Using Parametric and Residual-based Bootstrap to Assess the Absolute Goodness-of-fit for State Space Model

by

Fei Gu

ACKNOWLEDGEMENTS

I would like to thank Dr. Neal M. Kingston for his supervision throughout the entire process of this doctoral dissertation. I would also like to express my special thanks to Dr. Todd D. Little, the director of the Center for Research Methods and Data Analysis (CRMDA), for providing SAS Unix and the High Performance Computing (HPC) facility, and Dr. Paul Johnson, the associate director of the CRMDA, for teaching me how to use the HPC.

CHAPTER 1: INTRODUCTION

Analysis of interindividual variation versus analysis of intraindividual variation

A formidable collection of existing statistical models has been developed in the history of scientific research seeking results generalizable to a homogenous population, for example, analysis of variance, regression analysis, and factor analysis. In multilevel modeling (MLM), the population is assumed to be composed of different subpopulations, and subjects within each subpopulation are again considered homogenous. Apparently, the structure of interindividual variation has historically been the focus of statistical analysis, and there is a common assumption for all these interindividual approaches. That is, human individuals are homogenous in all relevant aspects. In other words, each individual has to obey exactly the same dynamical laws. Presumably, this focus can be attributed to the scientific ideal of finding nomothetic knowledge that should apply to all human individuals.

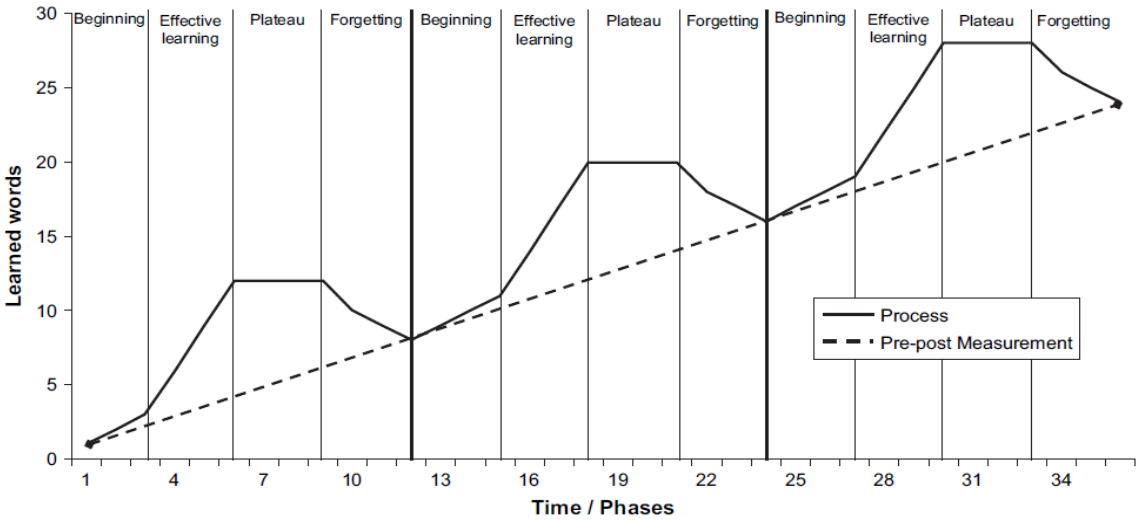
In test theory, the current dominant approach to statistical analysis "treats individual differences or, equivalently, the distribution of measurements over people" (Lord & Novick, 1968, p. 32). Particularly, the true and error scores primarily considered in test theory are those that deals with groups rather than individuals (selection rather than counseling). However, the original definitions of true and error scores rooted completely in the context of intraindividual variation. That is, "A mathematically convenient way of defining a true score is as the expected observed score with respect to the propensity distribution of a given person on a given measurement" and the propensity distribution "is a cumulative distribution function defined over repeated statistically independent measurements on the same person" (Lord & Novick, 1968, p.

30). The concept of error score comes straightforwardly: "The discrepancy between observed value and true value" (Lord & Novick, 1968, p. 31). The reason of this shifted treatment from the context of intraindividual variation to the basis of interindividual variation in classical test theory, as we know it, is that "it is not usually possible in psychology to obtain more than a few independent observations" (Lord & Novick, 1968, p. 30). Thus, it is immediately obvious that this raises a fundamentally important validity issue concerning the equivalence between the analysis of intraindividual variation and the analysis of interindividual analysis. As Molenaar (2003, 2004, 2008) articulated, the results obtained from one type of analysis are not generally equivalent to those obtained from the other type of analysis, unless there exists the very strict mathematical-statistical condition, i.e., ergodicity. As described by Van Rijn (2008), ergodicity is the condition that "the average of a stochastic process over time is equal to the average of the ensemble of stochastic processes at a single point in time" (p. 54). However, the ergodicity condition should never be expected in classical test theory that is based on a heterogeneous population (Molenaar, 2003, 2004, 2008).Therefore, the direct consequence of nonergodicity is that knowledge about the structure of interindividual variation in the population cannot be applied at the level of individual subjects making up this population, and vice versa (Molenaar & Ram, 2010).

Nonergodicity not only appears in test theory, but also in factor analysis and the analysis of developmental processes (Molenaar, 2004, 2007, 2008; Molenaar & Ram, 2009, 2010).Unfortunately, most statistical methods in psychology/psychometrics are applied to a collection of individuals rather than to a single subject, as Kratochwill (1978,p. 3) discussed from a historical perspective (cf. Molenaar, 2004; Van Rijn, 2008; Zu, 2008).An obvious reason for the remarkable lack of interest in a pure single-subject perspective in education and

psychology is that “until recently, we have lacked statistical methods that are appropriate for analyzing intraindividual data” (Molenaar, Sinclair, Rovine, Ram, & Corneal, 2009, p. 260).Despite this type of analysis not in a niche in mainstream psychometrics in the past, single-subject longitudinal models is not an alternative but "the epistemological necessity of idiography" (Molenaar, 2004, p. 204) to obtain valid results concerning individual development, learning performance, and so forth. In a hypothetical example of a student learning his English vocabulary over three days, Schmitz (2006) pointed out that researchers who adopt the common pretest-posttest design would not be able to find out the dynamics of the learning process; whereas a sequence of 12 measurements within the three days would provide a totally different insight into the characteristics of the whole picture (Figure 1). In fact, the field of single-subject research in other branches of sciences (including but not limited to econometrics, meteorology, and communications engineering) has grown rapidly, and psychologists would do well to study this carefully (Holtzmann, 1963, p. 199).

Figure 1. The learning of vocabulary: measures of reproduction for one individual before and after learning (broken line) and for a series of measurements (continuous line).



Note. This figure is reproduced from Schmitz (2006, Fig. 1).

In recent years, both substantive and methodological research devoted to longitudinal intraindividual analysis started to appear in the psychological literature (e.g., Browne & Nesselroade, 2005; Chow, Ho, Hamaker, & Dolan, 2010; Chow, Nesselroade, Shifren, & McArdle, 2004; Chow, Zu, Shifren, & Zhang, 2011; Du Toit and Browne, 2007; Ferrer & McArdle, 2003; Ferrer & Nesselroade, 2003; Hamaker, Dolan, & Molenaar, 2002, 2005; Hamaker & Dolan, 2009; Ho, Shumway, & Ombao, 2006; Molenaar, 1985, 1987; Molenaar, De Gooijer, & Schmitz, 1992; Molenaar & Nesselroade, 1998, 2009; Molenaar et al., 2009; Nesselroade, McArdle, Aggen, & Meyers, 2002; Sbarra & Ferrer, 2006; Shifren, Hooker, Wood, Nesselroade, 1997; Song & Ferrer, 2009; Van Buuren, 1997; Van Rijn, 2008; Wood & Brown, 1994; Zhang & Browne, 2006; Zhang & Chow, 2010; Zhang, Hamaker, & Nesselroade, 2008; Zu, 2008). Statistical models commonly used in these studies include multiple/multivariate/vector autoregressive moving average model, P-technique model, dynamic factor model, and state space model (SSM). Although not introduced into psychology until recently, SSM has been considered as a very flexible modeling approach to analyze intraindividual processes (e.g., Molenaar et al., 2009; Molenaar & Ram, 2010). Time series model, P-technique model, and dynamic factor model can all be expressed in the general state space form. Furthermore, SSM also subsumes other advanced modeling procedures, such as structural equation modeling (SEM) and MLM (Chow et al., 2010; Ho et al., 2006; MacCallum & Ashby, 1986; Otter, 1986).In one of the leading texts, Durbin and Koopman (2001) wrote that "In our opinion, the only disadvantages are the relative lack in the statistical and econometric communities of information, knowledge and software regarding these models" (p. 52).In this research, the linear time-invariant SSM will be discussed.

State space model: Introduction, software, mathematical formulation, and the Kalman filter algorithm

State space models have their origin in system theory and engineering, beginning with the groundbreaking paper of Kalman (1960). As discussed by Commandeur, Koopman, and Ooms (2011), the applications were initially (and still are) used to solve problems in astronautics related to accurately tracking the position and velocity of moving objects such as aircrafts, missiles, and rockets, and were later adapted to treat time series data in econometrics (Harvey, 1989). More recently, state space methods received growing attention from behavioral and psychological scientists because of the flexibility to both evaluate the measurement properties and the concurrent and time-lagged relationships of latent variables in developmental processes by combining factor analysis and time series analysis (e.g., Chow et al., 2010; Hamaker et al., 2005; Ho et al., 2006; Van Rijn, 2008). In general, state space methods provide an effective approach for substantive areas that generate intensive longitudinal data (e.g., electroencephalography, economic and financial time series, and functional magnetic resonance imaging). However, applications of state space model in the social sciences (except econometrics) are still uncommon to most behavioral and psychological scientists.

One of the important aspects for the application of statistical models is software implementation. Ho et al. (2006, p. 159) commented that "Though state-space modeling has become widespread over the last decade in economics and statistics, there has not been much flexible software for the statistical analysis of general models in the state-space form.". In this research, the SAS/IML (version 9.3) program provided by Gu and Yung (2012) is adapted for all the computational work. It was shown that this program is easy to use and flexible to be modified for many specialized purposes (Gu & Yung, 2012).

The linear SSM encompasses two equations, namely the measurement equation and the transition equation:

$$y_t = b_t + H_t z_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \Theta_t)$$
$$z_{t+1} = a_t + F_t z_t + \eta_t, \quad \eta_t \sim N(0, \Psi_t)$$

where y_t is a $p\times1$ vector of observations at time t, b_t is a $p\times1$ vector of intercepts at time t, H_t is a $p\times q$ loading matrix at time t, z_t is a $q\times1$ vector of latent state variables at time t, ε_t is a $p\times1$ vector of measurement errors (also referred to as innovations)at time t, a_t is a $q\times1$ vector of constants at time t, F_t is a $q\times q$ transition matrix at time t capturing the underlying dynamic processes, η_t is a $q\times1$ vector of transition noise at time t. For Gaussian SSM, ε_t and η_t are assumed to follow multivariate normal distribution with zero mean and covariance matrices of Θ_t and Ψ_t, respectively. Usually, measurement errors are assumed uncorrelated with each other, and thus, Θ_t is a diagonal matrix. The subscript t indicates all parameters (a_t, b_t, F_t, H_t, Ψ_t, and Θ_t,) are time-varying. In many practical applications, it is usually assumed that parameters do not change over time so that the subscript can be suppressed. The following linear time-invariant SSM is discussed.

$$y_t = b + Hz_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, \Theta)$$
$$z_{t+1} = a + Fz_t + \eta_t, \quad \eta_t \sim N(0, \Psi)\,.$$

The Kalman filter (KF) algorithm is used to provide the normal theory (Gaussian) maximum likelihood (ML) estimates via any reasonable optimization technique. Beginning with the initial state variables ($z_{1|0}$) and the associated covariance matrix ($P_{1|0}$), the KF algorithm uses two steps: the prediction step and the filtering step. In the prediction step, the conditional expectation of state variables and the corresponding covariance matrix are estimated at the current observation using all prior observations; then, these estimates are updated using the actual current

observation in the filtering step. Specifically, the prediction step is initialized at the first observation ($t = 1$)

$$z_{1|0} = a + Fz_{0|0}$$
$$P_{1|0} = FP_{0|0}F' + \Psi$$

with the one-step-ahead prediction error and its associated covariance matrix computed as

$$\hat{\varepsilon}_1 = y_1 - y_{1|0} = y_1 - (b + Hz_{1|0})$$
$$D_1 = HP_{1|0}H' + \Theta \quad ,$$

and then, these estimates are updated in the filtering step

$$z_{1|1} = z_{1|0} + P_{1|0}H'D_1^{-1}\hat{\varepsilon}_1$$
$$P_{1|1} = P_{1|0} - P_{1|0}H'D_1^{-1}HP_{1|0} \quad .$$

Taking the values of $z_{1|1}$ and $P_{1|1}$, the prediction step and the filtering step are recursively implemented at the second observation, and so on. For $t = 1, 2, \ldots, T$, the Kalman recursion can be written as

$$z_{t|t-1} = a + Fz_{t-1|t-1}$$
$$P_{t|t-1} = FP_{t-1|t-1}F' + \Psi$$
$$\hat{\varepsilon}_t = y_t - y_{t|t-1} = y_t - (b + Hz_{t|t-1})$$
$$D_t = HP_{t|t-1}H' + \Theta$$
$$K_t = P_{t|t-1}H'D_t^{-1}$$
$$z_{t|t} = z_{t|t-1} + K_t\hat{\varepsilon}_t = z_{t|t-1} + P_{t|t-1}H'D_t^{-1}\hat{\varepsilon}_t$$
$$P_{t|t} = P_{t|t-1} + K_tD_tK_t' = P_{t|t-1} - P_{t|t-1}H'D_t^{-1}HP_{t|t-1}$$

where K_t is called the Kalman gain matrix. After the KF cycles through all observations, $\hat{\varepsilon}_t$ and D_t ($t = 1, 2, \ldots, T$) are readily available to be substituted into the log-likelihood function based on the assumption of multivariate normal distribution given by Schweppe (1965), and this function is referred to as the prediction error decomposition (PED):

$$\mathrm{PED} = \frac{1}{2}\sum_{t=1}^{T}[-p\log(2\pi) - \log|D_t| - \varepsilon_t' D_t^{-1}\varepsilon_t] \quad .$$

Finally, ML estimates can be obtained by maximizing PED with respect to the parameters (*a*, *b*, *F*, *H*, Ψ, and Θ).

Why is assessing the absolute goodness-of-fit for state space model important?

In regression analysis, the coefficient of determination (i.e., R^2) is almost always presented as a measure of goodness-of-fit and as evidence that the model is a good one. In SEM, many fit indices have been developed in recent decades, and there are three types of fit indices commonly reported: absolute (e.g., chi-square index, goodness of fit index [GFI], root mean square residual), parsimony (e.g., adjusted GFI, parsimonious GFI, root mean square error of approximation [RMSEA], Akaike information criterion [AIC], Schwarz Bayesian criterion [SBC, sometimes referred to as Bayesian information criterion, BIC], McDonald centrality), comparative/incremental (e.g., Bentler comparative fit index [CFI], Bentler-Bonett non-normed index [NNFI], Bollen normed index Rho1, Bollen non-normed index Delta2). Among the three types of fit indices, the absolute goodness-of-fit, particularly the chi-square index, is of critical importance because it provides the basisfor the plausibility of a model. Also, many other fit indices in the other two types are derived from the chi-square index. Though the chi-square index is criticized for its sensitivity to sample sizes, it is undeniably to be considered the flagship in reporting SEM because many commonly reported fit indices are derived from the chi-square index as we can see from the equations of these derived indices (e.g., RMSEA, CFI, NNFI).

With the increasing importance and popularity of state space model in psychological applications, assessing goodness-of-fit for SSM is becoming a crucial issue because the plausibility of a model serves the basis of any meaningful substantive interpretation. However,

this area is somewhat underdeveloped as illustrated by the small amount of space (less than a page) devoted to this topic in Durbin and Koopman (2001, Section 7.4, p. 152).

In the literature on econometrics and time series, measures of goodness-of-fit are usually associated with forecasting/predictive errors due to the intrinsic purpose of most time series models (i.e., forecast/prediction). Additionally, econometrists are mostly interested in selecting a model from several competing models. However, when all competing models are fundamentally poor in terms of the absolute goodness-of-fit, no matter which model is selected relative to others, we do not have a good approximation to the sample data. Compared to regression analysis and SEM, there is a noticeable lack of research devoted to the absolute goodness-of-fit for SSM. Therefore, this dissertation focuses on this important, but yet underdeveloped, issue—assessing the absolute goodness-of-fit for SSM.

Goal: Using bootstrap to assess the absolute goodness-of-fit for state space model

Ideally, a fit index is expected to follow a certain probability distribution so that deviance of the index from its expected value can be evaluated on the basis of that probability distribution. In other words, the observed value of the index can be compared against a certain distribution to determine its likelihood. In SEM, under the assumption of multivariate normality in the observed data, or equivalently, the assumption of a joint Wishart distribution among the elements of the observed covariance matrix, the ML fit function yields an overall fit index that asymptotically follows a chi-square distribution when the model is correct in the population (Bollen, 1989, Appendix 4A & 4B).This makes it feasible to test the plausibility of a particular structural equation model. In state space model, however, no overall fit index is derived from the fit

function (i.e., PED) based on the one-step-prediction error and its covariance matrix obtained at each observation from the KF algorithm. Thus, no statistical test can be conducted under the null hypothesis that the specified model is correct in the population.As an alternative, when there is noinformation about population distribution, the sampling distribution of a statistic could be empirically derived through bootstrap techniques (Efron, 1979; Efron & Tibshirani, 1993).

Bootstrap mimics the sampling process by assuming that the distribution observed in the sample resembles the population distribution.The standard nonparametric bootstrap procedure includes the following four steps:

1. Fit a model to the sample.
2. Draw, with replacement, a random sample of the same size from the original sample.
3. Fit the same model to the bootstrap sample.
4. Repeat Steps 2 and 3 a large number of times, *B*, known as bootstrap replications, and obtain the sampling distribution of the statistic of interest.

After deriving the sampling distribution using the bootstrap method, the observed value of any statistic from the available sample can be compared to determine its likelihood. In the context of model evaluation, the validity of the model can be assessed without any distributional information in the population.

The standard nonparametric bootstrap procedure described above is not appropriate for time series data because it ignores the inherent lead-lag relationship in time. Two bootstrap procedures that are appropriate for SSM are used in this study. The first bootstrap method is parametric bootstrap (also known as Monte Carlo resampling method).The second bootstrap method is the residual-based bootstrap, first proposed by Stoffer and Wall (1991) to assess the precision of Gaussian ML estimates of the parameters of linear SSMs.

CHAPTER 2: REVIEW OF LITERATURE

At this time there is no literature related to assessing the absolute goodness-of-fit for state space model. This chapter reviews selected literature that mainly addresses methodological dissemination of state space model geared toward social and behavioral scientists. Also included in this chapter is a review of existing literature on assessing goodness-of-fit for state space model and on using the bootstrap procedure to assess goodness-of-fit in other contexts.

Recent literature of state space model related to education and psychology

Chow, Ho, Hamaker, and Dolan (2010) provided a comprehensive discussion of the similarities and differences of SSM and SEM through analytic comparisons and numerical simulations. They illustrated relative merits of SSM and SEM in addressing questions pertaining to intraindividual change and interindividual differences. Beyond these authors' contribution, MacCallum and Ashby (1986) and Otter (1986) also contributed early work on the equivalence between the two modeling approaches.

Hamaker, Dolan, and Molenaar (2005) presented the specific condition that must be satisfied to generalize results obtained from the interindividual level to the intraindividual level, and illustrated the analyses of intraindividual structure by fitting the P-technique models and multiple indicator vector autoregressive models in the state space framework to the empirical data collected from the Five Factor Model of personality.

In an introductory chapter on SSM, Ho, Shumway, and Ombao (2006) presented the flexible treatment of state space modeling for intensive longitudinal data.They illustrated two separate

applications of SSM. The first one is neural connectivity using fMRI data, and the second is traffic networking.

Molenaar and Nowell (2003) presented a nonlinear SSM using the extended KF and smoothing algorithm to fit the Schöner-Haken-Kelso model of human movement phase transitions to finger motions data. The extended KF and smoothing algorithm was later applied to the state space analyses of human developmental processes at the individual level (Molenaar, 2008; Molenaar, Sinclair, Rovine, Ram, & Corneal, 2009), a simulated 4-variate time series data (Molenaar & Ram, 2009), and a Monte Carlo simulation (Molenaar & Ram, 2010). In general, the results from these studies are promising. However, as Molenaar and Ram (2010) pointed out, the EKFIS computer program implementing the nonlinear SSM using the extended KF and smoothing algorithm"is not at all user-friendly, requiring writing and compiling separate Fortran subroutines", and they hoped that "the development of the EKFIS program along these lines will further improve its fidelity" (p. 30).

Song and Ferrer (2009) examined the finite sample properties of the KF and smoothing algorithm in a Monte Carlo simulation. Results indicated that parameter estimates are mostly asymptotically normal, accurate, precise and robust, especially for moderate and long time series. In addition, empirical example was provided by applying the state space methods on the daily affect data collected from a dating couple.

Van Rijn (2008) extended the state space methods for categorical time series data, and investigated the performance of the KF and smoothing algorithm. Specifically, it is demonstrated that the state space methodology can handle the analysis of both standard and dynamic item response theory (IRT) models in a straightforward manner.

Zhang, Hamaker, and Nesselroade (2008) compared the state space modeling technique using the KF algorithm to other three modeling approaches for estimating a dynamic factor model. The simulation results showed that all four methods yielded acceptable parameter estimates in almost all conditions. In their work, Zhang, Hamaker, and Nesselroade also discussed software programs implementing the four modeling approaches.

Zu (2008) compared the KF algorithm and the extended KF algorithm to track dynamics of both latent factors and time-varying coefficients in a dynamic factor model. The results demonstrated that the KF algorithm is robust to the type of model misspecification for estimating factor scores considered in the Monte Carlo simulation, and that certain parameter estimates are biased while others are not.

Literature on assessing goodness-of-fit for state space model

Literature on assessing goodness-of-fit for state space model is very limited, and often resides in the context of time series analysis (see Durbin & Koopman, 2001; Harvey, 1989). As discussed before in Section 1.3,because the purpose of most time series models is to predict, goodness-of-fit for time series model is usually associated with predictive errors, measured by the prediction error variance. Besides, goodness-of-fit criterion such as AIC and BIC is used to select a *comparatively* good model (see Harvey, 1989, section 2.6.3 and 5.5, for more details). One major limitation of using these strategies in model diagnosis and selectionis that the absolute goodness-of-fit is not ensured.In other words, a bad model will inevitably be selected if none of the competing models can provide acceptable fit to the data. In addition, published literature is almost always limited to univariate time series models. In other texts or related chapters for state space methods (e.g., Brockwell & Davis, 2002, Chapter 12; Chatfield, 2004, Chapter 10;

Fahrmeir & Tutz, 2001, Chapter 8; Harvey, Koopman, & Shephard, 2004; Lütkepohl, 2005, Chapter 18; Shumway & Stoffer, 2011, Chapter 6), the topic of goodness-of-fit is not mentioned at all. This lack of literature does present an excellent opportunity for researchers to explore and contribute to this body of knowledge.

Using bootstrap to assess goodness-of-fit

Some authors have used bootstrap, particularly parametric bootstrap, to assess goodness-of-fit in different contexts. Bone, Sharma, and Shimp (1989) illustrated how to implement the procedure by re-analyzing two previous studies in marketing and consumer research, and obtained the sampling distributions for some fit indices in SEM.Von Davier (1997) bootstrapped four goodness-of-fit statistics for sparse categorical data in a Monte Carlo study. He concluded that parametric bootstrap is a useful alternative approach with some examined statistics even if the data are very sparse. Parametric bootstrap has also been used to evaluate goodness-of-fit for IRT (Stone, Ankenmann, Lane, & Liu, 1993; Stone, 2000). However, no published work is available using bootstrap to assess goodness-of-fit for SSM.

CHAPTER 3: METHODS

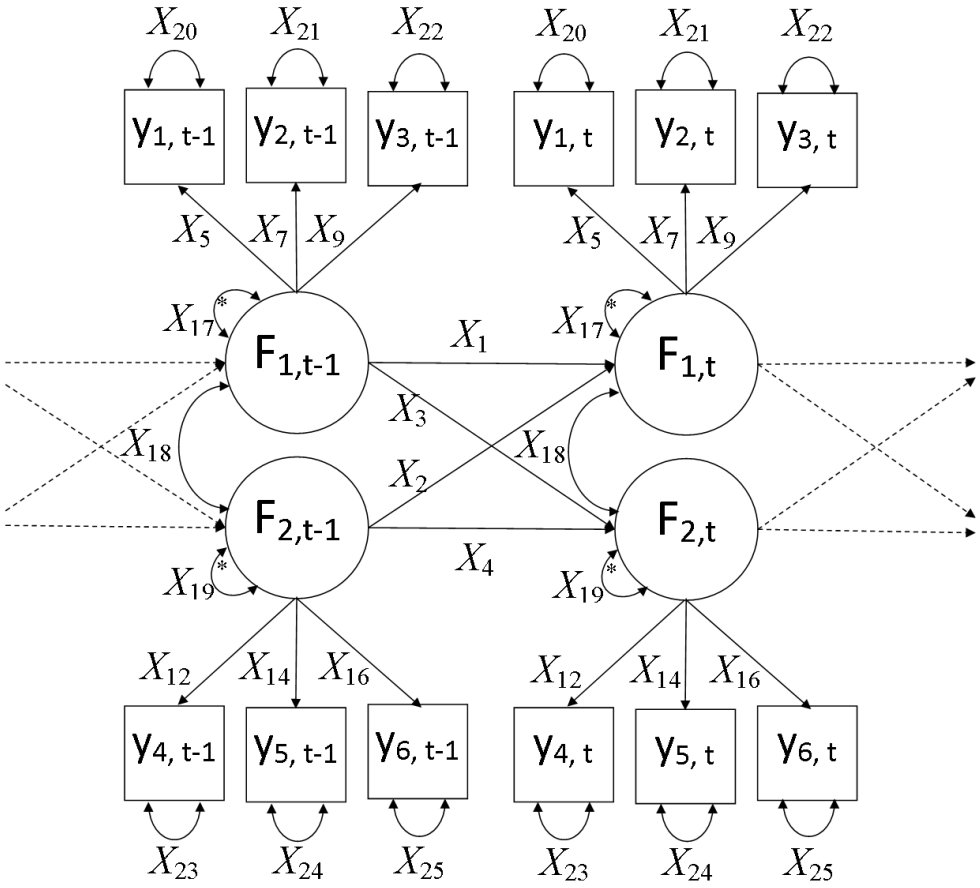
The model, the condition, and the data generation procedure

The model used in Zhang et al. (2008) is used in this research and is described as the following:

There are six observed variables and two factors in this model. The first three observed variables load on the first factor and the other three observed variables load on the second

factor. The factors have a one-lag autoregressive and cross-regressive structure, which means the first factor at the current time has a direct influence on the first and second factors at the next time, and so does the second factor (p. 379).

Figure 2.The simulated model with six observed variables and two factors. The factor scores have one-lag autoregressive structure.



Note. * indicated X_{17} and X_{19} are fixed to 0.36 for model identification.

Figure 2 portrays this dynamic factor model, in which the intercepts, *b*, and the constants, *a*, are both fixed to zero. The other parameter matrices are freely estimated, and the true values of the elements in these matrices are provided below (in matrix notation):

$$F = \begin{bmatrix} .8 & 0 \\ 0 & .8 \end{bmatrix}, \quad \Psi = \begin{bmatrix} .36^{*} & .18 \\ .18 & .36^{*} \end{bmatrix}, \quad H = \begin{bmatrix} 1 & 0^{*} \\ 1 & 0^{*} \\ 1 & 0^{*} \\ 0^{*} & 1 \\ 0^{*} & 1 \\ 0^{*} & 1 \end{bmatrix}, \quad \Theta = \begin{bmatrix} .1 & & & & & \\ & .1 & & & & \\ & & .1 & & & \\ & & & .1 & & \\ & & & & .1 & \\ & & & & & .1 \end{bmatrix}.$$

The asterisk beside the value indicates that the element in the matrix is fixed. For example, both diagonal elements in Ψ (representing variances of state variables) are fixed to 0.36 to identify the model.The condition controlled in this model is the length of the time series data (T = 50, 100, 500). In order to generalize the results, 100 random samples (N = 100) are generated in each condition from the population.

Given the parameter values and the time series length, the following steps are followed to generate each random sample:

1. Generate z_0 from the bivariate normal distribution *MultiNorm* ((0 0)', *diag*(100 100)).
2. Set the iteration number t = 1.
3. Generate η_t from the bivariate normal distribution *MultiNorm* ((0 0)', Ψ).
4. Calculate z_t using $z_t = a + Fz_{t-1} + \eta_t$.
5. Generate ε_t from the multivariate normal distribution *MultiNorm* ((0 0 0 0 0 0)', Θ).
6. Calculate y_t using $y_t = b + Hz_t + \varepsilon_t$.
7. Set $t = t + 1$ and return to Step 3.
8. Repeat Steps 3 to 6 until t > T + 1000.
9. Save the data from 1001 to T + 1000.

Bootstrap the fit function in state space model: parametric and residual-based bootstrap

For each random sample, two bootstrap procedures are applied to derive the sampling distributions of the PED function of a specified state space model. The first procedure, the parametric bootstrap, is essentially a Monte Carlo simulation in which the population parameters are estimates from the original sample and repeated bootstrap samples are simulated based on the data-generating mechanism implied by the specified model. Specifically, in a parametric bootstrap, the steps generating the random samples are followed to obtain a bootstrap sample except that the parameter estimates from fitting the model to the original sample are used in those steps. The underlying assumptions of parametric bootstrap are that the specified model is correct in the population and that the time series data conform to a certain distribution.

The second procedure, the residual-based bootstrap, is considered as a semi-parametric approach because population parameters are taken to be sample estimates, assuming the specified model is correct. This is the same as in a parametric bootstrap. On the other hand, random samples are drawn, with replacement, from the standardized residuals as in the standard nonparametric bootstrap. The residual-based bootstrap procedure is based on the SSM expressed in the innovations form:

$$\begin{aligned} \varepsilon_t &= y_t - Hz_{t|t-1} \\ D_t &= HP_{t|t-1}H^{'} + \Theta \\ K_t &= P_{t|t-1}H^{'}D_t^{-1} \\ z_{t+1|t} &= Fz_{t|t-1} + K_t\varepsilon_t \\ y_t &= Hz_{t|t-1} + \varepsilon_t \end{aligned} .$$

Let $\hat{\theta}$ denote the estimate of θ, the residual-based bootstrap is implemented as follows:

1. Standardize $\hat{\varepsilon}_t$ using $\hat{D}_t^{-1/2}\hat{\varepsilon}_t$, denoted as $\tilde{\varepsilon}_t$.

2. Draw, with replacement, a random sample from $\tilde{\varepsilon}_t$ to obtain $\tilde{\varepsilon}_t^*$.

3. Construct a bootstrap sample by using the following two equations:

$$z_{t+1|t} = \hat{F} z_{t|t-1} + \hat{F}\hat{K}_t \hat{D}_t^{-1/2} \tilde{\varepsilon}_t^*$$
$$y_t = \hat{H} z_{t|t-1} + \hat{D}_t^{-1/2} \tilde{\varepsilon}_t^* \quad .$$

The basic idea behind the residual-based bootstrap is that the standardized residuals are independent and identically distributed, and thereby exchangeable, after all the dynamic and measurement relationships have been accounted for by the model. This procedure, however, is not robust against model misspecification (c.f. Stoffer & Wall, 1991, 2004; Zhang & Chow, 2010).

For each original random sample, both bootstrap procedures are repeated 2000 times to obtain the sampling distributions of the PED function. Such a large number of bootstrap replications is chosen because the estimated percentiles will be used for hypothesis testing and confidence interval construction, as recommended by Yung and Chan (1999, p. 100). All computations are done by SAS 9.3 on Unix.

A power analysis

In order to examine the power of the bootstrap procedures in assessing the absolute goodness-of-fit, a power analysis will be conducted. Specifically, the transition matrix used in the model will be fixed, which reduces a SSM to a P-technique model (Cattell, Cattell, & Rhymer, 1947), and the P-technique models will be fitted to each random sample. Because the random samples are simulated from the SSM, fitting the P-technique model, which is the constrained SSM, will bring specification errors. Then, a decision can be made for each random sample based on the absolute goodness-of-fit by comparing the PED value from fitting the P-

technique model to the two sampling distributions derived from both bootstrap procedures. That is, if the PED value is inside the confidence interval constructed by the estimated percentiles, the plausibility of a model is supported. Otherwise, the model is considered a poor approximation to the sample data, and thus, it should be rejected. The power in each condition of both bootstrap procedures can be computed by dividing by 100 the number of models rejected.

Molenaar and Nesselroade (2009) presented some simulation results to demonstrate the recoverability of P-technique model. Specifically, the loading parameters and factor scores were recovered very satisfactorily from the P-technique models even though the transition matrix was incorrectly fixed in the P-technique models.

CHAPTER 4: RESULTS

Convergence

The numbers of convergent cases from both bootstrap approaches are provided in Table 1. It shows that the means of the convergent cases are higher when the length of time series gets longer. In addition, the means of the convergent cases from parametric bootstrap are consistently higher than those from the residual-based bootstrap, which indicates higher stability of the parametric bootstrap approach.

Table 1. Convergence cases from both bootstrap approaches in each condition.

	Parametric bootstrap			Residual-based bootstrap		
Length	50	100	500	50	100	500
Mean	1807.12	1883.09	1994.56	1754.23	1850.91	1987.53
SD	82.19	54.85	5.18	83.92	47.55	6.75

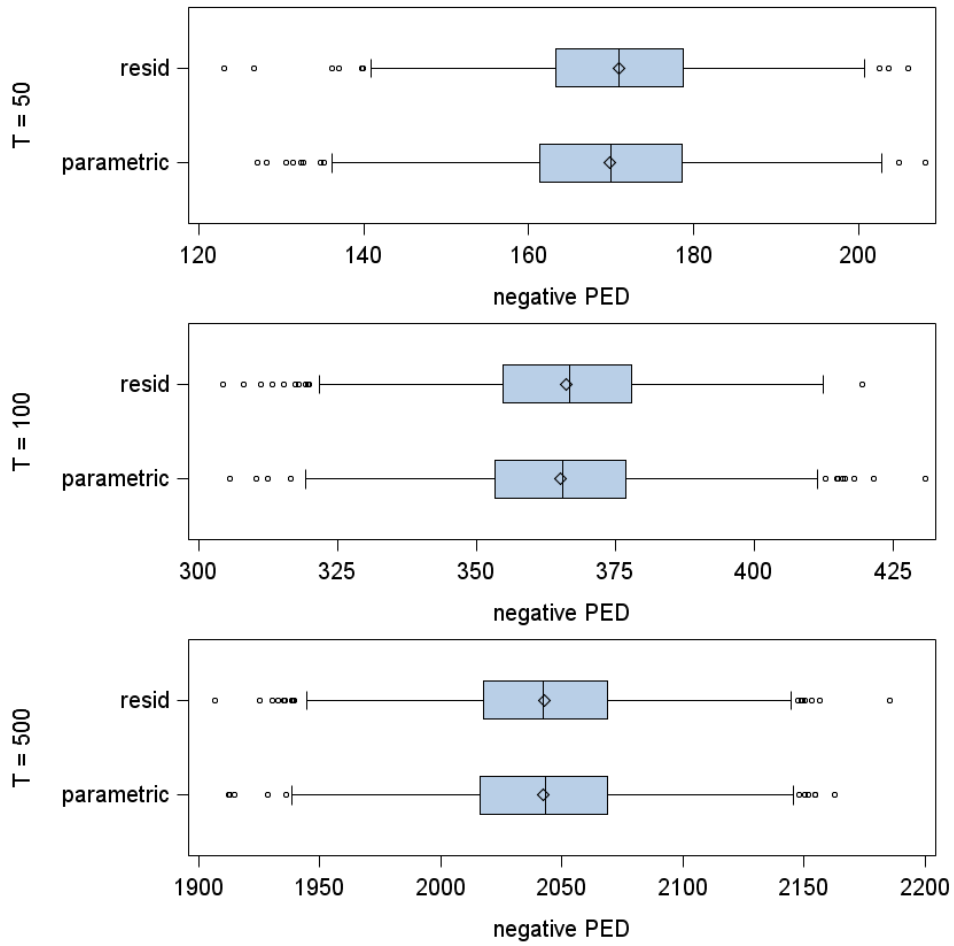
Convergence rates at length of 50 from both bootstrap procedures reported here are much higher than those reported by Zhang, Chow, and Ong (2011), where they compared a sandwich-type standard error estimator of a dynamic factor model to the moving block bootstrap approach. Specifically, they wrote, "proportions of nonconvergence… were higher than 20% in 59 of the first 100 simulation samples at $T = 50$" (p. 92). The discrepancy in the convergence rates from this study and Zhang et al. (2011) is largely due to the different modeling approaches between the state space model approach and the block Toeplitz approach in SEM. Discussions about the computational issue are provided in the last chapter.

Confidence intervals

By bootstrapping the simulated time series and fitting the true state space model to each bootstrap sample, the sampling distribution of the PED function, and thus the estimated percentiles, are derived. Because the value of the PED function is negative, the negative PED function is minimized in the program to find the ML estimates, which is equivalent to maximizing the positive PED function.

The sampling distributions derived from both bootstrap procedures are similar. In general, as the length of time series increases, the two sampling distributions get more and more similar. Specifically, the 5^{th} percentiles are a bit different from the two sampling distributions when the length is 50, and as are the 95^{th} percentiles. But this discrepancy almost disappears when the length reaches 100 and 500. Box plots from 3 simulated samples (one for each condition) are provided in Figure 3.

Figure 3. Box plots of the bootstrap estimates from the two bootstrap approaches.



Assessing model plausibility in the power analysis

Nonconverged results are excluded from the power analysis. For each simulated sample, the sample PED function values from fitting the true state space model and the misspecified P-technique model are compared to the 90% confidence interval (CI) constructed by the 5^{th} and 95^{th} percentiles from both bootstrap procedures. According to the judging rule, plausibility of a model is supported, though it is not a sufficient condition to ensure this conclusion, if the PED function value is inside the 90% CI; otherwise, the model plausibility is questionable.

Table 2 provides the results from the power analysis. Because the CIs from both bootstrap procedures are very similar, the outcomes are the same. Consistent with the intuitive expectation, all PED function values from fitting the true SSMs are within the 90% CI; whereas all PED

function values from fitting the misspecified P-technique models are also within the 90% CI[1], which indicates that there is no significant reduction in goodness-of-fit.

Table 2. Results of the power analysis.

	State space models			P-technique models		
Length	50	100	500	50	100	500
Not reject	91	94	100	87	95	99
Reject	0	0	0	0	0	0
Total	91	94	100	87	95	99

Computation time

Generally, bootstrap is very computationally intensive, and therefore time-consuming, because of the repetitive model fitting to each bootstrap sample. Using the program written in SAS/IML, fitting a state space model takes approximately 10 seconds, which in total will require about 33,333 hours (3 conditions * 100 samples * 2000 bootstrap replications * 2 bootstrap procedures * 10 seconds / 360 seconds per hour = 33,333.33 hours), equivalently 1,389 days, to obtain all the bootstrap percentiles if only a single computer is used. Fortunately, such tremendous computation tasks were finished in three days using the High Performance Computing (HPC) facility provided by the Center for Research Methods and Data Analysis (CRMDA).

CHAPTER 5: DISCUSSION

[1] A 90% CI was used instead of the more traditional 95% CI because estimation of the 2.5^{th} and 97.5^{th} percentiles is less stable than estimation of the 5^{th} and 95^{th}. An alternative approach could have been increasing the number of draws above the 2000 used, but this would have been time consuming (see section 4.4) and there is no reason to expect this would impact any of the conclusions of this study.

Assessing the absolute goodness-of-fit for state space model is necessary to properly interpret substantive results in real-data studies. To this end, this dissertation used two bootstrap procedures to derive the sampling distribution of the PED function from a state space model that is used to generate repeated random samples under different conditions of time series length. The PED function values from fitting the true state space model and the misspecified P-technique model to each random sample are compared to the CIs to assess the absolute goodness-of-fit. On a very positive note, none of the true state space models is rejected. Unfortunately none of the misspecified P-technique models were rejected either. The results indicate that both bootstrap procedures have no power to detect the specification error of constraining the transition matrix in the state space model considered in this study. Though the results are counterintuitive, they are consistent with the findings from Molenaar and Nesselroade (2009). That is, constraining the transition matrix in the state space model would not significantly affect factor loading and the factor score estimates. Similar findings were also reported by Chow et al. (2011) and Zu (2008), where these authors referred to the orthogonality of model parameters to explain this phenomenon. The concept of orthogonality of parameters is well known in regression analysis, and it is extended in the context of SEM by Yuan, Marshall, and Bentler (2003). Adopting the orthogonality testing methods proposed in Yuan et al. (2003) for parameters in state space model is a topic for future research.

Despite the results from the power analysis, models that are more complex than the one considered here may give opposite results. For instance, it is worth noting that the time series model considered in this study is stationary, and it is generally recognized that nonstationary time series can be treated by SSM. In time series analysis, stationarity is an important concept. The state vector in this example is stationary because all roots of the determinant equation

$$\left|I - \lambda F\right| = 0$$

lie outside the complex unit circle. A brief definition of weak stationary multivariate time series can be found, for example, in Molenaar et al. (p. 262, 2009) in the context of developmental psychology. A thorough discussion of stationarity is beyond the scope of this paper, but readers can refer to Lütkepohl (2005) for details. Further research on nonstationary time series data is a topic for future research.

Lastly, a practical purpose of fitting a P-technique model is its simple implementation in software packages. Conventional SEM packages (e.g., LISREL, SAS PROC CALIS) can be used directly for fitting a P-technique model, while fitting a SSM requires extended programming skills from the user to write his/her own program, as is realized in this dissertation. An alternative method to include the lagged structure in the SEM approach is to use the block Toeplitz matrix made up of the concurrent and lagged autocovariance/autocorrelation matrix. However, severe computational difficulty, causing possibly very high nonconvergence rate, can be a separate problem to overcome. For example, for a *p*-variate sample, a one-lag model will require the researcher to create a $2p \times 2p$ block Toeplitz matrix, and a two-lag model will need a $3p \times 3p$ block Toeplitz matrix, and so on. Apparently, for large *p*, a high-dimensional matrix needs to be inverted in the algorithm for SEM, which can be very difficult to handle even with today's computing power. Compared to the block Toeplitz approach, the KF algorithm for the SSM has better efficiency when higher orders of lag are introduced. Therefore, the SSM is generally preferred than the block Toeplitz approach in SEM.

REFERENCES

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Bone, P. F., Sharma, S., & Shimp, T. (1989). A bootstrap procedure for evaluating goodness-of-fit indices of structural equation and confirmatory factor models. *Journal of Marketing Research, 26*, 105-111.

Brockwell, P. J., & Davis, R. A. (2006). *Time series: Theory and methods*. New York, NY: Springer.

Browne, M. W., & Nesselroade, J. (2005). Representing psychological processes with dynamic factor models: Some promising uses and extensions of autoregressive moving average time series models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 415-451). Mahwah, NJ: Lawrence Erlbaum Associates.

Chatfield, C. (2004). *The analysis of time series: An introduction*. Boca Raton, FL: Chapman & Hall/CRC.

Chow, S.-M., Ho, M.-H.R., Hamaker, E. L., & Dolan, C. V. (2010).Equivalence and differences between structural equation modeling and state-space modeling techniques. *Structural Equation Modeling, 17,* 303-332.

Chow, S.-M., Nesselroade, J. R., Shifren, K., & McArdle, J. J. (2004). Dynamic structure of emotions among individuals with Parkinson's disease. *Structural Equation Modeling, 11*, 560-582.

Chow, S.-M., Zu, J., Shifren, K., & Zhang, G. (2011). Dynamic factor analysis models with time-varying parameters. *Multivariate Behavioral Research,46,* 303-339.

Commandeur, J. J. F., Koopman, S. J., & Ooms, M. (2011). Statistical software for state space methods. *Journal of Statistical Software, 41(1),* 1-18. Retrieved from http://www.jstatsoft.org/v41/i1/

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B,* 1-39.

Dolan, C. V. (2005). *MKFM6: Multi-group, multi-subject stationary time series modeling based on the Kalman filter*. Retrieved December 1, 2004 from http://users.fmg.uva.nl/cdoman/

Du Toit, S. H. C., & Browne, M. W. (2007). Structural equation modeling of multivariate time series. *Multivariate Behavioral Research, 42*,67-101.

Durbin, J., & Koopman, S. J. (2001). *Time series analysis by state space methods*. New York, NY: Oxford University Press.

Efron, B. (1979). Bootstrap methods: Another look at the jack-knife. *Annals of Statistics, 7*, 1-26.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall/CRC.

Fahrmeir, L. & Tutz, G. (2001).*Multivariate statistical modelling based on generalized linear models*. New York, NY: Springer.

Ferrer, E., & McArdle, J. J. (2003). Alternative structural models for multivariate longitudinal data analysis. *Structural Equation Modeling, 10,* 493-524.

Ferrer, E., & Nesselroade, J. R. (2003). Modeling affective processes in dyadic relations via dynamic factor analysis. *Emotion, 4*, 344-360.

Gu, F., & Yung, Y-F. (2012). A SAS/IML program using the Kalman filter for estimating state space models. *Behavior Research Methods*.

Hamaker, E. L.,& Dolan, C. V. (2009). Idiographic data analysis: Quantitative methods--from simple to advanced. In J. Valsiner et al. (eds.), *Dynamic process methodology in the social and developmental sciences* (pp. 191-216). © Springer Science+Business Media LLC.

Hamaker, E. L., Dolan, C. V., & Molenaar, P. C. M. (2002). On the nature of SEM estimates of ARMA parameters. *Structural Equation Modeling, 9,* 347-368.

Hamaker, E. L., Dolan, C. V., & Molenaar, P. C. M. (2005). Statistical modeling of the individual: Rational and application of multivariate stationary time series analysis. *Multivariate Behavioral Research, 40,* 207-233.

Hannan, E. J. & Deistler, M. (1988). *The Statistical Theory of Linear Systems*. New York, NY: John Wiley.

Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman filter*. Princeton, NJ: Princeton University Press.

Harvey, A. C., Koopman, S. J., &Shephard, N. (2004). *State space and unobserved component models: Theory and applications*. New York, NY: Cambridge University Press.

Ho, M.-H. R., Shumway, R., & Ombao, H. (2006). State-space approach to modeling dynamic processes: Applications in biological and social sciences. In T. A. Walls & J. L. Schafer (Eds.), *Models for intensive longitudinal data* (pp. 148-170). New York, NY: Oxford University Press.

Holtzmann, W. H. (1963). Statistical models for the study of change in the single case. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 199-211). Milwaukee, WI: University of Wisconsin Press.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering, 82(Series D),* 35-45.

Kratochwill, T. R. (1978) *Single subject research: Strategies for evaluating change*. New York, NY: Academic Press.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Berlin: Springer-Verlag.

MacCaullum, R. C., & Ashby, G. F. (1986). Relationships between linear systems theory and covariance structure modeling. *Journal of Mathematical Psychology, 30*, 1-27.

Molenaar, P. C. M. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika, 50*, 181-202.

Molenaar, P. C. M. (1987). Dynamic factor analysis in the frequency domain: Causal modeling of multivariate psychological time series. *Multivariate Behavioral Research, 22*,329-353.

Molenaar, P. C. M. (2003).*State space techniques in structural equation modeling: Transformation of latent variables in and out of latent variable models*. Retrieved from http://www.hhdev.psu.edu/hdfs/faculty/molenaar.html

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bring the person back into scientific psychology, this time forever. *Measurement, 2*, 201-218.

Molenaar, P. C. M. (2007). On the implications of the classical ergodic theorems: Analysis of developmental processes has to focus on intra-individual variation. *Developmental Psychobiology, 50*, 60-69.

Molenaar, P. C. M. (2008). Consequences of the ergodic theorems for classical test theory, factor analysis, and the analysis of developmental processes. In S. M. Hofer & D. F. Alwin (Eds.), *Handbook of cognitive aging: Interdisciplinary perspectives*. (pp. 90-104). Thousand Oaks, CA: Sage.

Molenaar, P. C. M., De Gooijer,J. G., & Schmitz, B. (1992). Dynamic factor analysis of nonstationary multivariate time series. *Psychometrika, 57*, 333-349.

Molenaar, P. C. M., & Nesselroade, J. R. (1998). A comparison of pseudo-maximum likelihood and asymptotically distribution-free dynamic factor analysis parameter estimation in fitting covariance-structure models to block-Toeplitz matrices representing single-subject multivariate time-series. *Multivariate Behavioral Research, 33*,313-342.

Molenaar, P. C. M., & Nesselroade, J. R. (2009).The recoverability of P-technique factor analysis. *Multivariate Behavioral Research, 44*, 130-141.

Molenaar, P. C. M., & Nowell, K. M. (2003). Direct fit of a theoretical model of phase transition in oscillatory finger motions. *British Journal of Mathematical and Statistical Psychology, 56*, 199-214.

Molenaar, P. C. M., Ram, N. (2009). Advances in dynamic factor analysis of psychological processes. In J. Valsiner et al. (eds.), *Dynamic process methodology in the social and developmental sciences* (pp. 255-268). © Springer Science+Business Media LLC.

Molenaar, P. C. M., Ram, N. (2010).Dynamic modeling and optimal control of intraindividual variation: A computational paradigm for nonergodic psychological processes. In S-M. Chow, E. Ferrer, & F. Hsieh (Eds.), *Statistical Methods for modeling human dynamics: An interdisciplinary dialogue*. (pp. 13-37). New York, NY: Routledge.

Molenaar, P. C. M., Sinclair, K. O., Rovine, M.J., Ram, N., & Corneal, S. E. (2009). Analyzing developmental processes on an individual level using nonstationary time series modeling. *Developmental Psychology*, 45, 260-271.

Nesselroade, J. R., McArdle, J. J., Aggen, S. H.,& Meyers, J. M.(2002). Dynamic factor analysis models for representing process in multivariate time-series. In D. S. Moskowitz & S. L.

Hershberger (Eds.), *Modeling intraindividual variability with repeated measures data: Methods and applications multivariate applications book series*. (pp. 235-265). Mahwah, NJ: Lawrence Erlbaum Associates.

Otter, P. (1986). Dynamic structure systems under indirect observation: Identifiability and estimation aspects from a system theoretic perspective. *Psychometrika, 51*, 415-428.

Sbarra, D. A., & Ferrer, E. (2006). The structure and process of emotional experience following nonmarital relationship dissolution: Dynamic factor analyses of love, anger, and sadness. *Emotion, 6*, 224-238.

Schmitz, B. (2006). Advantages of studying processes in educational research. *Learning and Instruction, 16*, 433-449.

Schweppe, F. (1965). Evaluation of likelihood functions for Gaussian signals. *IEEE Transactions on Information Theory, 11*, 61-70.

Selukar, R. (2011). State space modeling using SAS. *Journal of Statistical Software, 41*(12), 1-13. Retrieved from http://www.jstatsoft.org/v41/i12/

Shifen, K., Hooker, K., Wood, P, & Nesselroade, J. R. (1997). Structure and variation of mood in individuals with Parkinson's disease: A dynamic factor analysis. *Psychology and Aging, 12*, 328-339.

Shumway, R. H., & Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis, 3*, 253-264.

Shumway, R. H., & Stoffer, D. S. (2011). *Time series analysis and Its application*. New York, NY: Springer.

Song, H., & Ferrer, E. (2009). State-space modeling of dynamic psychological processes via the Kalman smoother algorithm: Rationale, finite sample properties, and applications. *Structural Equation Modeling, 16,* 338-363.

Stoffer, D. S., & Wall, K. D. (1991). Bootstrapping state-space models: Gaussian maximum likelihood estimation and the Kalman filter. *Journal of the American Statistical Association, 86,* 1024-1033.

Stoffer, D. S., & Wall, K. D. (2004). Resampling in state space models. In A. Harvey, S. J. Koopman, & N. Shephard (Eds.), *State space and unobserved component models: Theory and applications*. (pp. 171-202). New York, NY: Cambridge University Press.

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement, 37,* 58-75.

Stone, C. A., Ankenmann, R. D., Lane, S., & Liu, M. (1993). Scaling QUASAR's performance assessments. Paper presented at the 1993 annual meeting of the American Educational Research Association, Atlanta, GA.

Van Buuren, S. (1997). Fitting ARMA time series by structural equation models. *Psychometrika, 62,* 215-236.

Van Rijn, P. (2008). *Categorical time series in psychological measurement* (Unpublished doctoral dissertation). University of Amsterdam, Amsterdam, NL.

Von Davier, M. (1997). Boostrapping goodness-of-fit statistics for sparse categorical data: Results of a Monte Carlo study. *Methods of Psychological Research Online,* 2, 29-48.

Wei, W. W. S. (1990). *Time series analysis: Univariate and multivariate methods*. Redwood City, CA: Addison-Wesley.

Wood, P., & Brown, D. (1994). The study of intraindividual differences by means of dynamic factor models: Rationale, implementation, and interpretation. *Psychological Bulletin, 116,* 166-186.

Yuan, K.-H., Marshall, L. L., & Bentler, P. (2003). Assessing the effect of model misspecifications on parameter estimates in structural equation models. *Sociological Methodology, 33,* 241-265.

Yung, Y-F., & Chan, W. (1999). Statistical analyses using bootstrapping: Concepts and implementation. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research*. (pp. 81-105). Thousand Oaks, CA: Sage.

Zhang, G., & Browne, M. W. (2006). Bootstrap standard error estimates in dynamic factor analysis. *Multivariate Behavioral Research, 45,*453-482.

Zhang, G., & Chow, S-M. (2010). Standard error estimation in stationary multivariate time series models using residual-based bootstrap procedures. In P. C. M. Molenaar & K. M. Newell (Eds.), *Individual pathway of change: statistical models for analyzing learning and development* (pp. 169-182). Washington DC: American Psychological Association.

Zhang, G., Chow, S.-M., & Ong, A. D. (2011). A sandwich-type standard error estimator of sem models with multivariate time series. *Psychometrika, 76*, 77-96.

Zhang, Z., Hamaker, E. L., & Nesselroade, J. R. (2008). Comparisons of four methods for estimating a dynamic factor model. *Structural Equation Modeling, 15,* 377-402.

Zu. J. (2008). *Examing nonlinear changes of coefficients in time-varying dynamic factor models* (Unpublished master's thesis). University of Notre Dame, South Bend, IN.