2010). Lag-one autocorrelation in short series: Estimation and hypotheses testing. *Psicologica: International Journal of Methodology and Experimental Psychology*, *31*(2), 357–381.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639. doi:10.1111/1467-9868.00353

Stan Development Team. (2014). Rstan: the R interface to Stan, version 2.4. Retrieved from http://mc-stan.org/rstan.html

Stan Development Team. (2015). RStan: the R interface to Stan, Version 2.9. Retrieved from http://mc-stan.org/rstan.html

Stan Development Team. (2016). *Stan modeling language users guide and reference manual, version 2.9.0*. Retrieved from http://mc-stan.org/

Stoica, P., Friedlander, B., & Söderstorm, T. (1986). Least-squares, Yule-Walker, and overdetermined Yule-Walker estimation of AR parameters: a Monte Carlo analysis of finite-sample properties. *International Journal of Control*, *43*(1), 13–27. doi:10.1080/00207178608933446

Suls, J., Green, P., & Hills, S. (1998). Emotional reactivity to everyday problems, affective inertia and Neuroticism. *Personality and social psychology bulletin*, *24*(2), 127–136. doi:10.1177/0146167298242002

Swami, V. (2012). Mental health literacy of depression: Gender differences and attitudinal antecedents in a representative British sample. *PLoS ONE*, *7*(11). doi:10.1371/journal.pone.0049779

Tanaka, K. (1984). An Asymptotic Expansion Associated with the Maximum Likelihood Estimators in ARMA Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, *46*(1), 58–67. Retrieved from http://www.jstor.org/stable/2345462

Tanaka, K. & Maekawa, K. (1984). The sampling distributions of the predictor for an autoregressive model under misspecifications. *Journal of Econometrics*, *25*(3), 327–351. doi:http://dx.doi.org/10.1016/0304-4076(84)90005-8

Terui, N., Ban, M., & Maki, T. (2010). Finding Market Structure by Sales Count Dynamics - Multivariate Structural Time Series Models With Hierarchical Structure for Count Data. *Annals of the Institute of Statistical Mathematics*, *62*(1), 91–107. doi:10/1007/s10463-009-0244-2

Tiao, G. C. & Box, G. E. P. (1981). Modeling multiple time series with applications. *Journal of the American Statistical Association*, *76*(376), 802–816. doi:10.2307/2287575

Toni, C., Perugi, G., Frare, F., Mata, B., & Akiskal, H. (2004). Spontaneous treatment discontinuation in panic disorder patients treated with antidepressants. *Acta psychiatrica Scandinavica*, *110*(2), 130–137. doi:10.1111/j.1600-0047.2004.00347.x

Trull, T. J., Lane, S. P., Koval, P., & Ebner-Priemer, U. W. (2015). Affective dynamics in psychopathology. *Emotion Review*, *7*(4), 355–361. doi:10.1177/1754073915590617

Tugade, M. M., Fredrickson, B. L., & Feldman Barrett, L. (2004). Psychological resilience and positive emotional granularity: Examining the benefits of positive emotions on coping and health. *Journal of Personality*, *72*(6), 1161–1190. doi:10.1111/j.1467-6494.2004.00294.x

Van Apeldoorn, F. J., Van Hout, W. J. P. J., Timmerman, M. E., Mersch, P. P. A., & Den Boer, J. A. (2013). Rate of improvement during and across three treatments for panic disorder with or without agoraphobia: Cognitive behavioral therapy, selective serotonin reuptake inhibitor or both combined. *Journal of Affective Disorders*, *150*(2), 313–319. doi:10.1016/j.jad.2013.04.012

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. doi:10.1016/j.jrp.2004.05.006

Vehtari, A., Gelman, A., & Gabry, J. (2015). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. Retrieved from https://github.com/jgabry/loo

Walker, G. (1931). On Periodicity in Series of Related Terms. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *131*(818), 518–532. doi:10.1175/1520-0493(1931)59⟨277:OPISOR⟩2.0.CO;2

Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, *11*, 3571–3594. Retrieved from http://www.jmlr.org/papers/volume11/watanabe10a/watanabe10a.pdf

Watson, D., Clark, L. A., McIntyre, C. W., & Hamaker, S. (1992). Affect, personality, and social activity. *Journal of personality and social psychology*, *63*(6), 1011–1025. doi:10.1037/0022-3514.63.6.1011

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), 1063–1070. doi:10.1037/0022-3514.54.6.1063

West, K. D. & Wilcox, D. W. (1996). A Comparison of Alternative Instrumental Variables Estimators of a Dynamic Linear Model. *Journal of Business and Economic Statistics*, *14*(3), 281–293. doi:10.1080/07350015.1996.10524657

West, M. & Harrison, J. (1997). *Bayesian forecasting and dynamic models*. New York, NY: Springer Verlag.

Wetzels, R. & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic Bulletin & Review*, *19*(6), 1057–1064. doi:10.3758/s13423-012-0295-x

Wichers, M., Wigman, J. T. W., & Myin-Germeys, I. (2015). Micro-level affect dynamics in psychopathology viewed from complex dynamical system theory. *Emotion Review*, *7*(4), 362–367. doi:10.1177/1754073915590623

Wild, B., Quenter, A., Friederich, H.-C., Schild, S., Herzog, W., & Zipfel, S. (2006). A course of treatment of binge eating disorder: a time series approach. *ERV European Eating Disorders Review*, *14*(2), 79–87. doi:10.1002/erv.673

Young, L. C. (1941). on Randomness in ordered sequences. *The annals of mathematical statistics*, *12*(3), 293–300. doi:10.1214/aoms/1177731711

Yule, G. U. (1927). On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *226*, 267–298. doi:10.1098/rsta.1927.0007

Zhang, Z., Hamaker, E. L., & Nesselroade, J. R. (2008). Comparisons of Four Methods for Estimating a Dynamic Factor Model. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(3), 377–402. doi:10.1080/10705510802154281

Curriculum Vitae

Tanja Krone (1985) obtained both her Bachelor degree (2008) and her Master degree (2009) in Clinical Psychology at the University of Groningen. Afterwards, she continued to work as a research assistant in the research project of her masters internship. Research, and especially the analysis of research data, proved to be more interesting to her than the work of a clinical psychologist. As such, she enrolled in the specialization "Psychometrics and Statistics" of the research master program in Psychology of the University of Groningen. During her research master she taught statistics and methodology courses to Bachelor students and was part of the yearbook-committee. The highlight of her research master was a four-month period where she studied at the University of Oslo under Prof. Knut Hagtvet.

In 2011 she graduated cum laude on her research master, and headed straight into her PhD project. This PhD project was funded by a scholarship from the Netherlands Organisation for Scientific Research, awarded to her and her supervisors, Casper Albers and Marieke Timmerman. In her PhD project she studied time series analysis using several estimation methods and models, eventually focusing on Bayesian dynamic models. She also organized monthly research meetings for her department. Additionally, she taught practical classes with regard to statistics and methodology, and she supervised several bachelor theses. For her own eduction, she followed several graduate courses on topics such as Bayesian statistics, latent variable modeling and advising on research methods.

Dankwoord

De afgelopen jaren heb ik hard gewerkt aan hetgeen voor u ligt. Dit was mij niet gelukt zonder een aantal belangrijke mensen, en was een stuk minder plezierig geweest zonder een grote groep vrienden en collega's.

Ten eerste wil ik mijn begeleiders bedanken, Casper en Marieke. We hebben bijna vijf jaar intensief samengewerkt om dit resultaat te bereiken. Casper, het was zo fijn dat ik op elk mogelijk moment jouw kamer binnen kon stormen en jij aan een half woord genoeg had om mij te begrijpen. Mijn interesse, kennis en liefde voor de wondere wereld van Bayesiaanse analyse is voor een groot deel aan jou te danken. Marieke, je bent onontbeerlijk geweest voor het ontwikkelen van mijn academische schrijfkunsten. De discussies over zowel hele paragrafen als enkele woordjes hebben mij enorm gestimuleerd. De kritische blik die jij hebt op veel psychologische en statistische modellen, en op de mogelijkheden van de gebruikte statistiek, heeft gezorgd dat dit boekje zo goed is geworden als het nu is. De wekelijkse bijeenkomsten met ons drieën waren niet alleen een bron van motivatie en inspiratie, maar zorgden ook nog wel eens voor het nodige comic relief.

Onmisbaar waren natuurlijk mijn collega's. Rob, die als afdelingshoofd een rots in de branding was, en altijd bereid is te luisteren naar aanwezige problemen. Ik bedank mijn reeks aan kamergenoten: Tam, you are an awesome girl and a really easygoing roommate. It was an honour to be your paranimph. Mariska, Florian, Susan K. en Rivka die voor kortere danwel iets langere tijd mijn kamer deelden. Mijn uiteindelijke kamergenoot Nitin, die tevens mijn paranimf is. Nitin, our discussions on anything relating to science, religion or whatever we felt like were very stimulating. Although you distracted me a bit from serious work at times, you made me think. I like that in a roommate.

Verder wil ik natuurlijk Susan, Anja, Karin, Iris, Daniela en Lieke bedanken. Dames, ik vond het heerlijk om regelmatig even te kletsen over serieuze en minder serieuze dingen. Zonder jullie was mijn PhD-periode veel saaier geweest. Jorge, with whom I could always chit-chat about computers and other small stuff. En natuurlijk bedankt aan alle andere collega's die op de afdeling en in de buurt rond lopen of liepen en zo de afdeling een stukje beter maakten, met name Alwin, Don, Edith, Frans, Hanny, Henk, Iris S. en Richard.

Peter Kuppens wil ik graag bedanken voor de prettige samenwerking aan hoofdstuk 5. De combinatie van onze statistische benadering met jouw kennis van emotie psychologie heeft het zeker een beter hoofdstuk gemaakt.

Het IOPS is een belangrijk onderdeel geweest van mijn PhD-periode. Vooral de conferenties met de gezellige diners en de memorabele avonden. Alle PhD's die ik daar heb leren kennen en met wie ik heb gefeest, gegeten, cursussen heb gevolgd en heb geklaagd over journals, begeleiders en de wereld in het algemeen: bedankt. Ook wil ik Janneke bedanken. Ik heb je maar kort gekend, maar ik zal je nooit vergeten.

Uiteindelijk is er natuurlijk ook leven naast het werk, zelfs voor PhD-studenten. Voor ontelbare dames-avonden, sauna-dagen en andere ontspanning wil ik Carmen, Gitta, Judith en Marije bedanken. Voor klaverjassen tot we erbij neervielen: Hans, Jochem, Steven en Teije. Jullie bij elkaar vallen onder Selwerdtuig. Mijn soort tuig.

Verder een speciaal bedankje voor al mijn lunchpartners: Annika, Kelly, Meta, Roos, Staas, en al die mensen met wie ik verder heb geluncht. Niets breekt een lange dag zo goed als een heerlijke lunch. Maar vooral voor Boy. Mijn paranimf en mijn rots in de branding, wie ik altijd lastig kon vallen voor lunch, thee, taart of gewoon even klagen. Dat er nog maar veel songfestivals mogen volgen, waar we ook moge zijn, en nog ontelbaar heerlijke etentjes.

Mijn eeuwige dank gaat uit naar mijn moeder, die altijd in mij heeft geloofd, en mijn vader, die toch wel heel trots is op mij, naar het schijnt. Daarnaast is er nog mijn broertje die altijd bereid is mij te helpen om mijn vader en moeder in te maken met een potje klaverjas.

En vooral wil ik mijn belangrijkste steun en toeverlaat bedanken. Degene die mij thee brengt als ik moe ben en mijn boksbal vasthoudt als ik mijn frustratie kwijt moet. Degene die al jaren elk moment van elke dag voor me klaar heeft gestaan, en dat hopelijk nog lang doet. Tjitte, je bent het beste wat mij is overkomen en ik prijs mijzelf gelukkig dat ik elke dag naast jou wakker word.

O, en natuurlijk Flits en Odin. Geweldige excuus-katten en onuitputtelijke bron van kopjes.