Dynamic Structural Equation Models

Tihomir Asparouhov, Ellen L. Hamaker and Bengt Muthén

Version 2

April 27, 2017

Abstract

This paper presents a dynamic structural equation model (DSEM), which can be used to study the evolution of observed and latent variables as well as the structural equation models over time. DSEM is suitable for analyzing intensive longitudinal data (ILD) where observations from multiple individuals are collected at many points in time. The modeling framework encompasses previously published DSEM models and is a comprehensive attempt to combine time series modeling with structural equation modeling. The DSEM model is estimated with Bayesian methods using the MCMC Gibbs sampler and the Metropolis-Hastings sampler. We provide a detailed description of the estimation algorithm as implemented in the Mplus software package. DSEM can be used for longitudinal analysis of any duration and with any number of observations across time. Simulation studies are used to illustrate the framework and study the performance of the estimation method. Methods for evaluating model fit are also discussed. Continuous time modeling, uneven times of observations and subject-specific times of observations are discussed as well.

1 Introduction

In the last several years intensive longitudinal data with many repeated measurements from a large number of individuals have become quite common. These data are often collected using smart phones or other electronic devices

and are referred to as ambulatory assessments (AA), daily diary data, ecological momentary assessment (EMA) data, or experience sampling methods (ESM) data (cf. Trull & Ebner-Priemer, 2013). The accumulation of these types of data naturally leads to an increasing demand for statistical methods that allow us to model the dynamics over time as well as individual differences therein using intensive longitudinal data.

One of the most common methods for longitudinal analysis in the social sciences is growth modeling where an observed or latent variable is modeled as a function of time, for example, a linear function of time. The coefficients of the function, for example, intercept and slope, which determine the trajectory for the variable are subject-specific random effects. Frequently such growth models are expressed as multivariate models especially when the number of observations for each person is small, for example less than 10. Using a multivariate model allows us to introduce additional auto-correlations or add time-specific parameters such as time-specific residual variances. However, a multivariate model is not suitable for modeling longer longitudinal analysis such as 30 or more observations across time for several different reasons. The first reason is that the model can become computationally intensive for longer longitudinal data. A univariate model with 30 observations will require modeling the joint distribution for all 30 observations, i.e., a 30x30 variance covariance matrix. A bivariate model would require 60x60 variance covariance matrix and so on. The dimensions of the joint distribution increase rapidly and can easily become computationally prohibitive.

Longer longitudinal data provides further challenges for standard growth modeling. While it is reasonable to expect that the evolution of a variable across 10 time points can be quite well approximated by a linear, quadratic or cubic curve, it is unlikely that such an approach will be sufficient for 100 times of observations, simply because 4 parameters out of 10 degrees of freedom is much likelier to be approximately correct than 4 out of 100 degrees of freedom. This is also the reason why splines have gained momentum in modeling longer trajectories. However, splines or other smoothing techniques cannot be used for inference about the future as splines have no natural continuation outside of the observed period the way quadratic and linear trends do.

In longer longitudinal data the best predictors of a particular observation would typically be other observations taken around the same time. This is essentially the premise of time series analysis, i.e., allowing observations to be directly regressed on observations from the preceding periods. For longer

longitudinal analysis it will be much harder to predict the value of an observation at a particular time point simply using some global characteristics about the individual rather than to use characteristics about the individual that are relevant to that time period.

An alternative specification of a growth model is as a two-level model where each cluster consists of all the observations for one individual. Using cross-classified modeling this approach can be extended to allow time-specific random effects in addition to subject-specific random effects, see Asparouhov and Muthén (2016). Such an approach can accommodate longitudinal studies of any duration and number of observations. However, it does not accommodate autoregressive modeling where consecutive observations are directly related rather than through subject-specific effects. The framework that we describe here is a direct extension of the cross-classified ILD modeling framework described in Asparouhov and Muthén (2016). We simply add to that framework the ability to regress any variable, observed or latent, not only on other variables from the same time period but also from several of the previous periods.

The DSEM model described here can be viewed as the two-level extension of the dynamic structural models described in Molenaar (1985), Zhang and Nesselroade (2007) and Zhang et al. (2008). Time series models for observed and latent variables date back to Kalman (1960) and are applied extensively in engineering and econometrics. In most such applications, however, multivariate time series data of a single case (i.e., N=1) are analyzed. In contrast, the intensive longitudinal data discussed in this article is for a sample of individuals and the DSEM framework discussed here accommodates this more complex modeling need. Analyzing a random sample of individuals as usual allows us to make inference about individuals that are not in the sample, which is something that can not be done when a single individual is analyzed. Thus the DSEM framework will allow us to make inference for individuals outside of the sample as well as for future observations for individuals in the sample.

The DSEM framework is a powerful tool for exploring intensive longitudinal data as it combines four different modeling techniques: multilevel modeling, time-series modeling, structural equation modeling (SEM) and time varying effects modeling (TVEM). Each of these four techniques addresses different aspects of the data and is used to model different correlations that are found in such data. The multilevel modeling is based on correlations that are due to individual-specific effects. The time series modeling is based on

correlations due to proximity of observations. The SEM modeling is based on correlations between different variables. The TVEM modeling is based on correlations due to the same stage of evolution. The goal of the DSEM framework is to parse out and model these four types of correlations and thereby give us a fuller picture of the dynamics found in the ILD.

The outline of this article is as follows. First we present the general DSEM model and the model estimation using Bayesian methods. Next we discuss methods for model fit evaluation. We then illustrate the framework with multiple simulation studies and conclude with a summary discussion.

2 The general DSEM model

The general DSEM model consists of three separate models. The most general DSEM model is the cross-classified model which incorporates both individual- and time-specific random effects. The second most general model is the two-level DSEM model which incorporates individual-specific random effects only. This model could actually be the most popular and useful model as it is easier to estimate, identify and interpret. The third model is the single-level DSEM model for N=1, that is, a DSEM model-estimated with the time series data from a single individual, see Zhang and Nesselroade (2007). There are no random effects in the single-level model, that is, all model parameters are non-random. Here we describe the most general cross-classified DSEM model. The two-level and the single-level DSEM models are special cases of the cross-classified DSEM model.

Let Y_{it} be a vector of measurements for individual i at time t, where the i-th individual is observed at times $t = 1, 2, ..., T_i$. The cross-classified DSEM model of lag L begins with the following decomposition

$$Y_{it} = Y_{1,it} + Y_{2,i} + Y_{3,t}, \tag{1}$$

where $Y_{2,i}$ and $Y_{3,t}$ are individual-specific and time-specific contributions and $Y_{1,it}$ is the deviation of individual i at time t. All three components are latent normally distributed random vectors and are used to form three sets of structural equations - one on each level.

2.1 The between-level models

The second and the third level structural equation models take the usual form

$$Y_{2,i} = \nu_2 + \Lambda_2 \eta_{2,i} + K_2 X_{2,i} + \varepsilon_{2,i} \tag{2}$$

$$\eta_{2,i} = \alpha_2 + B_2 \eta_{2,i} + \Gamma_2 X_{2,i} + \xi_{2,i} \tag{3}$$

$$Y_{3,t} = \nu_3 + \Lambda_3 \eta_{3,t} + K_3 X_{3,i} + \varepsilon_{3,t} \tag{4}$$

$$\eta_{3,t} = \alpha_3 + B_3 \eta_{3,t} + \Gamma_3 X_t + \xi_{3,t}. \tag{5}$$

On each level the first equation is generally referred to as the measurement equation and the second equation is referred to as the structural equation. The vector $x_{2,i}$ is a vector of individual-specific time-invariant covariates and $x_{3,t}$ is a vector of time-specific but individual-invariant covariates. Similarly, $\eta_{2,i}$ is a vector of individual-specific time-invariant latent variables and $\eta_{3,t}$ is a vector of time-specific individual-invariant latent variables. The variables $\varepsilon_{2,i}, \xi_{2,i}, \varepsilon_{3,t}, \xi_{3,t}$ are zero mean residuals as usual and the remaining vectors and matrices in the above equation are non-random model parameters.

While the above equations do not include regressions among Y components such regressions are typically achieved by creating a latent variable equal to the Y variable, that is, the Y variable would be a perfect error-free indicator for a latent variable. Once such latent variables are included in the model the regression between the Y variables is specified as a regression between the corresponding latent variables using the structural equations. This is a simple way to reduce the number of matrices in the above equations and is somewhat of a tradition in the structural equation modeling literature, but it has no implication to model specification or estimation.

In the above specification we did not include level 2 and level 3 dependent variables but such variables are easy to accommodate as well. The vectors $Y_{2,i}$ and $Y_{3,t}$ can include not just the latent decomposition parts of the variables Y_{it} but can also include observed variables that are subject-specific or time-specific.

2.2 The within-level model

The within-level part of the DSEM model is described by the following equations which now include time-series components

$$Y_{1,it} = \nu_1 + \sum_{l=0}^{L} \Lambda_{1,l}\eta_{1,i,t-l} + \sum_{l=0}^{L} R_l Y_{1,i,t-l} + \sum_{l=0}^{L} K_{1,l} X_{1,i,t-l} + \varepsilon_{1,it} \quad (6)$$

$$\eta_{1,it} = \alpha_1 + \sum_{l=0}^{L} B_{1,l}\eta_{1,i,t-l} + \sum_{l=0}^{L} Q_l Y_{1,i,t-l} + \sum_{l=0}^{L} \Gamma_{1,l} X_{1,i,t-l} + \xi_{1,it}. \quad (7)$$

Here $x_{1,it}$ is a vector of observed covariates for individual i at time t and $\eta_{1,it}$ is a vector of latent variables for individual i at time t.

In the above equations the latent variables η, the dependent variables Y and the covariates X at times t, $t-1$, ..., $t-L$ can be used to predict the latent variables η and the dependent variables Y at time t. Including the lagged predictors X in the above equations is somewhat inconsequential, and we do this mostly for completeness. The covariate $X_{1,i,t-l}$ is not any different from the covariate $X_{1,i,t}$ because the model does not include distributional assumptions about the covariates X and is essentially a model for the conditional distribution of $[Y|X]$. Including the lagged covariates $X_{1,i,t-l}$ doesn't involve any special statistical consideration with one small exception of the initial unobserved values which we will address later.

Latent centering

The dependent variables Y, on the left and the right hand side of the above equations are not the actual observed quantities Y_{it} but rather the within-level component $Y_{1,it}$. These are sometimes referred to as the centered variables since $Y_{1,it} = Y_{it} - Y_{2,i} - Y_{3,t}$. The variables $Y_{2,i}$ and $Y_{3,t}$ can be interpreted as the mean for individual i and the mean for time t which are thus subtracted to form the pure realization for individual i at time t excluding any global effects specific for individual i and time t.

Centering is inconsequential for the variables on the left hand side of the equations but is important for the variables on the right hand side of the equations and is well established in the multilevel modeling literature, see Raudenbush and Bryk (2002). In principle one can use the corresponding observed sample means instead of the latent true means $Y_{2,i}$ and $Y_{3,t}$, however, that will produce biased estimates because the sample mean is different from

the true mean and has a sampling error which will be unaccounted for. In multilevel models this has been documented in Ludke et al. (2008) where the bias is shown to occur for the between-level estimates when the regression involves two separate variables. In time series models the bias has been documented in Nickell (1981) and Hamaker and Grasman (2015) where the bias occurs on the within-level and involves just one variable that is regressed on itself at the preceding time. In both cases the bias disappears as the cluster size increases and the difference between true mean and sample mean vanish.

Random slopes and loadings

In addition to the above equations we allow random slopes and loadings on the within-level. Every structural coefficient on the within-level can be a non-random model parameter or it can be a random parameter. Every within-level random parameter s can be decomposed as follows

$$s = s_{2,i} + s_{3,t}, \tag{8}$$

where $s_{2,i}$ is an individual-specific random effect, that is, an individual-specific latent variable which is an element of the vector $\eta_{2,i}$ modeled in the level 2 structural model. Similarly, $s_{3,t}$ is a time-specific random effect, i.e., a time-specific latent variable which is a part of the vector $\eta_{3,t}$ modeled in the level 3 structural model. An alternative way to present this model is to directly introduce the indices i and t in the structural parameters in Equations (6) and (7) as follows

$$Y_{1,it} = \nu_1 + \sum_{l=0}^{L} \Lambda_{1,lit}\eta_{1,i,t-l} + \sum_{l=0}^{L} R_{lit}Y_{1,i,t-l} + \sum_{l=0}^{L} K_{1,lit}X_{1,i,t-l} + \varepsilon_{1,it} \tag{9}$$

$$\eta_{1,it} = \alpha_{1,it} + \sum_{l=0}^{L} B_{1,lit}\eta_{1,i,t-l} + \sum_{l=0}^{L} Q_{lit}Y_{1,i,t-l} + \sum_{l=0}^{L} \Gamma_{1,lit}X_{1,i,t-l} + \xi_{1,it} \tag{10}$$

with the additional specification that every parameter varying with i and t is decomposed as in Equation (8).

Random residual variances

In addition to the above random effects we allow residual variances V on the within-level to be random parameters, i.e., the model parameters $Var(\varepsilon_{1,it})$ and $Var(\xi_{1,it})$ can be random as follows

$$V = Exp(s_{2,i} + s_{3,t}), \tag{11}$$

where $s_{2,i}$ is an individual-specific normally distributed random effect and $s_{3,t}$ is a time-specific normally distributed random effect and again these random effects are elements of the higher level latent variable vectors $\eta_{2,i}$ and $\eta_{3,t}$.

Note that random structural parameters such as loadings and slopes have a normal distribution, that is, are normally distributed random effects, while random residual variance parameters have a log-normal distribution. This is necessary to ensure that the variance parameters remain positive during the MCMC estimation. Note also that this random variance approach applies only to univariate variance parameters and it does not include random multivariate variance covariance matrices.

It is somewhat more difficult to construct random positive definite variance covariance matrices, based on random effects that could also be used in linear models such as (3) and (5) and remain positive definite for any individual and any set of covariates while at the same time be easy to interpret. However, it is possible to construct random variance covariance matrices by introducing factors with random variances or via random loadings Cholesky decomposition. Both of these approaches are somewhat more complex not just in implementation but also in interpretation as well.

Including moving-average terms

The DSEM model incorporates only the auto-regressive modeling as a time-series feature, but can easily accommodate the moving average modeling because it includes latent variable modeling. Consider for example the ARMA(1,1) model

$$Y_t = \mu + aY_{t-1} + \eta_t + b\eta_{t-1}. \tag{12}$$

The moving average part of this model is nothing more than a latent variable and its lagged 1 latent variable predicting the dependent variable Y. Thus the ARMA models are a special case of the DSEM model. Similarly accommodating ARIMA models amounts to fixing the regression coefficients of Y_t on Y_{t-l} to $(-1)^{l+1}\binom{m}{l}$, where m is the degree of integration. For example, fixing parameter a to 1 in Equation (12) yields the ARIMA(0,1,1) model.

Starting up the process

One final issue that should be specified in the above model is the fact that the variables $Y_{1,i,t-l}$, $X_{1,i,t-l}$ and $\eta_{1,i,t-l}$ can have a time index that is zero or negative in the above model. For example when $t \leq l$ the time index $t-l \leq 0$ appears in Equations (6) and (7). Such variables never appear in the model as dependent variables and thus we have to provide a specification of some kind. In this treatment we have chosen a method that is similar to the one used in Zhang and Nesselroade (2007). We treat all of these variables as auxiliary parameters that have their own prior. Such a prior could be difficult to specify in practical settings, however, and thus we propose the following method which estimates the prior during a burnin phase of the MCMC estimation. In the first iteration all the variables with non-positive time index are set to zero. After each MCMC iteration during the burnin phase of the estimation a new prior is computed as follows. The prior for $Y_{1,it}$ for $t \leq 0$ is set to be the normal prior with mean and variance the sample mean and variance of $Y_{1,it}$ over all $t > 0$ values. Similarly the prior is set for $X_{1,it}$ and $\eta_{1,it}$ for $t \leq 0$. Note that none of the burnin iterations are used to construct the final posterior distribution of the parameters. This is essential in order to preserve the integrity of the MCMC sequence. This method is easy to use and appears to be quite well tuned. It is the default option in Mplus and is based on 100 burnin iterations. Note that when the time series model is sufficiently large with 30 or more observations it is very unlikely that the prior specification affects the estimation. The effect of this prior tends to fade away beyond the first few time periods. However, when the number of time periods in the time series is small such as less than 20 one can expect that the prior will have some small effect on the estimates. The burnin phase prior estimation method we propose here appears to be working quite well even for short time series.

2.3 Categorical variables

Categorical variables can easily be accommodated in the above model through the probit link function. For each categorical variable Y_{ijt} in the model, $j = 1, ..., p$, taking the values from 1 to m_j, we assume that there is a normally distributed latent variable Y^*_{ijt} and threshold parameters $\tau_{1j}, ..., \tau_{m_j-1j}$ such that

$$Y_{ijt} = m \Leftrightarrow \tau_{m-1j} \leq Y^*_{ijt} < \tau_{mj}, \tag{13}$$

where $\tau_{0j} = -\infty$ and $\tau_{m_j j} = \infty$. The above definition essentially converts a categorical variable Y_{ijt} into an unobserved continuous variable Y^*_{ijt}. The model is then defined using Y^*_{ijt} instead of Y_{ijt} in Equation (1). Note that the τ parameters are non-random parameters while the random intercept parameters $Y_{2,ij}$ and $Y_{3,jt}$ provide a random and uniform shift for these threshold parameters, i.e., a certain degree of uniformity is assumed when the variable is ordered polytomous. Such an assumption does not exist for binary variables and when the variable is binary, depending on the structural model at hand the single threshold parameter can be replaced by a mean parameter for Y^*_{ijt}. This kind of parametrization yields a more efficient estimation by avoiding the slow mixing of Cowles (1996) algorithm for sampling thresholds, see also Asparouhov and Muthén (2010). This is also the reason why sometimes models with binary variables tend to be easier to estimate as compared to models with ordered polytomous variable with more than one category, despite the fact that ordered polytomous variables carry more information in the analysis and more information generally means better model identification and more precise estimation.

Note that while the categorical variable modeling given in Equation (13) relies on the underlying continuous variable Y^*_{ijt} the actual model application does not require such an interpretation. The variables Y^*_{ijt} are just a convenience for formulating the model. The model can equivalently be formulated without the underlying continuous variables Y^*_{ijt} and directly on the model-implied discrete probabilities $P(Y_{ijt} = m)$ using the model-implied probit regression. Thus even when underlying continuous variable is deemed an unacceptable concept from a substantive point of view the above model is applicable.

2.4 Continuous time dynamic modeling

The DSEM model as described this far applies to situations where the time variable can be scaled so that each person is observed at times $1, 2, ..., T_i$. This assumption is reasonable for example in daily diary applications where each subject is observed once a day. However, it is unrealistic in other applications where multiple observations are taken per day or in situations where observations are so dispersed that a daily scale is unrealistic. In many cases individual observations are taken at uneven time intervals or at random. The times of observations could be considered real values rather than integer values which would call for continuous time modeling.

We can resolve this problem by resetting the time variable using scaling, shifting, and rounding so that the continuous times of observations are well approximated by integer values. In its essence this process amounts to the following. Using a small value δ we divide the time line using an equally spaced grid where δ represents the length of the grid intervals. The times of observations are rounded to the nearest grid time point which thereby converts the continuous times of observations to integer times of observations. We then fill in the data with missing values for those integers that were not the nearest for an observed continuous time point. The complete details of the algorithm implemented in Mplus are given in Appendix A. Understanding the process of discretization is also important for a proper interpretation of the DSEM results.

2.5 Final remarks on the general DSEM model

For identification purposes, restrictions need to be imposed on the above general model. For example, mean structure parameters can exist only on one of the levels for most common situations, that is, ν_j will be fixed to 0 on 2 out of the 3 levels. Other identifying restrictions need to be imposed along the lines of standard structural equation models.

The above model is the time-series generalization of the time intensive model described in Section 8.3 of Asparouhov and Muthén (2016). The remarkable and daring features of this model are that longitudinal data of any length is allowed, an unlimited number of random effects can be estimated without a substantial computational burden, and that no two observations in the data are truly independent of each other, as the time series and subject-specific random effects correlate data within each subjects and the time-specific effects correlate data across subject.

The DSEM model is a two-level model, but because it is a multivariate model, it can be used to formulate three-level DSEM models where the first level is written in a multivariate wide format. This is particularly the case when the first level contains only a small number of observations. One such example is described in Jahng et al. (2008) where the three level structure is as follows: subjects, days, and observations within days. The number of observations within a day is typically a number smaller than 10 and thus can be represented with a 10 dimensional vector. Using this approach it is possible to model within-day autocorrelation structures and between-days autocorrelation structures, that is, construct three-level DSEM models.

The DSEM model estimation is implemented in Mplus Version 8, with three notable exceptions that may be resolved in future Mplus implementations. The three exceptions are as follows: a) the parameters R_l and Q_l can not be random for when $l = 0$; b) the parameters $\Lambda_{1,l}$, $B_{1,l}$ and the parameter in (11) can be random, but can not include a time-specific random effect; and c) for categorical variables the lagged variables $Y^*_{ij,t-l}$ are not a part of the model, that is, for categorical variables time series models can be built only through latent variables or other continuous dependent or independent variables.

In conclusion, the *cross-classified DSEM model* presented above allows us to study the evolution across time not just of the observed and latent variables but also of the structural model as well. The *two-level DSEM model* is a special case of the cross-classified DSEM model and eliminates $Y_{3,t}$, $X_{3,t}$ and $\eta_{3,t}$ variables from the model as well as Equations (4) and (5). In (1) the component $Y_{3,t}$ is eliminated and thus the main decomposition is the usual within/between decomposition that is fundamental to two-level structural equation models. The structural model is assumed to be time-invariant. However, this does not imply that the variable distribution is time invariant: Time-varying covariates, including the time variable t itself, can still be included in the model, and thus trends and growth models can be estimated in addition to the subject-specific time series models.

Furthermore, the *single-level DSEM model* is a special case of the two-level DSEM model and it essentially contains just one cluster and no random effects. Equation (1) reduces to $Y_{1,it} = Y_{it}$ and the variables $Y_{2,i}$, $X_{2,i}$ and $\eta_{2,i}$ are removed from the model as well as Equations (2) and (3). In fact, the model is completely specified only by Equations (6) and (7). Since we have just one cluster/individual in the model, the index i can be removed from the model.

Finally, note that the cross-classified DSEM model requires the time scale to be aligned across all individuals so that a time-specific effect $s_{3,t}$ has the same meaning for all individuals at time t. Not every ILD set is suitable for the cross-classified DSEM model. Consider for example an observational study in which time t is simply the time since the first observation was recorded; in this case, no particular effect may be expected at time t that applies to every subject in the study. On the other hand, if the study was on subjects that enrolled in a treatment and time t represents the time since enrollment in the treatment, it is natural to expect that time-specific effects at time t can exist and apply to all subjects in the data; in that case, the

cross-classified DSEM model can be used. Note that the two-level DSEM model has no particular requirements on the time scale and is thus suitable for any ILD analysis.

3 Model Estimation

The model estimation without the time-series features is described in Asparouhov and Muthén (2016) and Asparouhov and Muthén (2010). A substantial portion of that estimation algorithm also applies directly to the estimation of the DSEM model. We will summarize the general framework briefly and then we will provide details on the estimation that are specific to DSEM. The details that are not provided here can be found in Asparouhov and Muthén (2010) and are related to Bayesian estimation of SEM. Alternatively, these details can be found in Arminger and Muthén (1998) or Lee (2007).

The estimation is based on the MCMC algorithm via the Gibbs sampler. All model parameters, latent variables, random effects, between-level 2 components, between-level 3 components, and missing data are arranged into blocks. Each of these blocks is updated (new value is being generated) from the conditional distribution of that block, conditional on all other remaining blocks and the data. This process is repeated until a stable posterior distribution for all blocks is obtained. The goal of the block arrangement is to assure that each block has an explicit or manageable conditional distribution. In addition, the blocks are arranged in such a way that elements that are highly correlated are generated simultaneously as to improve the quality of the MCMC mixing. To achieve that we arrange the blocks to be as large as possible, while keeping the conditional distributions explicit. Then within each block we arrange the elements into the smallest possible sub-blocks that are conditionally independent and can be generated separately.

The MCMC estimation, unlike ML estimation, has the ability to absorb new modeling features easily, meaning that the estimation would not change dramatically when a new model feature is added. This is because the MCMC estimation is based on many conditional distributions rather than one joint distribution. Thus when a new feature is added to the model not all conditional distributions are affected. As an example consider the conditional distribution of the underlying Y_{it}^* for a categorical dependent variable. Computing the conditional distribution of Y_{it}^* can be done by the same method

we would apply without the time-series features of the model. Similarly the methodology for updating the threshold parameters is not changed by the time-series features of the model.

Let θ represent all non-random model parameters. We split θ in 3 blocks: intercepts, slope and loading parameters θ_1; variance, covariance and correlation parameters θ_2; and threshold parameters θ_3. Priors for each of these parameters have to be specified. Proper, improper, and informative conjugate prior specification for the various parameters are discussed in Asparouhov and Muthén (2010). Here we generally assume non-informative priors for all the parameters but informative priors can be facilitated as well in the MCMC estimation.

All unknown quantities in the DSEM model are placed in the following 13 blocks which are updated one at a time during the MCMC estimation

- B1: $Y_{2,i}$
- B2: All random slopes $s_{2,i}$
- B3: $Y_{3,t}$
- B4: All random slopes $s_{3,t}$
- B5: Other latent variables $\eta_{2,i}$ and $\eta_{3,t}$
- B6: Latent variables $\eta_{1,it}$, including initial conditions where $t \leq 0$
- B7: Missing variables Y_{it}
- B8: Initial conditions $Y_{1,it}$ and $X_{1,it}$ for $t \leq 0$
- B9: Threshold parameters for all categorical variables θ_3
- B10: Underlying variables Y^*_{it} for all categorical variables
- B11: Non-random intercepts, slope and loadings parameters θ_1
- B12: Non-random variance, covariance and correlation parameters θ_2
- B13: Random variance parameters

In certain cases, blocks can be combined to improve mixing quality and the speed of the computation. For example, if R_l and Q_l are non-random parameters, blocks B1 and B2 can be combined and blocks B3 and B4 can be combined. That is because the joint conditional distribution of B1 and B2 is normal. It is not normal if R_l and Q_l are random because Equation (9) will contain the product of elements of B1 and elements of B2. Similar logic applies to B3 and B4.

To complete the description of the MCMC estimation, the conditional distribution of each of the above blocks, conditional on all other blocks and the data, should be specified. The technical details of deriving these conditional distributions are given in Appendix B.

4 Model fit and model comparison

The easiest method for model comparison in the DSEM framework is to evaluate significance of individual parameters through the credibility intervals produced by the Bayesian estimation. This is particularly effective when models are nested and model comparison is essentially a test of significance of effects. However, in more complicated model comparisons such significance testing is not available. This section discusses the DIC and comparisons of sample and estimated quantities as methods for evaluating model fit and model comparison.

4.1 DIC

A commonly used criterion for model comparison in Bayesian analysis is the DIC that was first introduced by Spiegelhalter et al. (2002). The DIC can be computed when all the dependent variables are continuous using the usual formulas. The deviance is computed as

$$D(\theta) = -2\log(p(Y|\theta)), \tag{14}$$

where θ represents all model parameters and Y represents all observed dependent variables. The effective number of parameters p_D is computed as follows

$$p_D = \bar{D} - D(\bar{\theta}), \tag{15}$$

where $\bar{D}$ represents the average deviance across the MCMC iterations and $\bar{\theta}$ represents the average model parameters across the MCMC iterations. The

DIC criterion is then computed as

$$DIC = p_D + \bar{D}. \tag{16}$$

DIC can be used to compare any number of competing models, and these may be nested or not. The best model is the model with the lowest DIC value. The effective number of parameters p_D should generally be close to the size of the vector θ and is the penalty for model complexity of this information criterion.

Comparability of the DIC

Despite this seemingly clear definition, there is substantial variation in how the DIC is actually computed and defined (cf. Celeux, Forbes, Robert & Titterington, 2006). The source of the variation is the definition of θ, and in particular in depends on whether latent variables are treated as parameters or not. If a latent variable is treated as a parameter, it is a part of the vector θ and the likelihood used in the definition of the deviance is conditional on that latent variable. If a latent variable is not treated as a parameter, it is not a part of the vector θ and the likelihood used in the definition of the deviance is the marginal likelihood, that is, the latent variable has to be integrated out (see for a similar discussion in the context of the AIC: Vaida & Blanchard, 2005).

Consider for example a one-factor analysis model. If the factor is treated as a parameter, $p(Y|\theta)$ is the likelihood conditional on the factor where all indicators are independent of each other conditional on that factor. If the factor is not treated as a parameter, $p(Y|\theta)$ is computed without conditioning on the factor and instead using the model-implied variance covariance matrix where the indicators are not independent. These two different ways of computing the DIC will naturally produce different p_D and naturally will be on a completely different scale and incomparable, despite the fact that the model is the same. In more complicated models even more variation can occur as some latent variables can be included as parameters as some may not. This phenomenon makes the DIC somewhat trickier to use in latent variable rich DSEM models as one has to always check that the definitions of DIC are comparable.

Consider a different example that consist of 3 models. Model 1 is a two-indicator one-factor model example where we treat the factor as a parameter. Model 2 is the same as Model 1, but the variance of the factor is fixed to

zero, which is equivalent to the model of two independent indicator variables. Model 3 is the model of two correlated indicators without any factors. In this example the two different formulations of Model 2 yield the same DIC. Thus Model 2 DIC is comparable to Model 1 DIC. Model 2 DIC is also comparable to Model 3 DIC. However, Model 1 DIC is not comparable to Model 3 DIC (despite the fact that they are the same model), that is, model comparability is not transitive.

Stability of the DIC estimate

An additional complication that arises in the computation of DIC for the DSEM model is that when latent variables are treated as parameters the number of parameters p_D becomes so large and so many parameters have to be integrated through the MCMC iterations that the DIC precision is difficult to achieve. It is not unusual that convergence for the model parameters is easily achieved but stable DIC estimate require many more iterations, and there may be cases where it is practically infeasible to obtain a stable DIC estimate. In such cases, the imprecision that remains may be bigger than the DIC difference in the models we are trying to compare.

Therefore, we recommend to verifying that the DIC estimate has converged by running the MCMC estimation with different random seeds for the same model, and comparing the DIC estimates across the different runs to evaluate the precision of the DIC. Despite all these difficulties the DIC is the most practical way to compare models when simple parameter significance tests are not enough.

Formal definition of the DIC for the DSEM model

The definition of the DIC consists of the list of latent variables that are treated as parameters. As in Asparouhov and Muthén (2016), for DIC with two-level and cross-classified models all random effect variables such as random loadings, random slopes, random variances as well as the random intercept variables $Y_{2,i}$ and $Y_{3,t}$ are treated as parameters. In addition, any latent variable on the within-level that is lagged in a time series model is treated as a parameter, that is, any latent variable $\eta_{1,i,t}$ that is also used on the right hand side of Equations (6) and (7) in its lagged version $\eta_{1,i,t-l}$ is treated as a parameter. Clearly, this increases the number of parameters p_D of the DIC substantially, usually much more so than the between-level random effects.

A between-level random effect increases p_D by N while a within-level lagged latent variable increases p_D by $N \cdot T$. Similarly, the missing values for Y_{it} for every dependent variable that is lagged, meaning it is used on the right hand side of Equations (6) and (7) in lagged form, is also treated as a parameter.

Given that these variables are conditioned on, the variables Y_{it} are independent across time and persons, and the likelihood is computed as follows

$$log(P(Y|\theta)) = \sum_{i,t} log(P(Y_{it}|\theta)), \tag{17}$$

where $P(Y_{it}|\theta)$ is the likelihood for a single-level SEM model for individual i at time t. Thus, treating the lagged latent variables on the within-level and the lagged missing data as model parameters makes the computation of the DIC feasible.

To summarize, the DIC can be used to compare two or more DSEM models if the list of latent variables that are treated as parameters is the same, and it is provided as standard output when doing DSEM.

4.2 Comparing sample statistics and their corresponding model-estimated quantities

Another array of possibilities for evaluating model fit is to compare sample statistics and their corresponding model-estimated quantities. This is particularly effective for the two-level DSEM model. Let μ_i be the model-estimated mean for a single dependent variable Y for subject i. Let $\overline{Y_{i*}}$ be the sample mean for subject i, i.e., $\overline{Y_{i*}} = \sum_{t=1}^{T_i} Y_{it}/T_i$. From these two quantities we can compute the following statistics of model fit

$$R = Cor(\mu_i, \overline{Y_{i*}}) \tag{18}$$

$$MSE = \sum_{i=1}^{N} (\mu_i - \overline{Y_{i*}})^2/N. \tag{19}$$

Here R is the correlation between estimated and observed means across the clusters/individuals and MSE is the mean squared error of the estimated versus observed mean. If we compare two competing models, we want to select the model with smaller MSE and higher R as it will better represent the data. Note that such model fit evaluation is useful not just for two-level DSEM models but also for general two-level models.

It is important to realize that the above comparison is most reliable under the condition that there is no missing data. When data is missing and is missing at random (MAR) rather than completely at random (MCAR), the sample quantities $\overline{Y_{i*}}$ will not necessarily be the mean of Y in cluster i and the model-estimated μ_i could be the more accurate estimate for that mean. Cautious inference in the presence of missing data can still be made using R and MSE. However, undeniably these statistics are not as reliable as in the case of no missing data and discrepancy between model-estimated values and sample values could simply be the result of MAR and not MCAR missing data.

Note also that R and MSE can be computed for any observed model variable and statistic. For example, instead of the mean of Y we can compute the sample and model-estimated variance of Y. Another example is the covariance between two dependent variables $Cov(Y_1, Y_2)$, that is, computing the correlation R between the cluster-specific model-estimated covariance and the cluster-specific sample covariance. Yet another example that is particularly of interest for the two-level DSEM model is to compute the correlation R between the subject-specific sample autocorrelation for a variable Y and the subject-specific model estimated autocorrelation of Y across the subjects.

Because there are many variables and many different statistics, one can expect that the R and MSE statistics can potentially disagree about which model represents the data better. Empirical data applications will yield more insight on that topic and whether such a disagreement is common. Note also that the DSEM model offers many more subject-specific estimated quantities than the standard two-level SEM model without any random structural coefficients. For example, in the standard two-level SEM, estimated variances are not subject-specific so R would be zero for all models and no model comparison can be performed that way.

In Mplus the correlations R can be obtained for the means and the variance statistics within the Mplus between-level scatter plots simply by plotting the estimated gainst the sample quantities. The MSE is not reported in those plots but can easily be computed by saving the data of the plots and computing it in a separate step. In the Mplus residual output other estimated statistics, such as covariance and autocorrelations, can be found as well.

The model-estimated means, variances and covariances for the DSEM model are not computed as they are computed for the SEM model. The details on this computation are given in Appendix C.

5 Simulation Examples

In the following sections we illustrate the framework with several simulation studies that are also insightful in their own right.

5.1 Centering

In this section we show that the DSEM framework can be used to eliminate the dynamic panel bias, also known as Nickell's bias, see Nickell (1981). The example we use for this illustration is taken from Hamaker and Grasman (2015). The sample consist of N individuals observed at times $t = 1, ..., T$. We consider the univariate random autoregressive AR(1) model given by the following equation

$$Y_{it} = \mu_i + \phi_i(Y_{i,t-1} - \mu_i) + \xi_{it}. \tag{20}$$

The variable ξ_{it} is assumed to be white-noise with mean 0 and variance σ_w. The variables μ_i and ϕ_i have a bivariate normal distribution with mean parameters μ and ϕ, variances σ_{11} and σ_{22} and covariance σ_{12}. In the DSEM framework the above model can be estimated directly. The predictor $Y_{i,t-1} - \mu_i$ in Equation (20) is centered, that is, its mean is subtracted. Because the centering uses the true mean μ_i which is a latent variable, we call this centering the latent centering.

In contrast to the latent centering model we also consider the observed centering model

$$Y_{it} = \mu_i + \phi_i(Y_{i,t-1} - \overline{Y_{i*}}) + \xi_{it}, \tag{21}$$

where the predictor is now centered by the sample mean instead of the true mean for individual i. Model (21) can be estimated as a standard two-level regression model. However, that estimation produces Nickell's bias for the parameter ϕ because the model does not account for the error in the sample mean estimate of the true mean. Nickell (1981) also produced the following formula that approximates the bias

$$-\frac{1+\phi}{T-1}. \tag{22}$$

We conduct a simulation study to evaluate Nickell's bias, generating data according to model (20) and using the following parameter values $\mu = 0, \phi = 0.3, \sigma_{11} = \sigma_w = 3, \sigma_{22} = 0.01, \sigma_{12} = 0$. The variance σ_{22} is small so that the autocorrelation parameter ϕ_i remains in the $(-1, 1)$ range as it is a correlation

Table 1: Nickell's bias for ϕ=0.3

T	N	DSEM(latent centering)	Observed centering	Nickell's formula
10	100	0.025	-0.140	-0.144
20	50	0.006	-0.070	-0.068
30	30	0.008	-0.042	-0.045
50	50	0.000	-0.029	-0.027
100	100	-0.001	-0.014	-0.013

parameter. If the parameter ϕ_i exceeds that range, $Var(Y_{it})$ will increase with time to infinity. When data is generated for each person we need a starting value for the first time point that is generated. The standard way to resolve this ambiguity is to start at 0 but generate and discard the first few observations. That way the generated values stabilize and the effect of the original starting value of 0 is removed. We discard the first 10 values for each person.

In Table 1 we report the simulation results for various values of N and T using 100 simulated data sets for each combination of N and T. The results show that the DSEM latent centering approach resolves Nickell's bias and that the latent centering is superior to the observed centering. We also see that the bias is quite small for $T \geq 100$. The simulation study shows also that Nickell's formula predicts the bias quite accurately.

It was noted in Hamaker and Grasman (2015) that not centering the covariate also produces very good results for Nickell's bias, that is, we can replace Equation (21) with

$$Y_{it} = \mu_i + \phi_i Y_{i,t-1} + \xi_{it}. \tag{23}$$

We call this model the uncentered model. The model can also be estimated as a standard two-level regression model. The uncentered approach resolves Nickell's bias, however, it produces bias for the parameters on the between-level. In Table 2 we report the bias for $\sigma_{11} = Var(\mu_i)$ using the uncentered method and the DSEM method

We can make several conclusions from these results. The bias in the DSEM method for between-level parameter is driven by the number of subjects in the sample N and seems to disappear for $N \geq 100$. The DSEM bias

Table 2: Bias for $Var(\mu_i) = 3$

T	N	DSEM(latent centering)	Uncentered
10	100	-0.015	-1.637
20	50	0.217	-1.483
30	30	0.645	-1.256
50	50	0.378	-1.361
100	100	0.096	-1.508

is guaranteed to disappear asymptotically as the method is equivalent to the ML method for large N. It is also know that the variance of the between-level effect when $N < 100$ can be fine-tuned by using proper priors, see Browne and Draper (2006). For $N < 100$ the effect of the prior is not negligible and selecting a weakly informative prior can reduce the bias substantially. In this simulation study we used improper and uninformative priors. We can also conclude that the uncentered method yields distortion on the between-level and the bias seen in Table 2 does not disappear asymptotically.

This simulation uses a very simple DSEM model. The biases that we illustrated here for the observed centering method and the uncentered method will be difficult to track in more complicated models, especially because Nickell's bias can interact with Ludke's bias to further distort the model. We can also see clearly that the perils of the uncentered method are more dangerous from an estimation point of view as they remain in the model even with large samples.

Note also that the latent centering method used with DSEM is the only method that accommodates missing data and both the observed centering and the uncentered method are essentially not available when there are missing values in the data. The covariate can not be constructed when the data point is missing and that means that if $Y_{i,t-1}$ is missing the equation containing $Y_{i,t}$ would have to be removed as well. Thus if the data contain 20% missing data we have to remove another 20% that are next to the missing data. If the model we estimate is AR(2) we have to remove another 20% and the only data points that can be used for model estimation would be when 3 consecutive observations are all observed. This problem is in addition to the well know problems that occur when listwise deletion is used for dealing with missing data, particularly when the missing data is not MCAR.

5.2 Subject-specific variance

In regular multilevel analysis the within-level variance is generally estimated to be a cluster-invariant parameter. Even if that parameter is not cluster-invariant, the assumption of invariance generally does not affect the estimation of the structural parameters. However, for DSEM models that is not the case. Jongerling (2015) et al. show that ignoring the subject-specific variance can distort the structural parameters of the model particularly when the subject-specific variance is correlated with other random effects in the model. In this section we will reproduce this finding in the DSEM framework and discuss the general implications for DSEM modeling.

Consider the following simulation study based on the random autoregressive AR(1) model given by the following equations

$$Y_{it} = \mu_i + \varepsilon_{it} \tag{24}$$

$$\varepsilon_{it} = \phi_i \varepsilon_{i,t-1} + \xi_{it}, \tag{25}$$

where now we include subject-specific residual variance through the normally distributed random effect v_i

$$v_i = Log(Var(\xi_{it})). \tag{26}$$

The three random effects (μ_i, ϕ_i, v_i) in the above model are assumed to have an unrestricted multivariate normal distribution with mean $\nu = (2, 0.2, 0)$ and variance covariance Σ where $\sigma_{11} = 0.7$, $\sigma_{22} = 0.05$, $\sigma_{33} = 0.5$, $\sigma_{12} = \sigma_{13} = 0$. Since the covariance parameter between ϕ_i and v_i appears to be the most important parameter in this simulation study we use four different values for $\sigma_{23} = 0.15, 0.1, 0.05$, and 0. These four values correspond to the following correlation values 0.95 (high), 0.63 (medium), 0.31 (small) and 0 (none). In the simulation we use 100 replications, $N = 200$ and $T = 100$ for each value of σ_{23}. We generate data according to model (24-26) and analyze the data with the same model and the model (24-25) excluding the random variance variance effect, i.e., assuming subject invariant variance parameter. The results of the simulation are presented in Table 3.

We can make several conclusions from these results. The DSEM model without the random variance effect shows parameter bias and low coverage while the DSEM model with the random variance effect shows no bias and good coverage. Model parameter distortions are directly caused by the correlation between the random autoregressive parameter ϕ_i and the random

Table 3: Bias(coverage) for subject-specific variance simulation

parameter	$Cov(\phi_i, v_i)$	random variance	invariant variance
$E(\phi_i)$	high	.001(.97)	.040(.35)
$E(\phi_i)$	medium	.001(.98)	.028(.65)
$E(\phi_i)$	low	.001(.97)	.017(.83)
$E(\phi_i)$	none	.001(.96)	.007(.92)
$Var(\phi_i)$	high	.001(.97)	-.012(.47)
$Var(\phi_i)$	medium	.001(.93)	-.007(.78)
$Var(\phi_i)$	low	.001(.93)	-.004(.88)
$Var(\phi_i)$	none	.001(.94)	-.001(.91)

Table 4: Square root of the MSE for the random autoregressive parameters and the correlation between true and estimated random autoregressive parameters

	$Cov(\phi_i, v_i)$	DSEM random variance	DSEM invariant variance
SMSE	high	.255	.346
SMSE	medium	.293	.329
SMSE	low	.300	.316
SMSE	none	.300	.310
correlation	high	.96	.87
correlation	medium	.92	.89
correlation	low	.91	.90
correlation	none	.91	.90

variance parameter v_i. The higher that correlation is the bigger the distortions. Note that these two random effects are directly related via the following equation

$$Var(Y_{it}|i) = \frac{Exp(v_i)}{1-\phi_i^2}. \tag{27}$$

Because of that strict relationship one can expect in practical applications v_i and ϕ_i to be fairly highly correlated and therefore one can expect the DSEM model results with random variances to differ somewhat from the results without random variances. In that case we can assume that the DSEM model with random variances will yield the more accurate results.

The effect of ignoring the random variance parameters on the random autoregressive parameters is even more dramatic than the effect on the non-random parameters. To compare the estimated random autoregressive parameters with their true values we compute the square root of the mean squared error and the correlation between the estimated random autoregressive parameters and the true values

$$SMSE = \sqrt{(1/N)\sum_i (\hat{\phi}_i - \phi_i)^2} \tag{28}$$

$$correlation = Cor(\hat{\phi}_i, \phi_i) \tag{29}$$

The above quantities are computed for each of the 100 replications and the average values are reported in Table 4. The results show that the distortions in the estimates caused by ignoring the random variance effect go beyond simple inflation or deflation of the random parameters and the errors appear to have doubled from what they are for the non-random parameters. However, the cause of the increase in the SMSE error is somewhat more complex because it is not just due to the misspecification of the random variance effect. Consider for example the fact that the DSEM model with the random variance effect extracted a lot more information from the data and created v_i which is essentially a very good predictor for ϕ_i on the between-level. This will undeniably result in precision improvement for the ϕ_i estimates. Such a phenomenon exists of course not just for DSEM models but for regular two-level models as well, i.e., even when the non-random parameter estimates are not distorted, adding a random variance effect will improve the estimation of the other random effects particularly when the random variance effect is correlated with those other effects.

In multivariate DSEM models we can further consider modeling not just random variances but also random covariances and random correlations. The easiest way to model random covariance in the above framework is to model the covariance through a random variance of a common factor. However, it is not as easy to evaluate the effect of random covariance on the model estimates, because even if the factor covariance is not random the correlation between the variables is random when the variances are random. Some preliminary simulation studies, not reported here, indicate that the effect of random covariances might be more muted than those of random variances and might require much larger samples to detect. Further simulation studies are needed on this topic.

The DSEM framework can accommodate seamlessly a large number of random effects and thus using models with random variances and covariances in many situations should be the preferred choice as long as the MCMC convergence is unhindered. Because of the increase in the number of random effects, the likelihood of the model with these random variances and covariances will be less pronounced and in some cases the MCMC convergence will be much slower. This should be taken as an indication that there is not sufficient information in the data to identify the DSEM model with random variances and covariances. In such situations using cluster-invariant variances and covariances is not a poor choice by any means and unless these random variances and covariances are highly correlated with other random parameters we see that the effects are somewhat negligible.

5.3 ARMA(1,1) and the measurement error AR(1) model

The ARMA(1,1) time-series model is given by the following equation

$$Y_t = \mu + \phi Y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}. \tag{30}$$

The model has 4 parameters μ, ϕ, θ and $\sigma = Var(\varepsilon_t)$. The ARMA(1,1) process is stationary and invertible, see Green (2014), when the two parameters ϕ and θ are within the interval $(-1, 1)$ and generally when used in practical applications we expect these two parameters to be within that range. The model-implied variance for the ARMA(1,1) model is given by

$$Var(Y_t) = \sigma \left(1 + \frac{(\phi + \theta)^2}{1 - \phi^2} \right). \tag{31}$$

The model-implied first autocorrelation is given by

$$\rho(1) = \frac{(\theta + \phi)(1 + \theta\phi)}{1 + 2\theta\phi + \theta^2} \tag{32}$$

and for lag $l > 1$ the autocorrelation is given by

$$\rho(l) = \phi^{l-1}\rho(1). \tag{33}$$

It is shown in Schuurman et al. (2015) that this model is equivalent to the following measurement error AR(1) model under certain parameter restrictions

$$Y_t = \mu + f_t + \xi_t \tag{34}$$

$$f_t = \phi f_{t-1} + \epsilon_t. \tag{35}$$

We call this model the measurement error AR(1) model, that is, MEAR(1), because the latent variable f_t follows an AR(1) process but is not observed directly, rather, it is measured with error by the observed variable Y_t. This model is also sometimes referred to as AR(1)+WN (Granger & Morris, 1976), where WN stands for the white noise process representing the measurement error. The 4 parameters in this model are μ, ϕ, $\sigma_1 = Var(\xi_t)$ and $\sigma_2 = Var(\epsilon_t)$. The relationship between the parameters in the two models is as follows. The parameters μ and ϕ are unchanged while the MEAR(1) parameters σ_1 and σ_2 can be derived from the ARMA(1,1) parameters via the following equations

$$\sigma_1 = -\frac{\theta\sigma}{\phi} \tag{36}$$

$$\sigma_2 = (1 + \theta^2)\sigma + \frac{(1 + \phi^2)\theta\sigma}{\phi}. \tag{37}$$

The equivalence of the two models is subject to the parameter constraints that arise from the inequalities $\sigma_1 > 0$ and $\sigma_2 > 0$. Under the regularity conditions of ϕ and θ being in the interval (-1,1) the constraints can be further simplified to

$$\phi\theta < 0 \tag{38}$$

$$\phi + \theta > 0 \tag{39}$$

Every MEAR(1) model can be represented as an ARMA(1,1) model, while an ARMA(1,1) model can be represented by a MEAR(1) model when (36) and (37) produce positive variances or equivalently when (38) and (39) hold.

In the most common situation the autoregressive parameter ϕ will be positive. Let's assume for now the case of $\phi > 0$. In that case it is interesting to note for the MEAR(1,1) model that the autocorrelation parameters for the latent variable are always larger or equal to those for the observed variable

$$Cor(f_t, f_{t-l}) \geq Cor(Y_t, Y_{t-l}). \tag{40}$$

For the ARMA(1,1) model this is not the case and the corresponding statement

$$\phi^l > \rho(l) \tag{41}$$

is precisely equivalent to θ being negative which is the necessary condition for the ARMA(1,1) models to be equivalent to the MEAR(1) model, i.e., these constraints are no coincidences and have meaningful interpretations.

The MEAR(1) model is much easier to interpret than the ARMA(1,1) model, especially in the social sciences applications where measurement error is common. In cross sectional studies it is not possible to identify the measurement error model when there is only one measurement but as the MEAR(1) model clearly illustrates it is possible to do that in dynamic time-series models.

The MEAR(1)/ARMA(1,1) model is generally preferred to the AR(1) model in the econometrics literature as it offers more flexible autoregressive representation. The AR(1) model has an exponential decay of the autocorrelation function while the ARMA(1,1) autocorrelation decays slower. This is particularly important if we have to change time scale as is done with continuous time dynamic modeling. AR(1) hourly autocorrelation of 0.75 implies a daily autocorrelation of 0.001. With reliability of 0.8 the MEAR(1) model hourly autocorrelation of 0.75 implies a daily autocorrelation of 0.212. The AR(1) model implies that observations in two consecutive days would be approximately independent, while the MEAR(1) model implies that some lag relations will remain across consecutive days, which is a more realistic assumption. The difference in the decay of the autocorrelation is illustrated in Figures 1 and 2 which show typical decay for the autocorrelation for the AR(1) and ARMA(1,1) models.

Figure 1. AR(1) autocorrelation decay function

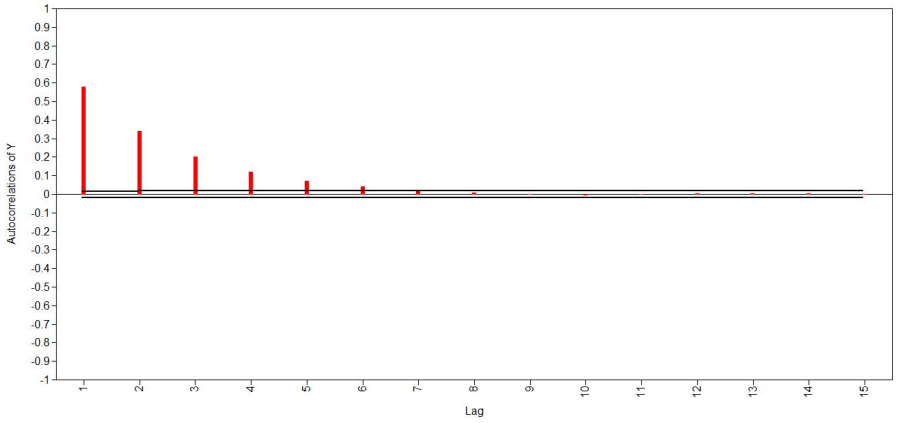
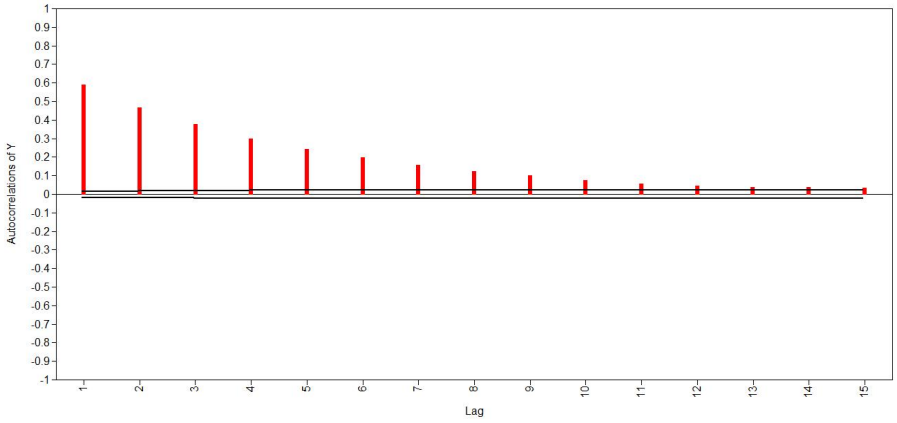


Figure 2. AR(1,1) autocorrelation decay function



In practical settings one can compute the sample autocorrelations and check if the decay is exponential or not and based on that decide if the AR(1) model is sufficent or the ARMA(1,1) should be explored. Of course there are many other possibilities such as the AR(2) model, the more general ARMA(p,q) model, or a MEAR(p) model which is a special case of the ARMA(p,p) model, see Granger and Morris (1976).

Next we conduct a brief simulation study to evaluate the performance of the estimation of the MEAR(1)/ARMA(1,1) model and to evaluate the sample size needed to obtain satisfactory estimates. We use the $N = 1$ case with $T = 100, 200, 300, 500$. We use 100 replications in all cases. The results

Table 5: Bias(coverage) AR(1) measurement error model / ARMA(1,1), N=1

parameter	True value	$T = 100$	$T = 200$	$T = 300$	$T = 500$
μ	0	-.09(.82)	-.01(.89)	-.04(.85)	-.02(.87)
ϕ	.8	-.07(.96)	-.04(.92)	-.03(.87)	-.01(.95)
σ_1	1	-.10(.97)	-.09(.94)	-.08(.88)	-.04(.90)
σ_2	1	.25(.95)	.17(.92)	.14(.91)	.08(.90)

are presented in Table 5. We use the MEAR(1) model formulation given in Equations (34) and (35).

We can make the following conclusions from these results. The estimation of the ARMA(1,1) model is more difficult than the estimation of the AR(1) model. Good estimation where the bias is small and the coverage is near or above 90% needs at least $T \geq 200$. The estimates are biased at $T = 100$ and coverage dropped to 82%. We can also conclude that to estimate a two-level ARMA(1,1) model, with all four of the parameters as random subject-specific parameters at least $T \geq 200$ is needed per person. If such a sample size is not available then one or two of the four ARMA(1,1) parameters should be held equal across individuals, i.s., should be non-random parameters. Most suitable are the two variance parameters σ_1 and σ_2 because the bias in Table 3 is smaller than the bias in Table 5 for $T = 100$. When parameters are held equal across individuals essentially the sample size used for the estimation changes from T to $N \cdot T$ and we can estimate a two-level ARMA(1,1) model with much fewer observations per person than we need for a single-level ARMA(1,1) model.

In the next simulation study we illustrate the two-level MEAR(1) model using categorical data. The DSEM framework has one limitation when it comes to categorical variables. Such variables can not be lagged on its own but only through a factor. The MEAR(1) model essentially resolves this problem as it includes such a factor already and thus we can estimate a univariate autoregressive model with a categorical variable. If we attempt to estimate a subject-specific auto-regressive model, such as the one in Equation (20), where the autoregressive parameter is subject-specific we will need a substantial sample size. Simulation studies, not reported here, indicate that for the N=1 case the MEAR(1) model needs a sample size of about 10000 for the binary case and about 1000 for an ordered polytomous case with 6

Table 6: Two-level ARMA(1,1) with binary variable, N=100, T=300

parameter	True value	Estimate(Coverage)
μ	0	0.00 (.95)
ϕ	.5	0.50(.78)
σ_w	1	1.01(.71)
σ_b	0.5	0.52(.94)

categories. This is the kind of sample size we would need per subject if we want to estimate a subject-specific two-level MEAR(1) model. Such a sample size, however, is not common in practical applications. If we estimate a two-level MEAR(1) model where the autoregressive parameter is not subject-specific then the data from the different subjects are combined and much fewer observations will be needed per subject. Our simulation study uses $N = 100$ individuals with $T = 300$ time points and 100 replications. Since the autoregressive coefficient is estimated at the population level, we essentially have $100 \cdot 300 = 30000$ observations to estimate this model which is sufficient.

The MEAR(1) model we estimate for the binary variable is given by

$$P(Y_{it} = 1) = \Phi(\mu_i + f_{it}) \tag{42}$$

$$f_{it} = \phi f_{i,t-1} + \xi_{it} \tag{43}$$

$$\mu_i \sim N(\mu, \sigma_b), \quad \xi_{it} \sim N(0, \sigma_w) \tag{44}$$

The function Φ is the standard normal distribution function. Note that for identification purposes the residual variance in Equation (34) is now fixed to 1. The model has four parameters μ, σ_b, ϕ and σ_w. The results of the simulation study are presented in Table 6. Parameter estimates appear to have no bias, however, some of the parameters have low coverage. This usually can be resolved by running longer MCMC chains. Here we used a minimum of 1000 MCMC iterations and convergence is determined by the PSR convergence criterion, see Asparouhov and Muthén (2010). Mixing with categorical variables is somewhat slower than with normally distributed variables and may require much longer MCMC chains. This simulation takes one minute per replication.

Let's also consider the model with ordered polytomous variables. Using ordered polytomous variables in practical applications is one way to deal

Table 7: Two-level ARMA(1,1) with ordered polytomous, N=100, T=100

parameter	True value	Estimate(Coverage)
τ_1	-3	-3.06 (.87)
τ_2	-1	-1.02 (.81)
τ_3	0	-0.01 (.79)
τ_4	1	1.01 (.75)
τ_5	3	3.05 (.81)
ϕ	.5	0.50(.93)
σ_w	1	1.09(.83)
σ_b	0.5	0.54(.94)

with non-normally distributed dependent variables. The model is given by the following equations

$$P(Y_{it} = j) = \Phi(\tau_{j+1} - \mu_i - f_{it}) - \Phi(\tau_j - \mu_i - f_{it}) \tag{45}$$

$$f_{it} = \phi f_{i,t-1} + \xi_{it} \tag{46}$$

$$\mu_i \sim N(0, \sigma_b), \quad \xi_{it} \sim N(0, \sigma_w) \tag{47}$$

The first $\tau_0 = -\infty$ and the last threshold $\tau_J = \infty$, where J is the number of categories of the observed variable. We conduct a simulation study using a 6 category variable, $N = 100$, $T = 100$ and 100 replications. The results are presented in Table 7. The parameter bias appears to be small and again we see some standard error underestimation that could potentially be resolved with running much longer MCMC chains. Each replication takes 2 minutes. Note here that because the outcome is ordered polytomous we were able to estimate the model only with $T = 100$ which is much smaller than we needed with the binary outcome. This is due to the fact that the ordered polytomous variable carries more information than the binary variable.

5.4 How to add a covariate in the MEAR(1) and AR(1) models

Note that the AR(1) model is nested within the MEAR(1) model. In Equation (34) if we set the parameter $Var(\xi_t) = 0$, i.e., if we set the measurement

error to zero the model becomes equivalent to the AR(1) model. The following discussion applies to both the AR(1) and the MEAR(1) models. There are three ways to add a covariate in the MEAR(1) model given in Equations (34) and (35). The covariate can be used in either of the two equations but it can also be used in both equations. In total we have three models. We call the following model the direct model

$$Y_t = \mu + f_t + \beta_1 X_t + \xi_t \tag{48}$$

$$f_t = \phi f_{t-1} + \epsilon_t. \tag{49}$$

The following model we call the indirect model

$$Y_t = \mu + f_t + \xi_t \tag{50}$$

$$f_t = \phi f_{t-1} + \beta_2 X_t + \epsilon_t. \tag{51}$$

The following model we call the full model

$$Y_t = \mu + f_t + \beta_1 X_t + \xi_t \tag{52}$$

$$f_t = \phi f_{t-1} + \beta_2 X_t + \epsilon_t. \tag{53}$$

The full model has a direct and an indirect effect from X on Y.

The first issue that we have to address is the fact that the full model is not identified in some special cases. The first case where the model is not identified is the case where the autoregressive parameter $\phi = 0$. In that case the model is a standard SEM model and the direct and indirect effects on Y are equivalent and therefore the full model which includes both effects is not identified. Note also that in the case of $\phi = 0$ the measurement error model is not identified as well with or without a covariate as the one indicator factor model is not identified in standard SEM.

Another case where the full model is not identified is the two-level MEAR(1) model where the covariate is time invariant. If the covariate is time invariant, then the indirect and the direct model become equivalent and the relationship between the parameters is as follows

$$\beta_2 = \frac{\beta_1}{1 - \phi}. \tag{54}$$

This relationship holds because when the covariate is time invariant it is essentially equivalent to the μ parameter. If the μ parameters is moved from

Equation (52) to Equation (53) it will also be divided by $(1-\phi)$. Because the indirect and the direct models are equivalent the full model is not identified.

Another covariate for which the indirect and direct model become equivalent and the relationship (54) holds is the case $X_t = t$, i.e., the linear growth model, see Hamaker (2005). We formulate this equivalence for the AR(1) model but the same holds for the MEAR(1) model. The direct linear growth AR(1) model is formulated as follows

$$Y_t = \gamma_0 + \gamma_1 t + \xi_t \tag{55}$$

$$\xi_t = \phi\xi_{t-1} + \varepsilon_t. \tag{56}$$

The indirect linear growth AR(1) model can be formulated as follows

$$Y_t = \beta_0 + \beta_1 t + \phi Y_{t-1} + \varepsilon_t. \tag{57}$$

The difference between the two models is that the autoregressive structure is imposed on the residuals variable ξ_t in the direct model while in the indirect model it is imposed on the observed variable Y_t. Simple algebraic manipulations show that the direct and indirect models are algebraically equivalent and the relationship between the parameters is as follows

$$\gamma_0 = \frac{\beta_0}{1-\phi} - \frac{\phi\beta_1}{(1-\phi)^2} \tag{58}$$

$$\gamma_1 = \frac{\beta_1}{1-\phi} \tag{59}$$

while the parameters ϕ and $Var(\varepsilon_t)$ remain unchanged. Here we conclude once again that if $X_t = t$ the full model is unidentified. Note also that the equivalence between the direct and the indirect linear growth models does not translate completely in two-level models with subject-specific random parameters particularly when there are covariates predicting the random effects β_j and ϕ. Linear relationship between a covariate and subject-specific β_j and ϕ will result in non-linear relationship of that covariate with γ_j. Thus the two-level indirect linear growth model is not equivalent to the two-level direct linear growth model which estimates a linear relationship between the covariate and γ_j.

Consider now the quadratic growth AR(1) model. The direct quadratic growth AR(1) model is

$$Y_t = \gamma_0 + \gamma_1 t + \gamma_1 t^2 + \xi_t \tag{60}$$

$$\xi_t = \phi\xi_{t-1} + \varepsilon_t. \tag{61}$$

The indirect quadratic growth AR(1) model is

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \phi Y_{t-1} + \varepsilon_t. \tag{62}$$

Simple algebraic manipulations show again that the models are algebraically equivalent and the relationship between the parameters is as follows

$$\gamma_0 = \frac{\beta_0}{1-\phi} - \frac{\phi\beta_1}{(1-\phi)^2} + \frac{\beta_2\phi(1+\phi)}{(1-\phi)^3} \tag{63}$$

$$\gamma_1 = \frac{\beta_1}{1-\phi} - \frac{2\phi\beta_2}{(1-\phi)^2} \tag{64}$$

$$\gamma_2 = \frac{\beta_2}{1-\phi} \tag{65}$$

while the parameters ϕ and $Var(\varepsilon_t)$ remain unchanged. Similar algebraic equivalence can be constructed with any polynomial growth AR(1) model. Again the equivalence between the direct and the indirect model implies that the full model is unidentified when βX_t represents a polynomial of t, i.e., if the predictor is a polynomial of t we can use it in either of the Equations (52) or (53) but not both. Another conclusion that we can make is that the simple relationship given in (54) where one simply divides the direct effect by $1-\phi$ to obtain the equivalent indirect effect doesn't hold except for the two cases of linear growth model and a between-level covariate. The relationship shown in the quadratic growth case is more complex. Not only that but we see that the relationship does not depend only on the covariate but also on what other covariates there are in the model. When $X_t = t$, the relationship between the direct and the indirect effect changed after we added another covariate t^2.

Let's consider now the fundamental difference between the direct, indirect and the full model using the conditional expectation $E(Y_t|X)$. For the direct model we have

$$E(Y_t|X) = \mu + \beta_1 X_t. \tag{66}$$

For the indirect model we have

$$E(Y_t|X) = \mu + \beta_2(X_t + \phi X_{t-1} + \phi^2 X_{t-2} + \phi^3 X_{t-3} + ...). \tag{67}$$

For the full model we have

$$E(Y_t|X) = \mu + \beta_1 X_t + \beta_2(X_t + \phi X_{t-1} + \phi^2 X_{t-2} + \phi^3 X_{t-3} + ...). \quad (68)$$

The interpretation is very clear. In the direct model the condition expectation depends only on the current value of X_t. This value may or may not depend on the prior values of X, but this is not a part of the model as we model only the conditional distribution of Y given X. Regardless of what the X_t process is, it is clear that the conditional expectation of Y_t can depend only on X_t and if there is any dependence on X_{t-1} it is only indirect through the effect of X_{t-1} on X_t.

The interpretation for the indirect effect is completely different. The current conditional expectation accumulates the effect of all prior values of X with diminishing influence when the model is stationary and $|\phi| < 1$. The power ϕ^l will converge to 0 as l increases and so will the effect of X_{t-l} on Y_t. The interpretation of the full model combines the two. It allows accumulated effect of X_t as well as a special direct effect exceeding the accumulating effect for the current value X_t. In practical applications we can determine the type of influence a covariate should have, accumulated v.s. direct, by estimating the full model and considering the significance of the two effects β_1 and β_2.

Next we illustrate the performance of the full model in a two-level simulation study. We will use the MEAR(1) model but the simulation results using the AR(1) model are similar. The full two-level MEAR(1) model is given by

$$Y_{it} = \mu_i + f_{it} + \beta_1 X_{it} + \xi_{it} \quad (69)$$

$$f_{it} = \phi f_{i,t-1} + \beta_2 X_{it} + \epsilon_{it}, \quad (70)$$

where μ_i is a between-level random effect with mean μ and variance σ_b. We generate and analyze 100 samples using this model and the following parameter values $\beta_1 = 0.3$, $\beta_2 = 0.4$, $\phi = 0.5$, $\mu = 0$, $\sigma_b = 0.7$, $Var(\xi_{it}) = Var(\epsilon_{it}) = 1$. The covariate X_{it} is generated from an AR(1) process with $Var(X_{it}) = 1$ and autocorrelation ϕ_x. We use three different values for ϕ_x: 0, 0.5 and 0.8. The sample consist of $N = 200$ individuals each observed at $T = 100$ times.

Table 8 contains the results of the simulation study for the structural parameters and the various values of ϕ_x. The results show that the parameter estimates are unbiased, the coverage is acceptable and the model is well identified.

Table 8: Two-level full MEAR(1) with covariate, N=200, T=100

parameter	ϕ_x	True value	Estimate(Coverage)
β_1	0	.30	.30(.87)
β_1	0.5	.30	.30(.96)
β_1	0.8	.30	.31(.89)
β_2	0	.40	.40(.87)
β_2	0.5	.40	.40(.93)
β_2	0.8	.40	.40(.90)
ϕ	0	.50	.50(.88)
ϕ	0.5	.50	.50(.93)
ϕ	0.8	.50	.50(.93)

Next we analyze the same data using the two-level direct and indirect MEAR(1) models. The results are presented in Tables 9 and 10. For the effect of the covariate in these tables we used 0.7 as this is the sum of the two effects, however, there is no true value since the model is misspecified. Also clearly the estimated effect for both direct and indirect model is not near 0.7 and is highly dependent on the autocorrelation parameter ϕ_x. Both direct and indirect models failed to capture well the covariate effect. In addition the autocorrelation parameter is distorted and coverage appears insufficient for both the indirect and the direct models. It appear that the level of distortion in the model parameters is directly related to how close the indirect or the direct model is to the full model. The further away these models are from the full model the bigger the biases.

It is also possible to estimate random direct and indirect effects in the full two-level MEAR(1) model in addition to a random autoregressive effect. Table 11 shows the results of a small simulation study with 100 replications, $N = 200$ and $T = 100$. Note that we are able to estimate this two-level random MEAR(1) model only with $T = 100$ due to the fact that only two of the four MEAR(1) parameters are subject-specific and the two variance parameters are not random. The fact that the model includes a covariate with two random effects does not appear to complicate the model estimation. The results show that the parameter estimates are unbiased, the coverage is acceptable and the model is well identified.

Table 9: Two-level full MEAR(1) with covariate analyzed as direct, N=200, T=100

parameter	ϕ_x	True value	Estimate(Coverage)
β_1	0	.70	.65(.00)
β_1	0.5	.70	.74(.07)
β_1	0.8	.70	.88(.00)
ϕ	0	.50	.50(.92)
ϕ	0.5	.50	.51(.85)
ϕ	0.8	.50	.52(.83)

Table 10: Two-level full MEAR(1) with covariate analyzed as indirect, N=200, T=100

parameter	ϕ_x	True value	Estimate(Coverage)
β_2	0	.70	.69(.92)
β_2	0.5	.70	.67(.21)
β_2	0.8	.70	.65(.07)
ϕ	0	.50	.36(.00)
ϕ	0.5	.50	.38(.00)
ϕ	0.8	.50	.41(.00)

Table 11: Two-level full MEAR(1) model with random effects, N=200, T=100

parameter	True value	Estimate(Coverage)
$E(\beta_{1i})$	0.3	.30(.91)
$E(\beta_{2i})$	0.4	.40(.91)
$E(\phi_i)$	0.2	.20(.88)
$Var(\beta_{1i})$	0.1	.10(.94)
$Var(\beta_{2i})$	0.1	.10(.95)
$Var(\phi_i)$	0.01	.01(.94)
$Var(\xi_{it})$	1	.98(.81)
$Var(\epsilon_{it})$	1	1.02(.82)

5.5 Dynamic Factor Analysis

Most of the dynamic factor analysis models considered previously have been for the case of $N = 1$, i.e., when a single-subject time series data are fitted with a factor analysis model across time. The DSEM framework described here includes dynamic factor analysis models for an entire population rather than a single subject only. The two most common dynamic factor models are the direct autoregressive factor score (DAFS) and the white noise factor score (WNFS) models, see Zhang and Nesselroade (2007). The DAFS model is given by the following equations

$$Y_t = \nu + \Lambda\eta_t + \varepsilon_t \tag{71}$$

$$\eta_t = \sum_{l=1}^{L} B_l\eta_{t-l} + \xi_t. \tag{72}$$

The WNFS model is given by the following equation

$$Y_t = \nu + \sum_{l=0}^{L} \Lambda_l\eta_{t-l} + \varepsilon_t. \tag{73}$$

The difference between the two models is clear. In the DAFS model only the current factor affects the observed variables while in the WNFS model the observed variables are also affected by the previous periods' factor values. In addition, the factor in the DAFS model is an AR(L) process, while the factors in the WNFS model are independent across time, i.e., the factors follow a white noise process. The implications for the observed variables are also different. The observed variables in the WNFS model follow a MA(L) process, while the observed variable in the DAFS model follows an ARMA(p,p) process. In fact the DAFS model for $L = 1$ is the MEAR(1) model for each factor indicator.

In practical applications inevitably the question arises of which one of the two factor analysis models should be used. We will add into this consideration the following hybrid DAFS+WNFS model that is nested above both the DAFS and the WNFS models

$$Y_t = \nu + \sum_{l=0}^{L} \Lambda_l\eta_{t-l} + \varepsilon_t \tag{74}$$

$$\eta_t = \sum_{l=1}^{L} B_l \eta_{t-l} + \xi_t \tag{75}$$

This model is considered also in Molenaar (2017). In fact the model is referred to as a DFM(p,q,L,L,0), where p refers to the number of observed variables in the factor model and q refers to the number of factors in the model. It is interesting to note that the hybrid DAFS+WNFS model is equivalent to a DAFS model where the factor follows an ARMA(L,L) process if the loadings Λ_l are proportional in the one factor model, or can be rotated into the same loadings in the multivariate case. Such a model would be referred in Molenaar (2017) terminology as a DFM(p,q,0,L,L) model. The hybrid model is also interesting because it illustrates how DSEM models differ from SEM models. In SEM models it is not possible to identify a model where a factor predictor is also a direct predictor for all indicator variables, while in the hybrid DAFS+WNFS this is possible.

In the following simulation study we illustrate the performance of the estimation method for a two-level hybrid DAFS+WNFS model. We use an $L = 1$ model with 5 indicators and 1 factor. We generate and analyze 100 samples with $N = 100$ and $T = 100$. The two level model also has a between-level factor model and the full model is given by the following equations

$$Y_{it} = Y_{1,it} + Y_{2,i} \tag{76}$$

$$Y_{1,it} = \Lambda_0 \eta_{1,t} + \Lambda_1 \eta_{1,t-1} + \varepsilon_{1,t} \tag{77}$$

$$\eta_{1,t} = \phi \eta_{1,t-1} + \xi_t \tag{78}$$

$$Y_{2,i} = \nu + \Lambda_b \eta_{2,t} + \varepsilon_{2,t} \tag{79}$$

We generate the data using the following parameter values which for simplicity are identical across the five indicators. For $j = 1, ..., 5$ we set $\lambda_{0,j} = 1$, $\lambda_{1,j} = 0.6$, $\theta_{1,j} = Var(\varepsilon_{1,t,j}) = 1$, $\phi = 0.4$, $\psi_1 = Var(\xi_t) = 1$, $\nu_j = 0$, $\lambda_{b,j} = 0.5$, $\psi_2 = Var(\eta_{2,t}) = 1$, $Var(\varepsilon_{2,t,j}) = 1$. In the estimation we fix the variance ψ_1 and ψ_2 to 1 for identification purposes. Table 12 contains the results of the simulation study for a selection of the model parameters. The estimates show no bias and good coverage is obtained.

Next we illustrate how the DIC criterion can be used for model selection. Using the same generated data we estimate the two-level DAFS model, the two-level WNFS model and the two-level hybrid DAFS+WNFS model. In all three models we use the correct one-factor model on the between-level. The

Table 12: Two-level hybrid DAFS+WNFS, N=100, T=100

parameter	True value	Estimate(Coverage)
$\lambda_{0,1}$	1	1.00(.92)
$\lambda_{1,1}$	0.6	0.60(.93)
$\theta_{1,1}$	1.0	1.00(.95)
ϕ	0.4	0.40(.95)
ν_1	0	0.00(.95)
$\lambda_{b,1}$	0.5	0.51(.94)
$\theta_{2,1}$	0.2	0.21(.97)

Table 13: DIC comarision

model	average DIC	smallest DIC value
two-level DAFS	150235	0%
two-level WNFS	149983	1%
two-level WNFS+DAFS	149813	99%

average DIC values across the 100 replications are given in Table 13. In each replication we compare the DIC across the three models and select the model with smallest value. In 99 out of 100 replication the correct WNFS+DAFS model had the smallest DIC value, i.e., the DIC performed well in identifying the correct model.

5.6 Subject-specific and uneven times of observations

In this section we illustrate the quality of the estimation when the times of observations vary across individuals and when they are unevenly spaced. We conduct two different simulation studies. The first study is based on a two-level DAFS AR(1) model and the second is based on a two-level AR(1) model.

The estimation algorithm described in Appendix A indicates that the quality of the estimation depends on the amount of missing data inserted between the observed values and how accurately the original times of observations are approximated by the integer grid that is used in the DSEM

Table 14: Two-level DAFS AR(1) with subject-specific times

percentage missing values	$\hat{\phi}$ (coverage) $\phi = 0.4$	convergence rate	comp time per replication in min
.80	.39(.95)	100%	1.5
.85	.39(.90	95%	2.5
.90	.35(.46)	55%	10
.95	.34(.55)	55%	18

estimation. The more accurate the approximation the more missing data will be inserted. In the first simulation study we want to see how the percentage of missing data affects the parameter estimates, convergence rates and speed of the estimation. We generate samples with 100 individuals using the same two-level AR(1) model we used in the previous section with the exception that we set Λ_1 to 0 so that the model is simply a DAFS AR(1) model rather than a hybrid. We generate T observations for each individual and mark m percent of these observations as missing at random. To be more precise, for each individual, each time point is marked as missing with probability m, and is removed from the data set. Simulation study will be conducted with four different m values: .80, .85, .90 and .95, i.e., the simulation study will have between 80% and 95% missing values. We also vary T as a function of m and we set $T = 60/(1 - m)$ which implies that on average after the missing values are removed each individual will have 60 observations taken at various uneven and unequal times. For each value of m we generate and analyze 20 data sets.

The results of this simulation study are given in Table 14. We report the average estimates and coverage for the autoregressive parameter ϕ for the within-level factor, the convergence rate for the estimation and the computational time per replication. The results show that in this model the quality of the estimation deteriorates as the amount of missing data reaches 90%. As the amount of missing data increases, the computational time increases, the number of convergence problems increases, and the quality of the estimates decreases in terms of bias and coverage. However, the results are acceptable for 80% or 85% missing values. Adding too many missing values between the observed data can destabilize the MCMC estimation.

In the next simulation study we will use the simpler two-level AR(1) model

$$Y_{it} = \mu_i + \phi(Y_{i,t-1} - \mu_i) + \varepsilon_{it} \tag{80}$$

$$\mu_i \sim N(\mu, v). \tag{81}$$

We use the following parameters to generate the data $\mu = 0$, $v = 0.5$, $Var(\varepsilon_{it}) = 1$, $\phi = .8$. We again use $N = 100$ and $T = 60/(1-m)$, where m is the percentage of missing data. In this simulation m takes just two values 0.80 and 0.95. The missing data is generated at random and that generates subject-specific times of observations. For example when $m = .95$, $T = 1200$ and each individual has approximately 60 observations that occur at times between 1 and 1200.

In this simulation we will vary the interval δ used in Appendix A. This interval is specified in Mplus using the tinterval option. We will estimate the two-level AR(1) model using different values of $\delta = 1, 2, 3, 4, 5, 10$. The case of $\delta = 1$ is the original time scale. As δ increases we use a more and more crude time scale, worsening the time scale approximation. Note also that we can not directly compare the models using different vales of δ. Denote by ϕ_j the estimated autocorrelation coefficient for $\delta = j$. This is also the autocorrelation of lag j for the original process and therefore $\phi_j = \phi_1^j$. To compare the models we will use $\phi_j^{1/j}$ which is the implied estimate for $\phi_1 = \phi = 0.8$. Note here that when we use $\delta > 1$ the data will be rearranged and a different amount of missing data will be inserted. Let's denote this missing data as m_2. This is the missing data that is being used in the analysis. For each of the values of m we generate 100 data sets and we analyze those with the various values of δ.

The results are presented in Table 15. In all cases the rate of convergence is 100%. This means that simpler models like the two-level AR(1) model can tolerate more missing data than the more complex models like the DAFS AR(1) model. We can also see from the results that the cruder the scale is the more biased the results are. The smaller the inserted amount of missing data is the more biased the estimates are. It is also somewhat clear that it will be impossible to establish a clear rule of thumb for δ and the amount of missing data that should be used. These quantities are probably going to remain specific to the particular examples. However, the trends are clear. The smaller δ is the better the estimates are. If δ is too small and the inserted missing data is too big the MCMC chain might experience convergence problems.

Table 15: Two-level AR(1) with subject-specific times. Estimates and coverage for ϕ and amount of missing data m_2 during the analysis.

m	δ	$\phi = 0.8$	m_2
.80	1	.80(.91)	.80
.80	2	.81(.31)	.58
.80	3	.83(.00)	.38
.80	4	.84(.00)	.18
.80	5	.86(.00)	.05
.80	10	.92(.00)	.00
.95	1	.80(.85)	.95
.95	2	.81(.57)	.90
.95	3	.82(.20)	.85
.95	4	.83(.00)	.80
.95	5	.84(.00)	.74
.95	10	.88(.00)	.49

In practical applications when estimating an AR(1) model to verify that a particular value of δ is sufficiently small one can simply compare the results for the autocorrelation parameter using δ and $\delta/2$. If $\phi_\delta \approx \phi^2_{\delta/2}$ we can conclude that δ is sufficiently small. If that is not approximately true then we should interpret that result as evidence that δ should be decreased or that the AR(1) model does not hold. In fact we can test that method with our simulated data for the case of $m = .80$. The estimate for ϕ using $\delta = 1$ is $\phi_1 = 0.8002$. The estimate for ϕ using $\delta = 0.5$ is $\phi_{0.5} = 0.8943$ and the estimate for $\phi^2_{0.5} = 0.7998$, which confirms that $\delta = 1$ is sufficiently refined as it yields the same model as the more precise $\delta = 0.5$. Note here that if the model is a more complicated time-series model, rather than a simple AR(1) model, the connection between the time series model for Y_t and the model for Y_{2t} is much more complicated. This problem is somewhat compounded by the fact that such a question has not been of interest in the econometric literature while it is of interest in the social sciences and this DSEM framework particularly for the purpose of addressing subject-specific and uneven times of observations.

Overall it appears that the optimal amount of inserted missing data

should be somewhere between 80% and 95%, depending on how complex the model is. This corresponds to 5% to 20% present data and covariance coverage as reported in the Mplus output. In a practical setting one should of course consider interpretability in the choice of δ. If times of observations are recorded on "days" metric, choosing δ to represent one day is the most natural choice and it will preserve the interpretability of the model.

It is also worth noting here that when δ values increase to a sufficiently large value the amount of missing data converges to 0% which means that the time scale is completely ignored and the times of observations are set to $1, 2, ...$, i.e., are assumed to be consecutive. In our example this happened for $m = 0.80$ and $\delta = 10$. The estimate of the autocorrelation coefficient is 0.92 which is $\phi_{10}^{0.1}$, i.e., the raw estimate of the autocorrelation is $\phi_{10} = 0.45 \approx 0.92^{10}$. This is the autocorrelation that one would get by estimating the data and ignoring the subject-specific and uneven times of observations. Such an estimate of course is quite different from the true value of 0.8.

There are many other ways to deal with subject-specific and uneven times of observations. For example, continuous time modeling can be performed using Brownian motion theory, or using dynamic models based on differential equations, see Deboeck and Preacher (2016). Another possible approach is to use the times between consecutive observations in the model to reflect the strength of the relationship between the observations, i.e., having the autoregressive parameters depend on the distance between the observations. Yet another method is to use the same approach of missing data insertion but to change the algorithm described in Appendix A. As described there the algorithm focuses on global time scale matching. A different algorithm that focuses on matching consecutive time differences could potentially yield more accurate results. Such alternative algorithms can easily be studied with Mplus by preprocessing the continuous times of observations before employing the DSEM analysis. Clearly this is a vast research topic and there are many possibilities for improving the treatment described here. The main advantage of the method we chose is that it can fit smoothly in the general framework, apply to all models, and work fairly well as the above simulations show.

5.7 Time-specific effects

In this section we illustrate the TVEM feature of the DSEM framework with an ARMA(1,1) model with a covariate where the random random intercept

and random slope evolve over time. We will use the MEAR(1) version of the ARMA(1,1) model. The model is given by the following equations

$$Y_{it} = \mu_t + Y_i + \beta_t X_{it} + f_{it} + \varepsilon_{it} \tag{82}$$

$$f_{it} = \phi f_{i,t-1} + \xi_{it} \tag{83}$$

In this model Y_i is a subject-specific random effect, while μ_t and β_t are time-specific random effects. We generate a single data set with 500 individuals each observed at times 1,2,...,50, using the following parameters $\theta_b = Var(Y_i) = 0.5$, $\theta_w = Var(\varepsilon_{it}) = 0.5$, $\phi = 0.5$ and $\psi = Var(\xi_{it}) = 1.2$. We generate the covariate X_{it} from a standard normal distribution. The time-specific effects μ_t and β_t are generated from arbitrary functions of time. In this simulation we use a logarithmic function for μ_t and a quadratic function for β_t as follows

$$\mu_t = g_1(t) = log(t) \tag{84}$$

$$\beta_t = g_2(t) = a + bt + ct^2 = 0.001 \cdot t \cdot (50 - t). \tag{85}$$

We can estimate this model in the DSEM framework assuming that μ_t and β_t are normally distributed random effects with distributions $N(\mu, v_\mu)$ and $N(\beta, v_\beta)$. This is technically an incorrect assumption because μ_t and β_t are not time invariant, $E(\mu_t|t) = g_1(t)$, $Var(\mu_t|t) = 0$, $E(\beta_t|t) = g_2(t)$, $Var(\beta_t|t) = 0$. Nevertheless, we can use the DSEM framework to estimate the above model as a first step in an exploratory fashion. Because there are 500 observations at each time point, the prior assumptions, $N(\mu, v_\mu)$ and $N(\beta, v_\beta)$, for these two random effects will have only a minor if any effect on the estimates. The estimates of μ_t and β_t will be dominated by the data. Table 16 contains the results for the non-random parameters of this analysis. The estimates are near the true values and the credibility intervals contain the true value for all four parameters. Figures 3 and 4 show the estimated values of μ_t and β_t compared with the true values given by $g_1(t)$ and $g_2(t)$. The estimated values trace the true curves well. The correlation between the true and estimated values for μ_t is 0.993 and for β_t it is 0.953. The SMSE for μ_t is 0.157 and for β_t it is 0.057.

Table 16: Exploratory TVEM-DSEM

Parameter	True Value	Estimate(95% Credibility Interval)
ϕ	0.5	0.523(0.496,0.549)
θ_w	0.5	0.541(0.462,0.617)
θ_b	0.5	0.537(0.461,0.628)
ψ	1.2	1.155(1.060,1.250)

Figure 3. Estimated v.s. True value for μ_t

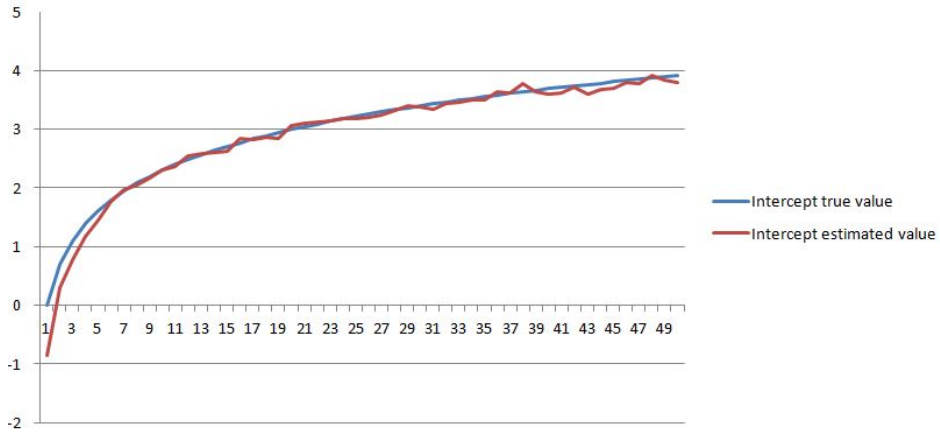
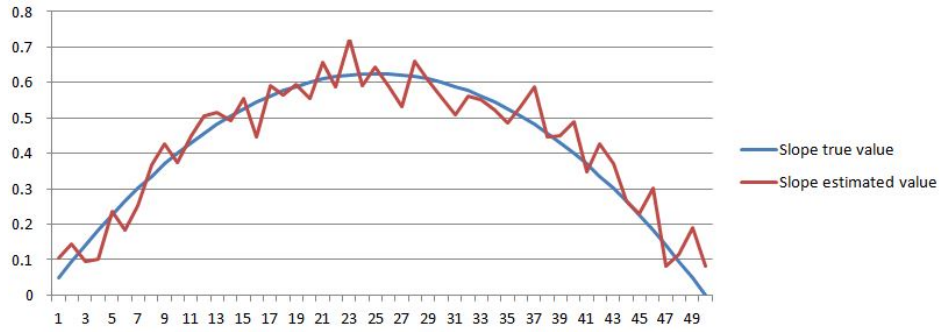


Figure 4. Estimated v.s. True value for β_t



Given the clear trends that are established from the exploratory TVEM-DSEM analysis, the next step of the analysis is to incorporate these trends

in the DSEM model by creating predictors of μ_t and β_t that account for the trends. The predictors are essentially smoothing curves for the estimated values obtained in the exploratory analysis. Such curves can be constructed through a separate algorithm, using the estimated μ_t and β_t values, or using multiple imputed values for μ_t and β_t. The smoothing can be done through polynomial functions or splines as in Buja, Hastie, and Tibshirani (1989). These smoothed curves can be entered into the DSEM model as predictors of μ_t and β_t.

Alternatively the smoothing can be performed within the DSEM framework as follows. We add time-specific predictors of μ_t and β_t based on the shapes of the trends. Given the estimated values we add $log(t)$ as the predictor for μ_t and t and t^2 as the predictors of β_t so that it is modeled as a quadratic function. Thus we augment the model given in (82) and (83) with the following two equations with some added scaling for the predictors.

$$\mu_t = a_1 + a_2 log(t) + \xi_{1,t} \tag{86}$$

$$\beta_t = a_3 + a_4(0.05t) + a_5(0.001t^2) + \xi_{2,t} \tag{87}$$

The results of this analysis are presented in Table 17. All parameters are estimated well and the true values are within the credibility intervals. The only exception is the $Var(\xi_{1,t})$ parameter where the lower end of the credibility interval is 0.002, slightly above the true value of 0, but clearly there is no support for a meaningful non-zero variance. The estimated random effects for μ_t and β_t are plotted against the true values in Figure 5 and 6. Clearly the estimates are improved particularly for the β_t values. The correlation between the true and estimated values for μ_t is 0.999 and for β_t it is 0.997. The SMSE for μ_t is 0.133 and for β_t it is 0.019.

Table 17: TVEM-DSEM accounting for the trends

Parameter	True Value	Estimate(95% Credibility Interval)
ϕ	0.5	0.516(0.482,0.542)
θ_w	0.5	0.522(0.417,0.600)
θ_b	0.5	0.540(0.465,0.627)
ψ	1.2	1.179(1.081,1.307)
a_1	0	-0.390(-0.516,0.021)
a_2	1	1.114(0.985,1.148)
a_3	0	-0.023(-0.077,0.031)
a_4	1	1.027(0.935,1.126)
a_5	-1	-1.005(-1.099,-0.917)
$Var(\xi_{1,t})$	0	0.005(0.002,0.014)
$Var(\xi_{2,t})$	0	0.001(0.000,0.002)

Figure 5. Estimated v.s. True value for μ_t accounting for the trends

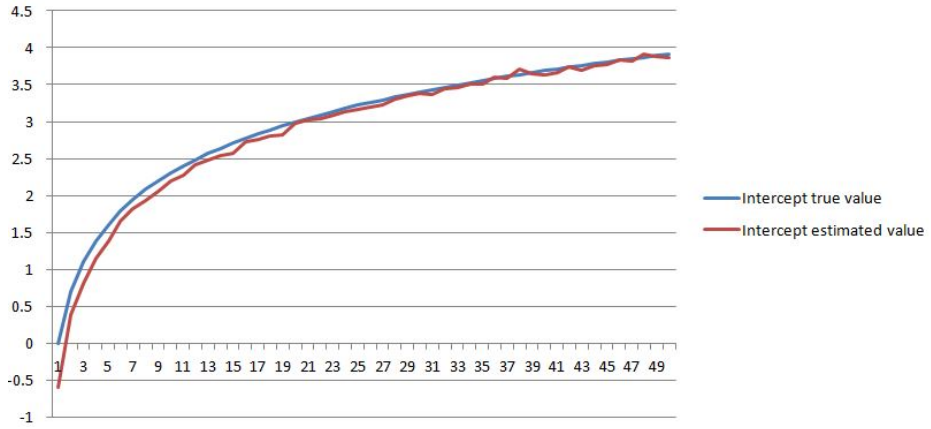
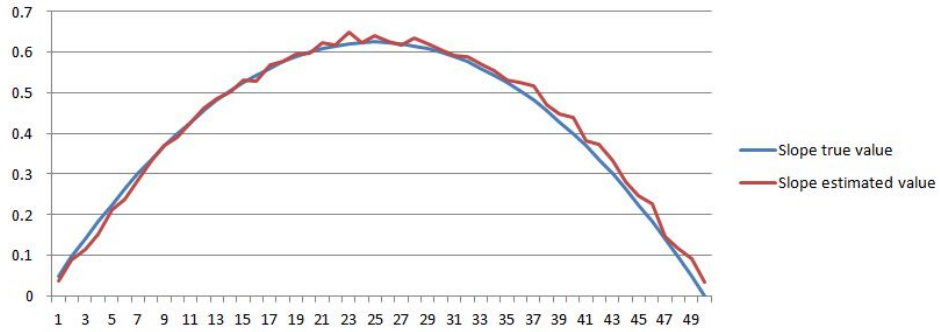


Figure 6. Estimated v.s. True value for β_t accounting for the trends



Given that the time-specific random effects have nearly zero residual variance, we can remove the random effects $\xi_{1,t}$ and $\xi_{2,t}$ from Equations (86) and (87). If we do so the model can be estimated simply as a two-level DSEM model rather than a cross-classified DSEM model as follows

$$Y_{it} = a_1 + a_2 log(t) + Y_i + (a_3 + a_4(0.05t) + a_5(0.001t^2))X_{it} + f_{it} + \varepsilon_{it} \quad (88)$$

$$f_{it} = \phi f_{i,t-1} + \xi_{it}. \quad (89)$$

The coefficients a_4 and a_5 are the interaction effects of X_{it} with t and t^2. The results for this analysis are presented in Table 18. All parameter estimates are very close to the true values. Note that in this model the effects μ_t and β_t are now smooth curves with no error term, which is how we generated the data. Because the parameter estimates are so close to the true values these curves are virtually indistinguishable from the true value curves. The correlation for both estimated effect v.s. true value is 1 and the SMSE are now further reduced to 0.021 and 0.017.

6 Conclusion

The DSEM framework builds on the econometric literature and advancements in time series modeling as well as the progress that has been made previously in single-level dynamic structural modeling as well as the progress that has been made in the area of multilevel structural equation modeling. The DSEM framework allows us to combine time series models for a population of subjects. One of the strengths of the framework is that it allows subject-specific structural and autoregressive parameters. These parameters

Table 18: Two-level TVEM-DSEM accounting for the trends

Parameter	True Value	Estimate(95% Credibility Interval)
ϕ	0.5	0.528(0.503,0.554)
θ_w	0.5	0.555(0.487,0.628)
θ_b	0.5	0.536(0.462,0.626)
ψ	1.2	1.136(1.043,1.224)
a_1	0	-0.030(-0.136,0.094)
a_2	1	1.003(0.967,1.028)
a_3	0	-0.020(-0.068,0.026)
a_4	1	1.025(0.941,1.111)
a_5	-1	-1.004(-1.088,-0.923)

can be used further for structural modeling on the population level, i.e., they can be predicted by subject-specific variables or they can be used as predictors of other such variables. Just as important is the total opposite. In the DSEM framework autoregressive and structural parameters can be chosen to be non-random, i.e., invariant across subjects in the population. When the number of time points is within the mid length range of 10 to 100, which is the most common range in the social sciences, parameters invariant across subjects are essential in expanding model complexity beyond what is accessible with single-level DSEM models. The inclusion of non-random parameters gives us the ability to combine data across the population to obtain more accurate time-series and structural parameters. But perhaps the real strength of DSEM is the fact that it seamlessly can accommodate random and non-random parameters at the same time, not just to improve the quality of the estimation and the quality of the statistical methodology effort to match the data and the models, but also to use data analysis to find answers to real life questions hidden in the data.

References

[1] Arminger, G. and Muthen, B. (1998) A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. Psychometrika, 63, 271-300.

[2] Asparouhov, T. & Muthén, B. (2016). General random effect latent variable modeling: Random subjects, items, contexts, and parameters. Forthcoming in the edited book "Advances in Multilevel Modeling for Educational Research: Addressing Practical Issues Found in Real-World Applications", 155-182. Information Age Publishing.

[3] Asparouhov, T. & Muthén, B. (2010). Bayesian analysis using Mplus: Technical implementation. Technical Report. Version 3. http://statmodel.com/download/Bayes3.pdf

[4] Buja, A. Hastie, T. J., Tibshirani, R. (1989). Linear Smoothers and Additive Models, Annals of Statistics, 17, 453-555.

[5] Browne, W & Draper, D. (2006) A comparison of Bayesian and likelihood-based methods for fitting multilevel models. Bayesian Analysis, 1, 473-514.

[6] Celeux, G., Forbes, F., Robert, C. P. & Titterington, D. M. (2006). Deviance informaiton criteria for missing data models. Bayesian Analysis, 1, 651-674.

[7] Deboeck, P. & Preacher, K. (2016) No Need to be Discrete: A Method for Continuous Time Mediation Analysis. Structural Equation Modeling, 23, 61-75.

[8] Granger, C.W.J. & Morris, M.J. (1976). Time series modelling and interpretation. Journal of the Royal Statistical Society, Series A, 139, 246-257.

[9] Greene, W. H. (2014). Econometric Analysis (7th Edition). Prentice Hall. New Jersey.

[10] Hamaker, E.L. (2005) Conditions for the equivalence of the autoregressive latent trajectory model and a latent growth curve model with autoregressive disturbances. Sociological Methods and Research, 33, 404 - 418.

[11] Hamaker E.L. and Grasman R.P.P.P. (2015) To center or not to center? Investigating inertia with a multilevel autoregressive model. Frontiers in Psychology., 5, 1492.

[12] Jahng S., Wood, P. K.,& Trull, T. J., (2008) Analysis of Affective Instability in Ecological Momentary Assessment: Indices Using Successive Difference and Group Comparison via Multilevel Modeling. Psychological Methods, 13, 354-375.

[13] Jongerling J, Laurenceau J.P., Hamaker E. (2015). A Multilevel AR(1) Model: Allowing for Inter-Individual Differences in Trait-Scores, Inertia, and Innovation Variance. Multivariate Behavioral Research, 50, 334-349.

[14] Lee S.Y. (2007) Structural Equation Modelling: A Bayesian Approach, London: John Wiley & Sons.

[15] Ludtke, O., Marsh, H.W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. Psychological Methods, 13, 203-229.

[16] Molenaar, P.C.M. (2004). A Manifesto on Psychology as Idiographic Science: Bringing the Person Back Into Scientific Psychology, This Time Forever. Measurement: Interdisciplinary Research and Perspectives, 2, 201-218.

[17] Molenaar, P.C.M. (2017) Equivalent Dynamic Models. Multivariate Behavioral Research, 52, 1-17.

[18] Nickell, S. (1981). Biases in dynamic models with fixed effects. Econometrica: Journal of the Econometric Society, 1417-1426.

[19] Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed.). Thousand Oaks: Sage Publications.

[20] Schuurman N., Houtveen J., Hamaker E. (2015) Incorporating measurement error in n = 1 psychological autoregressive modeling. Frontiers in Psychology, 6, 1038.

[21] Trull, T. & Ebner-Priemer, U. (2014) The Role of Ambulatory Assessment in Psychological Science, Current Directions in Psychological Science, 23, 466-470.

[22] Vaida, F. & Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. Biometrika, 92, 351-370.

[23] Zhang Z. and Nesselroade J. (2007) Bayesian Estimation of Categorical Dynamic Factor Models, Multivariate Behavioral Research, 42, 729-756.

[24] Zhang Z., Hamaker E. and Nesselroade J. (2008) Comparisons of Four Methods for Estimating a Dynamic Factor Model, Structural Equation Modeling, 15, 377-402.

7 Appendix A: Continuous time DSEM modeling

In this section we describe the algorithm implemented in Mplus for approximating a continuous time DSEM model with a discrete time DSEM model. Every continuous function $f(t)$ can be approximated by a step function. Let δ be a small number. The function $f(t)$ can be approximated by a step function $f_0(t) = f_j = f(j \cdot \delta)$ when $j \cdot \delta - \delta/2 < t \leq j \cdot \delta + \delta/2$. The smaller the step interval δ the better the approximation. Based on this same principle we can approximate a continuous time DSEM model with a discrete time DSEM model.

7.1 Step 1: Rescaling the time variable

Suppose that individual i is observed at times t_{ij}, for $j = 1, ..., T_i$. We replace the value t_{ij} with an integer value $\hat{t}_{ij} = [t_{ij}/\delta]$, where $[t]$ denotes the smallest integer value not smaller than t, i.e., $\hat{t}_{ij}$ is the integer value for which

$$(\hat{t}_{ij} - 1)\delta < t_{ij} \leq \hat{t}_{ij}\delta. \tag{90}$$

Essentially, we first rescale the time variable by multiplying it by $1/\delta$ and then rounding it up to the nearest integer. Thus for all t_{ij} falling in the interval $(0, \delta]$, $\hat{t}_{ij} = 1$, for all t_{ij} falling in $(\delta, 2\delta]$, $\hat{t}_{ij} = 2$ and so on. Using this approach we convert any real time values t_{ij} to the integer time values $\hat{t}_{ij}$. At that point the standard DSEM modeling can be used. For all integer values that are not observed, missing data is assumed, that is, for individual i and integer time value t which is not equal to any of the $\hat{t}_{ij}$ we assume that the data is missing or not recorded. This is not really an assumption but is a way to properly record the data so that the observations are recorded for every integer.

If the δ value in the above algorithm is not sufficiently small it is very likely that two or more t_{ij} values for individual i will appear in the interval $((n-1)\delta, n\delta]$. This will result in several values $\hat{t}_{ij}$ being assigned the value n which is not an acceptable outcome as we can use just one observation for time n. To resolve this problem we apply the following algorithm. For individual i all t_{ij} are placed in the intervals $((n-1)\delta, n\delta]$ following (90). Starting with the smallest n for which the interval $((n-1)\delta, n\delta]$ contains multiple values, we determine the closest empty interval to that interval

and we shift one of the overflow values towards that interval, preserving the original order of t_{ij}. That means that each interval from the overflow interval to the empty interval shifts one value in the direction of the overflow interval. This algorithm approximately minimizes

$$\sum_j (\hat{t}_{ij} - t_{ij}/\delta)^2 \tag{91}$$

over integer and unequal values $\hat{t}_{ij}$ in most common situations. In some situations the above algorithm won't quite minimize the above objective function but it will come fairly close to minimizing it. Full minimization may be too intricate to accomplish in general because of the discrete optimization space. This algorithm as implemented in Mplus would report $max|\hat{t}_{ij} - t_{ij}/\delta|$ if this quantity is greater than 5, which means that an observation had to be shifted more than 5 intervals away from it original assignment. This would suggest that the discretized grid constructed for that value of δ is too crude to be considered a good approximation and a smaller value of δ should be used.

7.2 Step 2: Time shift transformation

The next step of the time transformation is a time shift transformation. There is a fundamental difference between the cross-classified DSEM model and the two-level DSEM model that comes into play here. In cross-classified DSEM models we estimate time-specific effects and this can be meaningful only if the time scale is aligned between individuals. In cross-classified DSEM, time t for individual $i = 1$ should have the same meaning as as time t for individual $i = 2$, for example, the number of days since an intervention that both individuals received, so that the same time-specific random effect $s_{t,3}$ applies. Such an alignment of time is not needed for the two-level DSEM model and this is why the time shift transformation is different for the two models.

For cross-classified DSEM models we compute $T_0 = min_{i,j}(\hat{t}_{ij})$ and we shift the time so that we start at 1, $\hat{\hat{t}}_{ij} = \hat{t}_{ij} - T_0 + 1$. At least one individual is observed at time 1 and this is the earliest time an observation was made in the sample. Missing values are recorded for all individuals and time points not in the set $\hat{\hat{t}}_{ij}$. For each individual the missing values beyond the last

observed value are not analyzed. This time-shift is done differently for two-level DSEM models. We compute $T_{0i} = min_j(\hat{t}_{ij})$, i.e., we find the first observed value for each individual i and shift each individual by that value so that every individual starts at 1, i.e., $\hat{\hat{t}}_{ij} = \hat{t}_{ij} - T_{0i} + 1$. This minimizes that amount of missing data that will have to analyzed and imputed in the MCMC estimation. Again all missing data after the last observed value is not analyzed. The difference in the time shift transformation is that in the cross-classified model we shift the time uniformly across all individuals while in the two-level model the time scale is shifted for each individual separately.

7.3 How to choose δ

The transformation is determined by time interval δ. The smaller this value is, the more precise the approximation. However, the smaller the value is the more missing data will be interspersed between the observed data. This will cause the MCMC sequence to converge slower. It will also cause the model to lose some precision. Consider for example trying to estimate the daily autocorrelation ϕ_d by first estimating the hourly autocorrelation ϕ_h using an AR(1) model. The relationship between the two is given by $\phi_d = \phi_h^{24}$. If $\phi_d = 0.75$ then $\phi_h = 0.988$. A small error in the estimation of ϕ_h, say 0.987, results in bigger error for ϕ_d as it will be estimated to 0.73. Thus model imprecision is amplified for smaller δ.

The selection of δ should be driven by three principles. First is the interpretability. Using natural δ values such as an hour, a day, a 2-day interval, a week, a month would improve the interpretability as supposed to say a time metric such as 1.3 days. The second consideration is the amount of missing data resulting in this process. The missing data should be no more than 90% to 95% of the data. More missing data than that will likely yield a slow converging MCMC estimation which potentially can produce bigger error in the estimation than the discrete time approximation for larger δ values. The third consideration should be that δ needs to be small enough so that the original times are well approximated. Using a large value of δ will result in $\hat{\hat{t}}_{ij} = j$ in two-level DSEM models, that is, the information in the original times t_{ij} is completely ignored and all observations are assumed equally spaced.

There is one further consideration that applies only to the cross-classified DSEM model. The smaller the δ value is the more time periods there will

be. Since the DSEM model estimates time-specific random effects for each interval, it is desirable that each period has at least several observations, which act as measurements for the time-specific effects. A simple rule of thumb would be to have at least 3 observations per random effect at each time point. Thus the δ value should not be chosen to be so small as to reduce the number of observations below that level.

8 Appendix B: DSEM model estimation

Here we describe the conditional distributions for each of the 13 blocks given in Section 3. These are used in the MCMC estimation to update each block. The conditional distributions we are interested in are the conditional distributions for each block conditional on all other blocks and the data.

Consider the conditional distribution of B1. Given that we condition on B3 the variables $Y_{3,t}$ are considered known. All other random and non-random slopes and loadings are also considered known. Let $Y'_{1,it} = Y_{it} - Y_{3,t}$. Equation (6) can be expressed as

$$Y'_{1,it} - Y_{2,i} = (I - R_0)^{-1}\nu_1 + \sum_{l=0}^{L}(I - R_0)^{-1}\Lambda_{1,l}\eta_{1,i,t-l} + \sum_{l=1}^{L}(I - R_0)^{-1}R_l(Y'_{1,i,t-l} - Y_{2,i}) + \sum_{l=0}^{L}(I - R_0)^{-1}K_{1,l}X_{1,i,t-l} + (I - R_0)^{-1}\varepsilon_{1,it} \tag{92}$$

or equivalently

$$Y'_{1,it} - (I - \sum_{l=1}^{L}(I - R_0)^{-1}R_l)Y_{2,i} = (I - R_0)^{-1}\nu_1 + \sum_{l=0}^{L}(I - R_0)^{-1}\Lambda_{1,l}\eta_{1,i,t-l} + \sum_{l=1}^{L}(I - R_0)^{-1}R_lY'_{1,i,t-l} + \sum_{l=0}^{L}(I - R_0)^{-1}K_{1,l}X_{1,i,t-l} + (I - R_0)^{-1}\varepsilon_{1,it} \tag{93}$$

where I denotes the identity matrix. The conditional distribution of $Y_{2,i}$ is now determined by the log-likelihood of the above equation in conjunction with Equation (2). Denote $F(Y_{2,i})$

$$F(Y_{2,i}) = \sum_t L(Y'_{1,it}|*) + L(Y_{2,i}|*), \tag{94}$$

where $L(Y'_{1,it}|*)$ is the log-likelihood expression of Equation (93) and $L(Y_{2,i}|*)$ is the likelihood expression of Equation (2). Since all these equations are for normal distributions, the conditional distribution of $Y_{2,i}$ is given by

$$Y_{2,i} \sim N(F''^{-1}F'(0), F''^{-1}), \tag{95}$$

where F' and F'' denote the first and the second derivative of the log-likelihood function F. Note that since F is a quadratic function of $Y_{2,i}$ the second derivative is a constant matrix that does not depend on the value of $Y_{2,i}$. Another way to compute this is as follows. Let $M = (I - \sum_{l=1}^{L}(I - R_0)^{-1}R_l)$. The conditional distribution of $MY_{2,i}$ can be computed as follows. The variable $MY_{2,i}$ is the random intercept of a two-level model where the within-level model is given by (93) and the between-level model is given by (2) multiplied by M so that $MY_{2,i}$ is the dependent variable on the between-level as well. The conditional distribution of the random intercept in a standard two-level model is well-known. If the conditional distribution of $MY_{2,i}$ is $N(m, v)$ then the conditional distribution of $Y_{2,i}$ is $N(M^{-1}m, M^{-1}v(M^{-1})^T)$.

The conditional distribution of B2 is similar. Conditional on all other blocks, $Y_{2,i}$ and $Y_{3,t}$ are considered known, which means that $Y_{1,it}$ is known, as well as $\eta_{1,it}$. The joint conditional distribution of all $s_{2,i}$ come from Equations (9) and (10) as well as a reformulation of Equation (3) which expresses $s_{2,i}$ as a dependent variable on the left hand side and other variables on the right hand side. Denote again by F the log-likelihood function

$$F(s_{2,i}) = \sum_t L(Y_{1,it}|*) + \sum_t L(\eta_{1,it}|*) + L(s_{2,i}|*), \tag{96}$$

where $L(Y_{1,it}|*)$ is the log-likelihood contribution of Equation (9), written directly as it is expressed in that equation, $L(\eta_{1,it}|*)$ is the log-likelihood contribution of Equation (10), also written directly as it is expressed in that equation and $L(s_{2,i}|*)$ is the log-likelihood contribution of Equation (3). All these distributions are normal and thus the function F is again a quadratic function of $s_{2,i}$ and the conditional distribution can be obtained as in (95)

$$s_{2,i} \sim N(F''^{-1}F'(0), F''^{-1}). \tag{97}$$

There is a key assumption in this procedure, which can be viewed also as a model restriction. Equations (9) and (10) can be used directly to write the likelihood only under the assumption that there are no non-recursive interactions in the model. That is to say that the dependent variables $Y_{1,it}$

can not appear in a cyclical fashion in these equations, i.e., no two components of that vector, say, $Y_{1,it1}$ and $Y_{1,it2}$ can simultaneously be predictors of each other. Also longer cyclical regressions involving 3 or more variables can not appear in the model. Such a restriction is needed to preserve the quadratic form of F and to preserve the integrity of the likelihood obtained directly from these equations. If the equations are non-recursive then F is not quadratic and those equations can not be used directly to write the log-likelihood. When the equations are recursive they can be ordered in such a way that the $[Y_{1,it1}|Y_{1,it2}, Y_{1,it3}...][Y_{1,it2}|Y_{1,it3}...]...$ conditional distributions are expressed precisely by Equation (9). The same applies to Equation (10).

The conditional distribution of block B3 is slightly more complicated than the conditional distribution of block B1. Let $Y'_{1,it} = Y_{it} - Y_{2,i}$ Equation (6) can be expressed as

$$Y'_{1,it} - Y_{3,t} = (I - R_0)^{-1}\nu_1 + \sum_{l=0}^{L}(I - R_0)^{-1}\Lambda_{1,l}\eta_{1,i,t-l} + \sum_{l=1}^{L}(I-R_0)^{-1}R_l(Y'_{1,i,t-l} - Y_{3,t-l}) + \sum_{l=0}^{L}(I-R_0)^{-1}K_{1,l}X_{1,i,t-l} + (I-R_0)^{-1}\varepsilon_{1,it} \tag{98}$$

or equivalently

$$Y'_{1,it} - (Y_{3,t} - \sum_{l=1}^{L}(I-R_0)^{-1}R_l Y_{3,t-l}) = (I-R_0)^{-1}\nu_1 + \sum_{l=0}^{L}(I-R_0)^{-1}\Lambda_{1,l}\eta_{1,i,t-l} + \sum_{l=1}^{L}(I - R_0)^{-1}R_l Y'_{1,i,t-l} + \sum_{l=0}^{L}(I - R_0)^{-1}K_{1,l}X_{1,i,t-l} + (I - R_0)^{-1}\varepsilon_{1,it} \tag{99}$$

It is clear from this equation that the conditional distribution of $Y_{3,t}$ is determined not just by the above equation at time t, i.e., the level 3 cluster at time t but also by the above equation at times $t + 1$, ..., $t + L$. It is also clear that the conditional distribution of Y_{3,t_1} is not independent of the conditional distribution of Y_{3,t_2}. Therefore computing the joint distribution of all $Y_{3,t}$ becomes computationally infeasible. We resolve this problem by breaking down block B3 into separate blocks, one for each time t and we consider the conditional distribution of $Y_{3,t}$ not just conditioned on all other

blocks but also on all other $Y_{3,t'}$ where $t' \neq t$. Denote by $F(Y_{3,t})$

$$F(Y_{3,t}) = \sum_{l=0}^{L} \sum_{i} L(Y'_{1,i,t+l}|*) + L(Y_{3,t}|*), \tag{100}$$

where $L(Y'_{1,it}|*)$ is the log-likelihood expression of Equation (99) and $L(Y_{3,t}|*)$ is the likelihood expression of Equation (4). Since all these equations are for normal distributions the conditional distribution of $Y_{3,t}$ is given by

$$Y_{3,t} \sim N(F''^{-1}F'(0), F''^{-1}), \tag{101}$$

where F' and F'' denote the first and the second derivative of the log-likelihood function F.

Another way to compute this posterior distribution is as follows. Denote by $B_0 = I$, $B_l = -(I - R_0)^{-1}R_l$. The random intercept of (99) is $A_t = \sum_{l=0}^{L} B_l Y_{3,t-l}$. Suppose that the conditional distribution of that random intercept A_t computed from the data in that cluster is $N(m_t, v_t)$, excluding a between-level model. The conditional distribution of $Y_{3,t}$, conditional on all other $Y_{3,t'}$ where $t' \neq t$ is given by

$$Y_{3,t} \sim N(Dd, D), \tag{102}$$

where

$$D = \left(\Sigma_3^{-1} + \sum_{l=0}^{L} B_l^T v_{t+l}^{-1} B_l \right)^{-1} \tag{103}$$

$$d = \Sigma_3^{-1}\mu_3 + \sum_{l=0}^{L} B_l^T v_{t+l}^{-1} \left(m_{t+l} - \sum_{n=0, n\neq l}^{L} B_l Y_{3,t+l-n} \right), \tag{104}$$

where $N(\mu_3, \Sigma_3)$ is the implied distribution for $Y_{3,t}$ from Equation (4). The above equations apply for $t \leq T - L$ where $T = max(T_i)$. When $t > T - L$ the equations get reduced by $L - T + t$ because the index of the equation is greater than the largest t in the model, i.e., equations with time index greater than T do not exist as no data is observed beyond time T.

The conditional distributions of block B4 is obtained the same way as the conditional distribution of block B2. Level 2 and level 3 simply reverse roles. The conditional distribution of block B5 is as in Step 1 in Section 2.4 in Asparouhov and Muthén (2010). Conditional on blocks B1-B4 and the generated values for these variables, the level 2 and level 3 models become

essentially like multiple groups in single-level modeling. The two levels are independent of each other, the within-level model, and the observed data Y_{it}. Therefore the single-level approach in Asparouhov and Muthén (2010) applies. We reproduce this step here for completeness. Consider the single-level SEM model

$$y = \nu + \Lambda\eta + Kx + \varepsilon \tag{105}$$

$$\eta = \alpha + B\eta + \Gamma x + \zeta \tag{106}$$

The conditional distribution is given by

$$[\eta|*] \sim N(Dd, D), \tag{107}$$

where

$$D = \left(\Lambda^T\Theta^{-1}\Lambda + \Psi_0^{-1}\right)^{-1} \tag{108}$$

$$d = \Lambda^T\Theta^{-1}(y - \nu - Kx) + \Psi_0^{-1}B_0^{-1}(\alpha + \Gamma x), \tag{109}$$

where $B_0 = I - B$, I is the identity matrix, $\Theta = Var(\varepsilon)$, and $\Psi_0 = B_0^{-1}Var(\zeta)(B_0^{-1})^T$.

The conditional distribution of block B6 requires some additional computations. Because $\eta_{1,it}$ are not independent across time they can not be generated simultaneously in an efficient manner as that will require computing the large joint conditional distribution of $\eta_{1,it}$ for all t. Therefore block B6 is essentially split into separate blocks, one for each t. Thus we update the within-level latent variable one at a time starting at $\eta_{1,i,1-L}$, $\eta_{1,i,2-L}$,...,η_{1,i,T_i}, where T_i is the last observation for individual i. We need to construct the conditional distribution of $\eta_{1,it}$ conditional on all the other blocks and all the other $\eta_{1,it}$, for times different from t. Given all the other blocks, $Y_{1,it}$ is observed. The conditional distribution of $\eta_{1,it}$ is somewhat different at the end and at the beginning of that sequence so first we consider the case where t is in the middle, more specifically $0 < t < T_i - L$. The latent variable $\eta_{1,it}$ conditional distribution is determined by Equation (6) and (7) at time t, $t+1$, ..., $t+L$. In total there are $2L+2$ equations that affect the conditional distribution. We combine all these equations into one big model for $\eta_{1,it}$, that consists of one structural equation: (7) at time t, and $2L+1$ measurement equations for $\eta_{1,it}$: Equation (6) at time t and Equations (6) and (7) at times $t+1$, ..., $t+L$. Using this larger model the conditional distribution is obtained again as in Equation (107). For $t > T_i - L$ the conditional distribution

is obtained similarly. However, since there are no observations beyond time T_i there will be only $2(T_i - t + 1)$ equations in the enlarged model, again one structural equation and $2(T_i - t) + 1$ measurement equations. For $t \leq 0$ we also have fewer equations due to the fact that there are no observations before $t = 1$. There are only $2(L + t)$ measurement equations in the model, and the prior specification for $\eta_{1,it}$ for $t \leq 0$ takes the role of the structural equation.

In block B7 similar considerations are taken into account. Missing values are imputed one at a time and in a sequential order. Block B8 is actually done at the same time as block B7 because one can interpret the initial conditions as missing values, i.e., at times $t \leq 0$, $Y_{1,it}$ and $X_{1,it}$ can be viewed as missing values. Note here that conditional on all other blocks, the missing values of Y_{it} are essentially the missing values of $Y_{1,it}$, that is once the missing values for $Y_{1,it}$ are imputed, the values of Y_{it} are obtained by (1) since $Y_{2,i}$ and $Y_{3,t}$ are known, i.e., are conditioned on. Let's first consider the missing value $Y_{1,it}$ in the middle of the sequence $0 < t < T_i - L$. In non-time series analysis we impute the missing value from the univariate conditional normal distribution obtained from the within-level model, see Section 4 in Asparouhov and Muthén (2010). Since conditional on $X_{1,it}$ the multivariate joint distribution of $Y_{1,it}$ and $\eta_{1,it}$ is normal then one of these variables conditional on all other variables has a univariate normal distribution, and that distribution is used for missing value imputation. However, in the time-series model we consider here $Y_{1,it}$ variable is used in $2L+2$ Equations (6) and (7) at times t, $t+1$, ..., $t + L$. A missing variable in this context is nothing more than a unobserved latent variable. Therefore the procedure we outlined above for conditional distribution for block B6 applies here as well. As in block B6 at the end of the sequence for $t > T_i - L$ or at the beginning of the sequence for $t \leq 0$ the number of equations used for the computation decreases and for $t \leq 0$ the structural equation, where the missing value is the dependent variable is replaced by the prior specification. This applies both for $Y_{1,it}$ and $X_{1,it}$ when $t \leq 0$. Note that the missing data treatment is likelihood based and thus will guarantee consistent estimation as long as the missing data is MAR.

Blocks B9-B12 are all implemented as in Asparouhov and Muthén (2010). Conditional on all latent variable the DSEM model is essentially a 3-group single-level SEM and the procedures for single-level SEM apply directly. Finally let's consider block B13. The conditional distribution of the random effects from Equation (11) are not explicit and we use the Metropolis-Hastings algorithm to generate values from that distribution. Suppose that $Y_{1,it1}$ has a

random residual variance $\sigma_i = Exp(s_{2,i})$, where $s_{2,i}$ is a normally distributed random effect. Suppose that the current value of that random effect is s_0. A new proposed value s_1 is drawn from a normal distribution $N(s_0, V)$ where V is referred to as the proposal distribution variance. We then compute the acceptance ratio as follows

$$R = \frac{P(s_{2,i} = s_1|*) \prod_t P(Y_{1,it1}|\sigma_i = Exp(s_1))}{P(s_{2,i} = s_0|*) \prod_t P(Y_{1,it1}|\sigma_i = Exp(s_0))}, \tag{110}$$

where $P(s_{2,i} = s_j|*)$ is the likelihood of $s_{2,i}$ obtained from Equation (3) conditional on all other variables in that equation and $P(Y_{1,it1}|\sigma_i = Exp(s_j))$ is the likelihood of $Y_{1,it1}$ obtained from Equation (6) conditional on all other variables in that equation. The proposed value s_1 is accepted with probability $min(1, R)$. If the value is rejected the old value s_0 is retained. The proposal distribution variance V is chosen to be a small value such as 0.1 and is adjusted during a burnin stage of the estimation to obtain optimal mixing, i.e., optimal acceptance rate in the Metropolis-Hastings algorithm. The optimal acceptance rate is considered to be between .25 and 0.50. To preserve the integrity of the MCMC chain the jumping distribution variance is not changed beyond the burnin iterations and those iterations are discarded and not used in the posterior distribution. Under these conditions the above Metropolis-Hastings algorithm generates $s_{2,i}$ from the correct conditional distribution. Random variances for latent factors are estimated similarly. This concludes the description of the MCMC estimation of the DSEM model.

What is hidden in the above description of the estimation is the computational times to estimate the model. Depending on the particular details of the model, the conditional distributions may or may not be invariant across subject or invariant across time. The more random structural parameters there are that vary across subject and time, the less invariance there is in the conditional distributions described above. Generally speaking for the two-level DSEM model most of the conditional distributions are invariant across time. Thus the conditional means and variances depend only on sufficient statistics of the data and are easily computed. For the cross-classified DSEM model even when a single structural parameter varies across time and subject, the structural SEM model given in Equations (9) and (10) changes for every i and t and a separate computation is required. This generally results in substantial increase in the computational time. Paired with the slower convergence, that stems from the fact that the model is more flexible and

the cross-classified DSEM model can become substantially more computationally intensive than a two-level DSEM model.

9 Appendix C. Computing the model-estimated subject-specific means and variances

In this section we provide details on how model-estimated subject-specific means and variances can be computed for the DSEM model. The main assumption in such a computation is the assumption of stationarity. Any autoregressive process in the model has to be stationary, that is, over time the distribution of the variables in the autoregressive process stabilizes.

To compute the subject-specific model-estimated mean and variance implied by the two-level DSEM model we start with Equation (1) assuming no time-specific component $Y_{3,t}$, that is,

$$E(Y_{it}|i) = E(Y_{1,it}|i) + Y_{2,i} \tag{111}$$

$$Var(Y_{it}|i) = Var(Y_{1,it}|i). \tag{112}$$

The estimated subject-specific variance is simply the estimated within-level subject-specific variance while the estimated subject-specific mean is the sum of $Y_{2,i}$, which is estimated within the MCMC estimation, and the within-level estimated mean. Thus, we can focus on Equation (6) and (7).

Let Z represent the variables in these equations that are involved in an autoregressive model. Let's assume the following autoregressive model for Z_t

$$Z_t = \mu + \sum_{l=1}^{L} A_l Z_{t-l} + \zeta, \tag{113}$$

where $\Sigma = Var(\zeta)$. Assuming stationarity of this model, the mean of Z_t is

$$E(Z_t) = \left(I - \sum_{l=1}^{L} A_l \right)^{-1} \mu \tag{114}$$

Let $\Gamma_j = Cov(Z_t, Z_{t-j})$. The variance of Z_t, Γ_0, is computed from the Yule-

Walker equations, see Greene (2014)

$$\begin{bmatrix} \Gamma_0 & \Gamma_1^T & \Gamma_2^T & ... & \Gamma_L^T \\ \Gamma_1 & \Gamma_0 & \Gamma_1^T & ... & \Gamma_{L-1}^T \\ \Gamma_2 & \Gamma_1 & \Gamma_0 & ... & \Gamma_{L-2}^T \\ ... & ... & ... & ... & ... \\ \Gamma_L & \Gamma_{L-1} & \Gamma_{L-2} & ... & \Gamma_0 \end{bmatrix} \begin{bmatrix} I \\ -A_1^T \\ -A_2^T \\ ... \\ -A_L^T \end{bmatrix} = \begin{bmatrix} \Sigma \\ 0 \\ 0 \\ ... \\ 0 \end{bmatrix} \tag{115}$$

These equations can been used to compute the model parameters A_j from the sample autocovariances Γ_j, however, we do the opposite. As the model parameters are known, we solve these equations for the model-implied Γ_j, which have a total of $Lp^2 + p(p+1)/2$ parameters, where p is the size of the vector Z. The above system is over identified as it has $(L+1)p^2$ equations. To make it just identified we remove the $p(p-1)/2$ upper diagonal of the first row. Note that this method yields not just the model-estimated variance for the dependent and latent variables but also the model-estimated autocorrelations of lags $1, ..., L$.

If there is a trend in the data it should be modeled outside of the autoregressive process for the Yule-Walker computations to apply. In Mplus the Yule-Walker computation is done within the residual output option. To be more clear, model estimation does not require the trend to be modeled outside of the autoregressive process. In fact in some cases, such as linear growth, modeling the trend within the autoregressive process or outside of the autoregressive process make no difference, see Section 5.4, and the models are equivalent reparameterizations of each other. But the Yule-Walker computation outlined above does require model trends to be outside of the autoregressive process because the autoregressive part of the DSEM model is assumed stationary.

Three Mplus output options are based on the Yule-Walker equations: tech4, residual and standardization, and therefore are only valid when the autoregressive part of the model is stationary. In some cases the Mplus program will automatically detect and report non-stationarity for some of the subjects in the population simply because the model-implied subject-specific variance estimates are negative or the model-implied subject-specific variance-covariance matrices are not positive definite. Even if the Yule-Walker equations produce positive definite variance-covariance matrices and the Mplus program does not produce non-stationarity warnings, the stationarity assumption may still not hold and that could results in incorrect

model-implied estimates for the means and variances. Thus stationarity assumption should be carefully inspected before the model-implied estimates are used.