Some notes on Bayesian time series analysis in psychology

Krone, Tanja

Publication date:
2016

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.