# Comparison of Estimation Procedures for Multilevel AR(1) Models

Tanja Krone *, Casper J. Albers and Marieke E. Timmerman

*Department of Psychometrics and Statistics, Heymans Institute, University of Groningen, Groningen, Netherlands*

To estimate a time series model for multiple individuals, a multilevel model may be used. In this paper 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.

Keywords: time series analysis, autocorrelation, Bayesian MCMC, multisubject, maximum likelihood estimation, simulation study

## 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 and Laurenceau, 2013). The data provided through ecological momentary assessment, often denoted as intensive longitudinal data (Hamaker et al., 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 et al., 2015), dynamics of depression (Kuppens et al., 2010; Erbas et al., 2014; Kashdan and Farmer, 2014), and the relation between affect and stress (Scott et al., 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 and 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 and 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 and 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 et al. (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 et al. (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 et al. (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 and 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 (Huitema and McKean, 1991; DeCarlo and Tryon, 1993; Arnau and Bono, 2001; Solanas et al., 2010; Krone et al., 2015). 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., 2015). 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., 2015). This is consistent with the advice of a lower bound of 50 time points for any time series modeled with an AR(1) model, as given by Box and Jenkins (1976).

In this paper, 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 et al. (2015) compared the 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 paper. 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 and Zeger, 1986). Although these methods have the benefit of being faster than iterative methods, i.e., MLE and 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 Maximum Likelihood Estimation (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 paper (Krone et al., 2015).

In this paper 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 Markov-Chain Monte Carlo (Bayesian MCMC) estimation. In the next part of this paper, 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.

## 2. THE MULTILEVEL AUTOREGRESSIVE LAG 1 MODEL

The ML-AR(1) model we use is a random coefficients model (e.g., Snijders and 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 and Jenkins, 1976):

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

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

In this paper 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 paper.

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:

$$\mu_n = \gamma_{0,0} + U_{0,\,n}, \qquad\qquad (2)$$
$$\phi_n = \gamma_{0,1} + U_{1,\,n}, \qquad\qquad (3)$$

with:

$$U_{0,\,n} \sim N(0, \sigma_{U_{0,\,n}}), \qquad\qquad (4)$$
$$U_{1,\,n} \sim N(0, \sigma_{U_{1,\,n}}). \qquad\qquad (5)$$

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, $\sigma_e$, 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:

$$y_{t,n} = \gamma_{0,0} + \gamma_{0,1}(y_{t-1,n} - \gamma_{0,0} - U_{0,\,n}) + U_{0,\,n} +$$
$$U_{1,\,n}(y_{t-1,n} - \gamma_{0,0} - U_{0,\,n}) + e_{t,\,n}, \quad e_{t,\,n} \sim N(0, \sigma_e)(6)$$

## 2.1. Estimation Methods
### 2.1.1. 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 and 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 et al., 1977), also used in latent variable modeling, may be used. However, in this paper we will assume that the full data is available.

### 2.1.2. 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 et al., 1953; Hastings, 1970) that allows for an efficient estimation of the parameters (Gelman 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.

## 2.2. Procedure
In this study, we aim to examine the comparative quality of MLE and Bayesian MCMC estimation for the autocorrelation

**FIGURE 1 | Flowchart of the study design of the simulation study.**
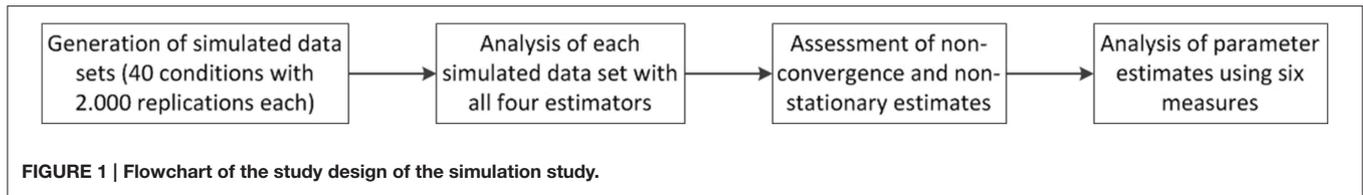
parameter in random and fixed ML-AR(1) models. This results in four estimators which will be compared: MLE-F, MLE-R, BAY-F, and BAY-R. For the Bayesian MCMC estimations, we use the program Rstan (Stan Development Team, 2014). For the estimation of the MLE-R, we use the package lme4 for R (Bates et al., 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. SIMULATION STUDY

### 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 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 (5) and (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 (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 $\phi_n$ were then drawn from a truncated and rescaled normal distribution with range $-1$ to 1, to ensure the resulting time series were stationary:

$$\phi_n \propto N(\gamma_{0,1}, \sigma_{U_{1,n}})\tau(-1, 1). \tag{7}$$

### 3.1.1. 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 2000 replications of a single simulation condition, using 5000 iterations, taking $\gamma_{0,0} = 0.00$, $\sigma_{U_{1,n}} = 0.40$, $T = 10$ and $N = 10$ (see Equation 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 and Yang, 1994), which has shown to better perform than the flat prior for single case AR(1) models (Krone et al., 2015). This prior does not need hyperparameters.

We tested several hyperparameters for the prior distributions of $\mu_n$ and $\sigma_e$ for the fixed model, and $\gamma_{0,0}$, $\sigma_{U_{0,n}}$, $\sigma_e$ and $\sigma_{U_{1,n}}$ of the random model, as shown in **Table 1**. Our parameter of

**TABLE 1 | Different combinations of priors tested to see their influence on the posterior results.**

| Test | Fixed model | | Random model | | | |
|---|---|---|---|---|---|---|
| | $\mu_n$ | $\sigma_e$ | $\mu$ | $\sigma_\mu$ | $\sigma_e$ | $\sigma_\phi$ |
| 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 | | | $N \sim (0, 2)$ | $\Gamma \sim (2, 2)$ | $\Gamma \sim (2, 2)$ | $\Gamma \sim (1, 2)$ |

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 1**, as these gave the best results.

### 3.1.2. 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 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 and Rubin, 1992; Stan Development Team, 2014). We deemed the improvements brought by a higher number of iterations negligible, thus we continued using 3000 total iterations, of which 1500 were burn-in.

### 3.1.3. 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 1500 replications

**TABLE 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.**

| Iterations | | Mean | Percentage of $\hat{R}$ above: | | | |
|---|---|---|---|---|---|---|
| Total | Burn-in | $\hat{R}$ | 1.05 | 1.1 | 1.5 | 1.7 |
| 3000 | 1500 | 1.01 | 2.53 | 0.89 | 0.02 | 0.00 |
| 4000 | 2000 | 1.01 | 1.97 | 0.68 | 0.02 | 0.00 |
| 10,000 | 5000 | 1.00 | 1.54 | 0.75 | 0.01 | 0.00 |
| 10,000 | 8000 | 1.01 | 2.13 | 0.70 | 0.01 | 0.00 |
| Final analyses: 3000 iterations with 1500 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 |

for all used methods, being stable for all at 2000 replications. For example, the standard deviations of the estimated mean $\gamma_{0,1}$ or $\phi_n$, depending on estimation method, over replications was lower than 0.01 at 2000 replications for all used estimators. Therefore, the number of replications per condition is set to $R = 2000$. Given that we have 40 conditions, this amounts to $40 \times 2000 = 80,000$ datasets generated.

### 3.1.4. 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.

### 3.2. 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 (6), the statistics obtained with the random and fixed estimators in their respective notations as in Equations (6) and (1).

### 3.2.1. 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.

### 3.2.2. 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:

$$\text{bias} = \left( \frac{1}{R} \sum_{r=1}^{R} \hat{\gamma}_{0,1_r} \right) - \gamma_{0,1}, \qquad (8)$$

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 $\frac{1}{N} \sum_{n=1}^{N} \hat{\phi}_n$.

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 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 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.

### 3.2.3. 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:

$$\text{bias} = \left( \frac{1}{R} \sum_{r=1}^{R} \hat{\sigma}_{U_{1,\,n}} \right) - \sigma_{U_{1,\,n}}. \qquad (9)$$

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**. For the random estimators, the bias for $N = 25$ is smaller than the bias for $N = 10$. The bias is smaller
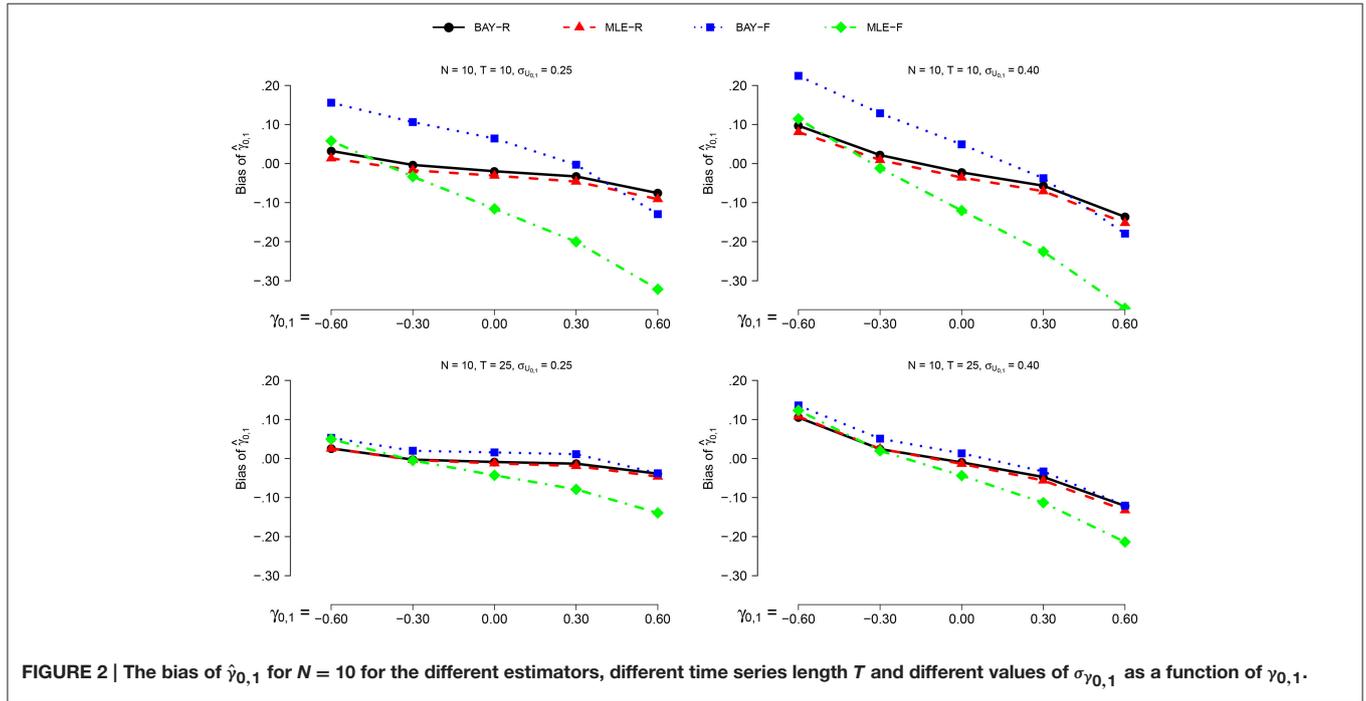


FIGURE 2 | The bias of $\hat{\gamma}_{0,1}$ for $N = 10$ for the different estimators, different time series length $T$ and different values of $\sigma_{\gamma_{0,1}}$ as a function of $\gamma_{0,1}$.
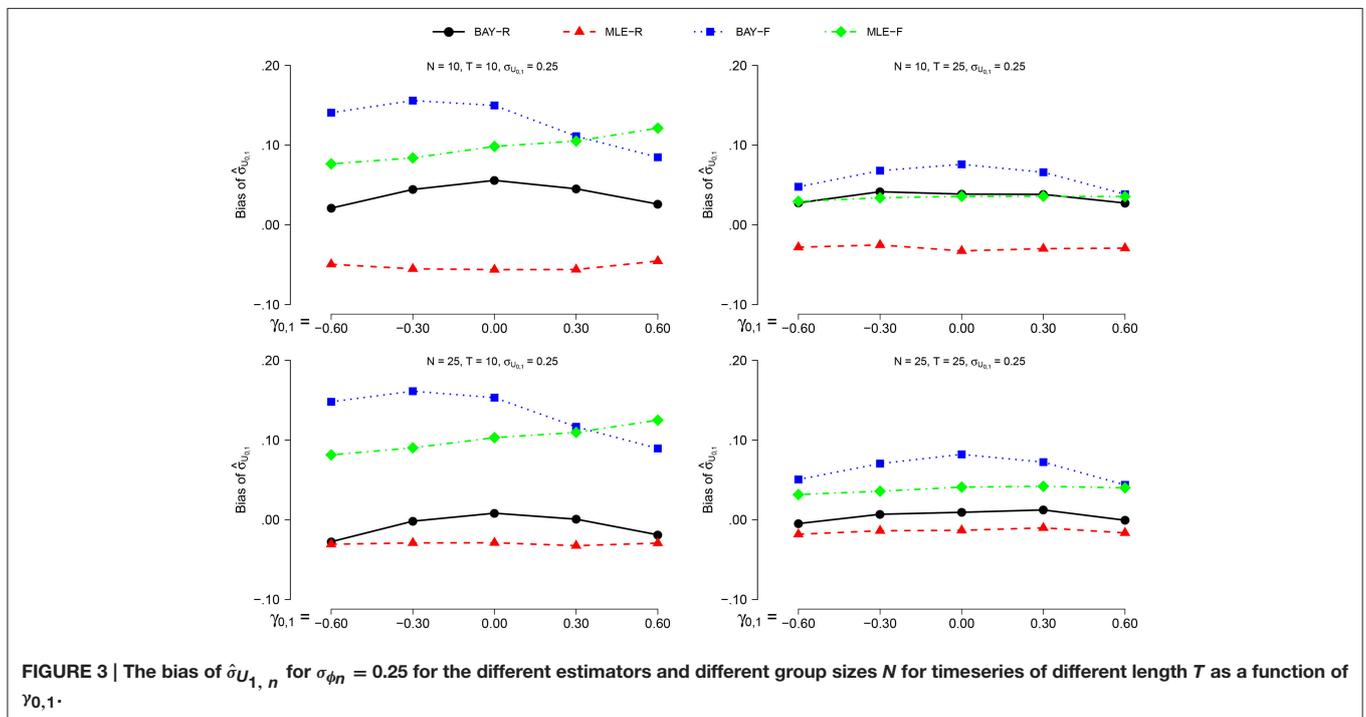


FIGURE 3 | The bias of $\hat{\sigma}_{U_{1,\,n}}$ for $\sigma_{\phi_n} = 0.25$ for the different estimators and different group sizes $N$ for timeseries of different length $T$ as a function of $\gamma_{0,1}$.

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.

### 3.2.4. 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 4**. The $SD(\hat{\gamma}_{0,1})$ is smaller for $N = 25$ compared to $N = 10$, 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.

### 3.2.5. 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:

$$\text{bias of SE}\,(\hat{\gamma}_{0,1}) = \left(\frac{1}{R}\sum_{r=1}^{R} SE(\hat{\gamma}_{0,1_r})\right) - SD(\hat{\gamma}_{0,1}), \qquad (10)$$

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 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 $T$ 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.

### 3.2.6. 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 $\alpha$ in the condition with $\gamma_{0,1} = 0.00$; and the power in all other conditions.

For frequentist methods, testing $H_0 : \gamma_{0,1} = 0$ vs. a two-sided alternative at significance level $\alpha$, 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:

$$\hat{\gamma}_{0,1} \pm t^*_{(1-\alpha);df\,=\,N-2}SE(\hat{\gamma}_{0,1}), \qquad (11)$$

where $SE(\hat{\gamma}_{0,1})$ is obtained as explained in Section 3.2.5. 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 6**, **7**. The actual $\alpha$ 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 $\alpha$ 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 $\alpha$ consistently at or around 0.05, while the MLE-R has an actual $\alpha$ that is too high for $\sigma_{U_{1,\,n}} = 0.40$, namely at 0.10. The fixed estimators have an actual $\alpha$ at or even below 0.01, rather than the desired 0.05.

### 3.2.7. 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 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. 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
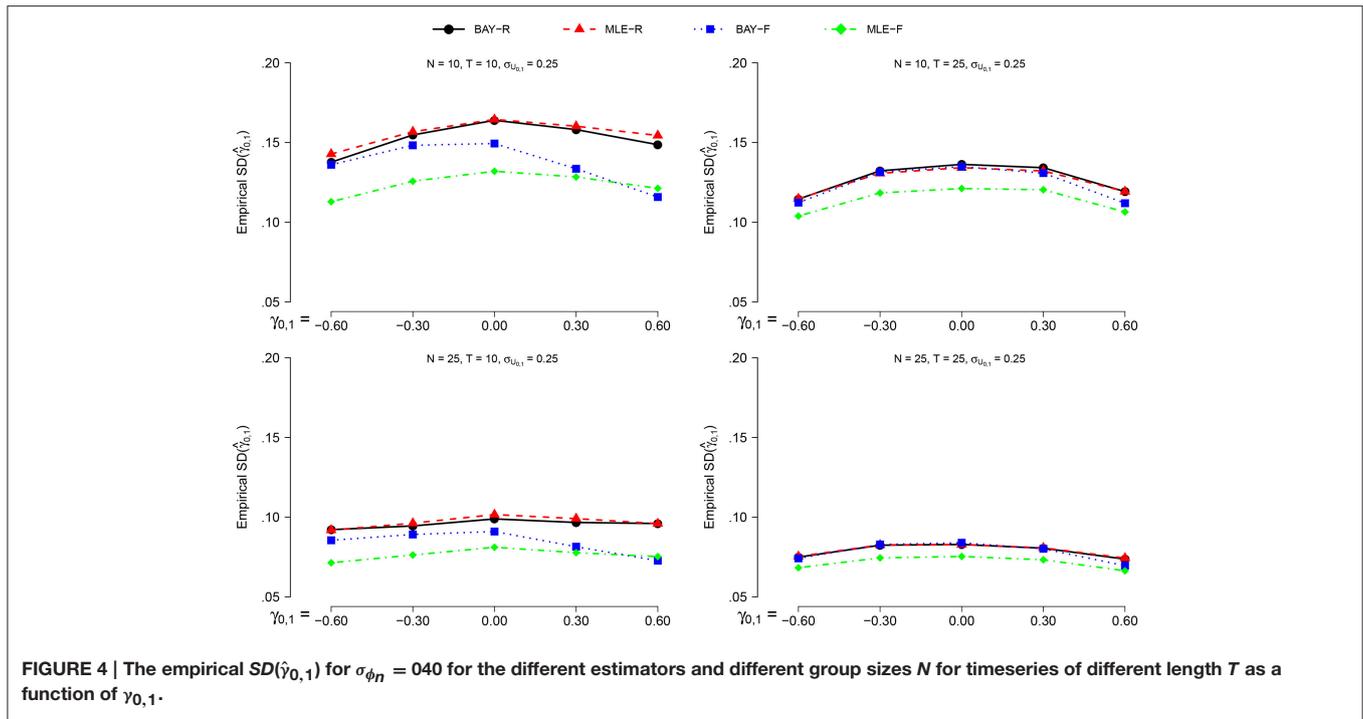
**FIGURE 4 | The empirical $SD(\hat{\gamma}_{0,1})$ for $\sigma_{\phi n} = 040$ for the different estimators and different group sizes $N$ for timeseries of different length $T$ as a function of $\gamma_{0,1}$.**
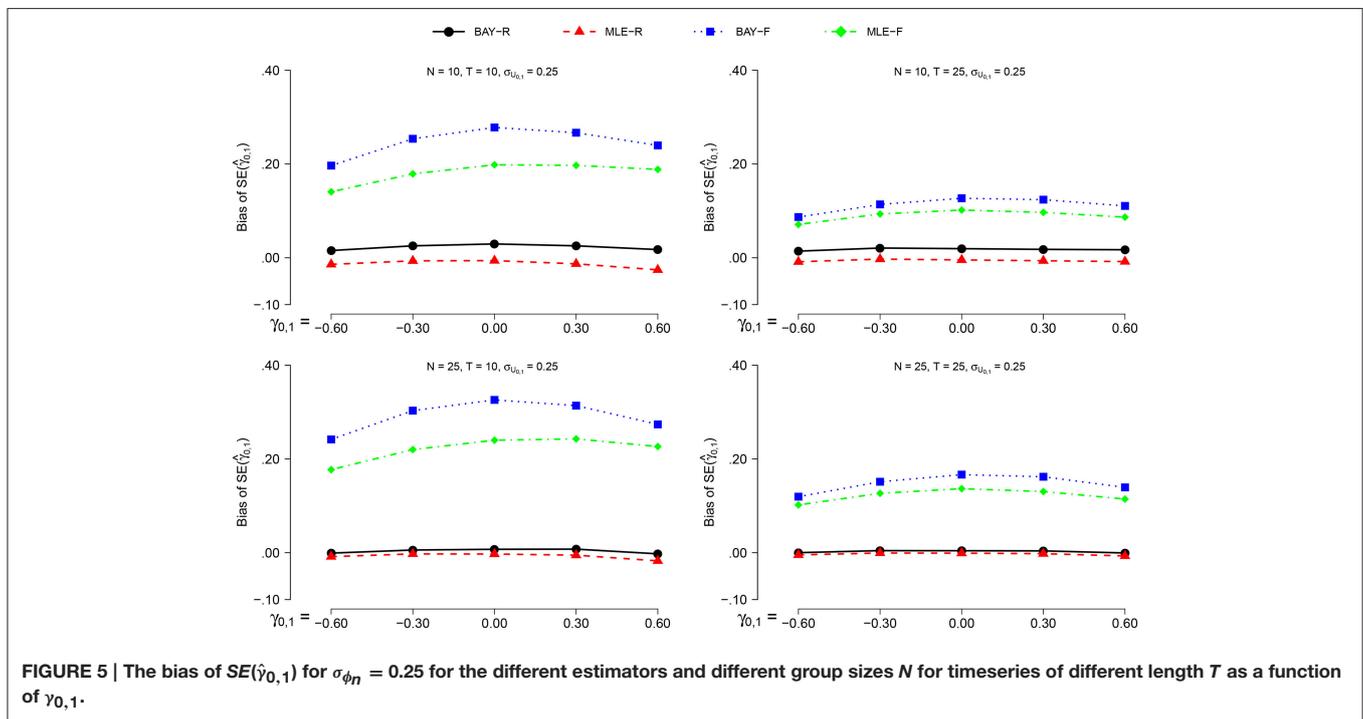


**FIGURE 5 | The bias of $SE(\hat{\gamma}_{0,1})$ for $\sigma_{\phi n} = 0.25$ for the different estimators and different group sizes $N$ for timeseries of different length $T$ as a function of $\gamma_{0,1}$.**

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 $\sigma_{U_{1,n}}$ 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 $SE(\hat{\gamma}_{0,1})$.

## 4. DISCUSSION AND CONCLUSIONS

In this paper 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., 2015). We combined
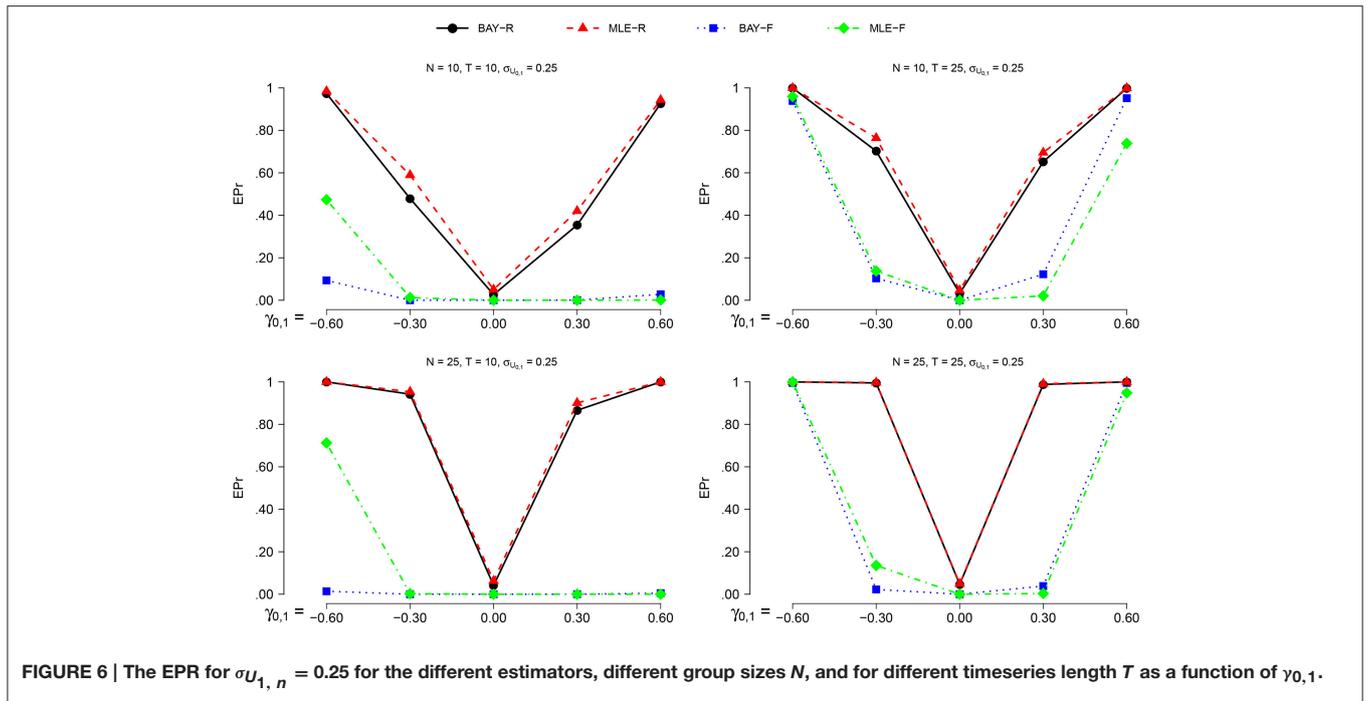
**FIGURE 6 | The EPR for $\sigma_{U_{1,\,n}} = 0.25$ for the different estimators, different group sizes $N$, and for different timeseries length $T$ as a function of $\gamma_{0,1}$.**
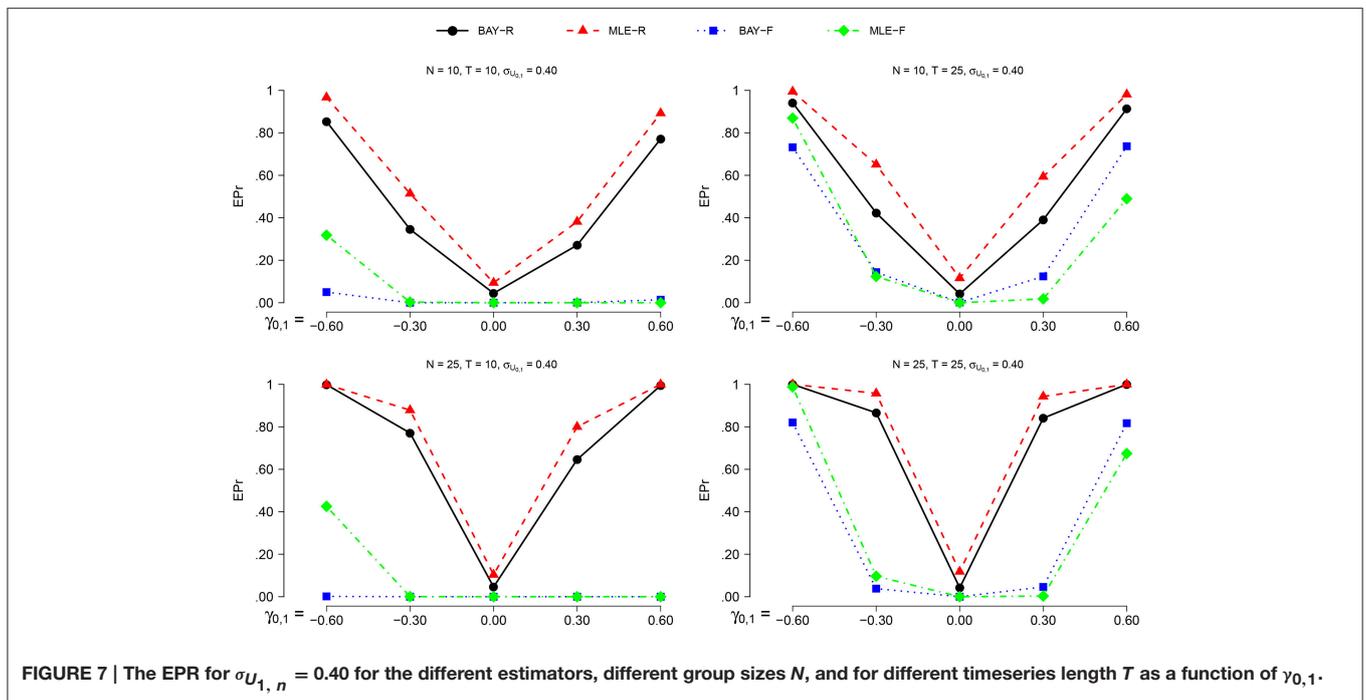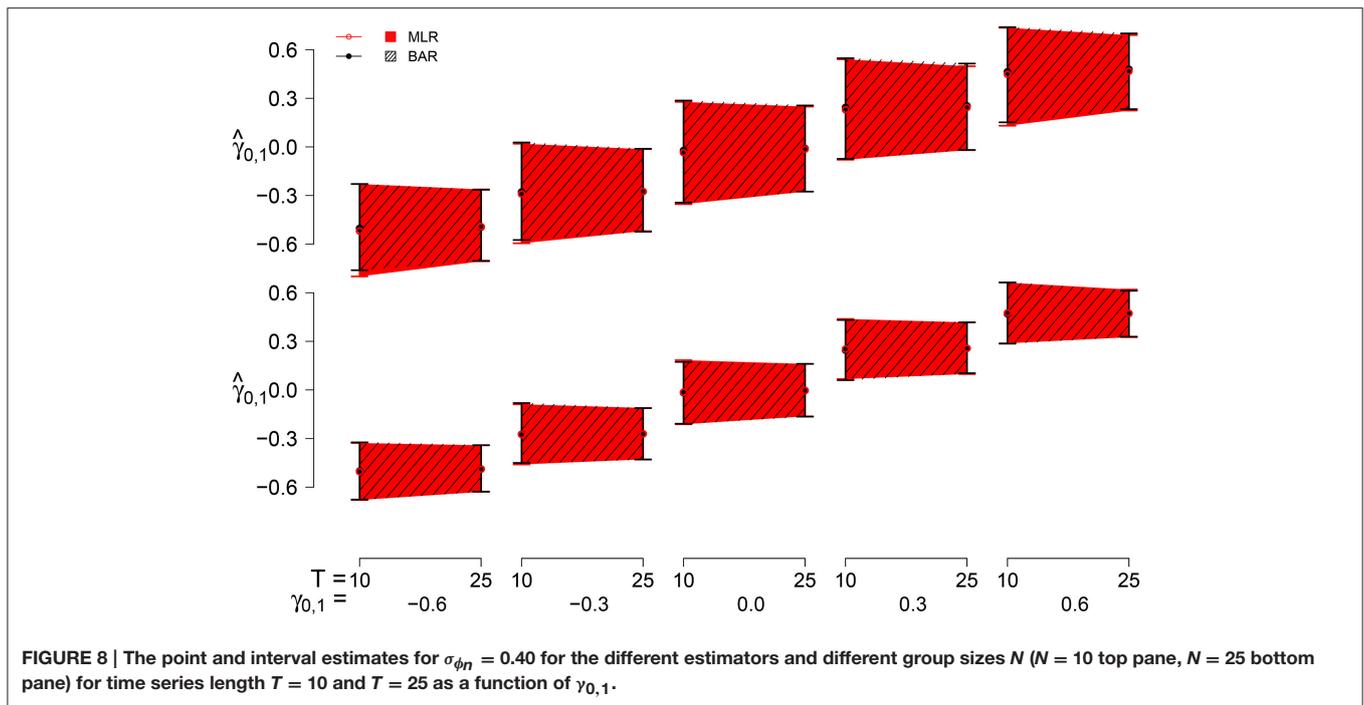


**FIGURE 7 | The EPR for $\sigma_{U_{1,\,n}} = 0.40$ for the different estimators, different group sizes $N$, and for different timeseries length $T$ as a function of $\gamma_{0,1}$.**

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 $\alpha$ as obtained with the fixed estimators,

**FIGURE 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 $\gamma_{0,1}$.**

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 $\alpha$ 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 $\phi_n$ 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 $\phi_n$, $\gamma_{0,1}$, 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 and McKean, 1991; DeCarlo and 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 and Jenkins, 1976; Krone et al., 2015).

The aim of this paper 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 6) is also affected by the auto-correlation. Schuurman et al. (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 paper. 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 and Maekawa, 1984; Kunitomo and Yamamoto, 1985; Krone et al., 2015). 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 and 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 et al., 2011) and data following a Poisson distribution (Terui et al., 2010). However, the theoretical framework to estimate a SSM with any distribution in the exponential family is available (Petris et al., 2009; Durbin and Koopman, 2012). A Bayesian interpretation of the state space model is found in the Bayesian dynamic model (West and 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.

## AUTHOR CONTRIBUTIONS

TK: Main contributing author, CA and MT: co-authors of the study-design, substantial contributions to the manuscript and revisions thereof.

## FUNDING

## REFERENCES

Arnau, J., and Bono, R. (2001). Autocorrelation and bias in short time series: an alternative estimator. *Qual. Quant.* 35, 365–387. doi: 10.1023/A:1012223430234

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

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

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

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

Box, G. E. P., and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. San Francisco, CA: Holden-Day.

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

Bryk, A. S., and Raudenbush, S. W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: SAGE Publications.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* 16, 1190–1208. doi: 10.1137/0916069

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B Methodol.* 39, 1–38.

Durbin, J., and Koopman, S. J. (2012). *Time Series Analysis by State Space Models*. Oxford, UK: Oxford University Press.

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

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

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–472. doi: 10.1214/ss/11770 11136

Goldstein, H. (2011). *Multilevel Statistical Models, 4th Edn*. Hoboken, NJ: Wiley.

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

Hamaker, E. L., Ceulemans, E., Grasman, R. P. P. P., and Tuerlinckx, F. (2015). Modeling affect dynamics: state of the art and future challenges. *Emot. Rev.* 7, 316–322. doi: 10.1177/1754073915590619

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J.*

*Am. Stat. Assoc.* 72, 320–338. doi: 10.1080/01621459.1977.104 80998

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

Hox, J. J. (2010). *Multilevel Analysis*. Quantitative methodology, 2nd Edn. New York, NY: Routledge.

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

Jongerling, J., Laurenceau, J.-P., and Hamaker, E. L. (2015). A multilevel AR(1) model: allowing for inter-individual differences in trait-scores, inertia, and innovation variance. *Multivariate Behav. Res.* 50, 334–349. doi: 10.1080/00273171.2014.1003772

Kashdan, T. B., and 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, 629–638. doi: 10.1037/a0035796

Krone, T., Albers, C. J., and Timmerman, M. E. (2015). A comparative simulation study of AR(1) estimators in short time series. *Qual. Quant*. doi: 10.1007/s11135-015-0290-1. [Epub ahead of print].

Kunitomo, N., and Yamamoto, T. (1985). Properties of predictors in misspecified autoregressive time series models. *J. Am. Stat. Assoc.* 80, 941. doi: 10.1080/01621459.1985.10478208

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

Liang, K.-Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73, 13–22.

Lodewyckx, T., Tuerlinckx, F., Kuppens, P., Allen, N. B., and Sheeber, L. (2011). A hierarchical state space approach to affective dynamics. *J. Math. Psychol.* 55, 68–83. doi: 10.1016/j.jmp.2010.08.004

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087–1092. doi: 10.1063/1.1699114

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

Powell, M. J. D. (2009). *The BOBYQA Algorithm for Bound Constrained Optimization Without Derivatives*. Cambridge, UK: University of Cambridge.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

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

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

Shiffman, S., Stone, A. A., and Hufford, M. R. (2008). Ecological momentary assessment. *Annu. Rev. Clin. Psychol.* 4, 1–32. doi: 10.1146/annurev.clinpsy.3.022806.091415

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

Solanas, A., Manolov, R., and Sierra, V. (2010). Lag-one autocorrelation in short series: estimation and hypotheses testing. *Psicologica* 31, 357–381.

Stan Development Team (2014). *Rstan: The R Interface to Stan, Version 2.4*.

Suls, J., Green, P., and Hills, S. (1998). Emotional reactivity to everyday problems, affective inertia and Neuroticism. *Pers. Soc. Psychol. Bull.* 24, 127–136. doi: 10.1177/0146167298242002

Tanaka, K., and Maekawa, K. (1984). The sampling distributions of the predictor for an autoregressive model under misspecifications. *J. Econ.* 25, 327–351. doi: 10.1016/0304-4076(84)90005-8

Terui, N., Ban, M., and Maki, T. (2010). Finding market structure by sales count dynamics - multivariate structural time series models with hierarchical structure for count data. *Ann. Inst. Stat. Math.* 62, 91–107. doi: 10.1007/s10463-009-0244-2

Walker, G. (1931). On periodicity in series of related terms. *Proc. R. Soc. Lond. A* 131, 518–532.

West, M., and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. New York, NY: Springer Verlag.

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.