State Space Techniques in Structural Equation Modeling

Transformation of latent variables in and out of latent variable models

Peter C.M. Molenaar

June 2003

Second version

Preface

Structural equation modeling is an exciting and very productive approach. In the hands of Jöreskog, Sörbom, Bentler, Browne, Arminger, Satorra, Muthen, and many others, structural equation modeling has evolved into a mature and encompassing technique for scientific data analysis. I have had the privilege of teaching introductory and advanced courses in structural equation modeling for more than twenty years in many places of the world. This has always been a rewarding endeavor: structural equation modeling is eminently suited to accommodate theoretical insights about the data generating process. Moreover, it significantly furthers understanding of the commonalties between seemingly disparate techniques like (M)AN(C)OVA, regression analysis, factor analysis, and hybrid variants thereof. Although my own background is in signal analysis, I always have been involved in structural equation modeling too. Therefore I gladly accepted a kind offer by Jan de Leeuw, editor of this series, to write a monograph on structural equation modeling and state-space modeling. This would bring together two of my favorite fields of research.

In the realm of signal analysis and time series analysis, state-space modeling occupies a similar position as structural equation modeling does in multivariate statistical analysis. The state-space model can be regarded as a kind of canonical model ranging over almost all linear models used in signal analysis and time series analysis. Moreover, the formal structure or layout of the state-space model bears a close relationship to the general structural equation model. Yet, this close formal relationship between state-space modeling and structural equation modeling has hardly been noticed in the published literature, let alone been exploited to the benefit of one (or both) of them. I do not know the reasons for this mutual neglect, but perhaps differences in scientific context, jargon and technical detail may have played a role. State-space modeling is a typical engineering approach, usually applied to analyze the sequentially dependent behavior of a single system across many time points. In contrast, structural equation modeling is a social scientific approach, usually applied to the behavior of many systems observed at one or a few time points. Issues such as feedback and optimal control play an important role in state-space modeling, while they appear to be absent in structural equation modeling. Transformation techniques like the Fourier transform or the wavelet transform are heavily used in state-space modeling, but hardly figure in structural equation modeling. And these are just a few of the differences existing between the two approaches.

Yet, despite these differences, the commonalties between structural equation modeling and state-space modeling are manifold and intriguing. First and foremost, the common formal structure of state-space and structural equation models opens up the possibility to use linear algebraic results obtained for state-space models in the context of structural equation models. The algebraic results concerned constitute the field of realization theory and mainly involve the possible different ways in which the same behavior of a system can be realized by distinct model structures. As I will show in chapter 2 of this book, application of results from realization theory to structural equation models leads to quite surprising and, in my view, interesting outcomes. It will be shown, among other things, that the latent variables in latent growth curve models, common factor models and latent simplex models can be removed from these models. Hence these models can be rewritten as equivalent models lacking latent common factors or random weights. To the best of my knowledge this has not been accomplished before in the field of structural equation modeling. The possibility to remove latent factors and their likes from structural equation models has many interesting implications and casts a new light on the status of latent variables.

While the formal and statistical differences between structural equation modeling and state-space modeling do not appear to be fundamental, the same cannot be said of their fields of application. Structural equation modeling is mainly analysis of between-subject (between-system) variation, while state-space modeling is mainly analysis of within-subject (within-system) variation. An important field of research in dynamical systems theory is concerned with the relationship between results obtained in analyses of between-system variation and within-system variation. This part of dynamical systems theory is called ergodic theory. The classical ergodic theorems imply that analysis of between-system variation and within-system variation only yield equivalent results if the systems concerned obey strong restrictions. In chapter 3 I will argue that many populations figuring in social scientific research do not obey these restrictions. This lack of ergodicity has many consequences, perhaps the most profound one concerns classical test theory. This will be discussed at some length in chapter 3.

I have written this book with an audience of structural equation modelers in mind. Any attempt to apply aspects of mathematical systems theory to structural equation modeling requires the introduction of formal concepts and tools that may be new to many members of this audience. As to this, one can follow at least two distinct approaches: one in which the required formal concepts and tools are presented as completely as possible, or one in which this material is dealt with to the minimum degree possible. The first approach is in danger of distracting (or stretching) the attention of readers so much that the main ideas get lost in a forest of details. The second approach is in danger of providing too little detail, leading to incomplete understanding of the main ideas. I have chosen to introduce formal tools and concepts from mathematical systems theory and time series analysis with a minimum degree of elaboration. The reason is, that these concepts and tools mainly are used in an algebraic sense to manipulate structural equation models and transform these models into equivalent representations. For such formal manipulations it is not required to understand the various uses and interpretations of these tools and concepts. In addition, I will give ample references to the literature where all this is covered in great detail.

Most empirical illustrations given in this book are of the following form. a) A probability model is specified (always under the simplest possible assumptions of Gaussian random variables). b) The true covariance matrix under this model is generated. c) Model fits are carried out on this covariance matrix. In this way the equivalence of distinct models fitted in step c) can be illustrated. To prove their equivalence, algebraic derivations are carried out that sometimes become quite tedious. But the bottomline is, that (almost) all models considered in this book are linear Gaussian models, whether of the state-space or structural equation variety, where the latter can be handled by means of standard structural equation modeling software. In this book I use the Lisrel program for that purpose.

In rough outline, the contents of this book are as follows. In chapter 1 it is proven that the regression estimator for factor scores in a standard (cross-sectional) factor model is equivalent to the Kalman filter for the estimation of states in a state-space model. Even readers who are not particularly interested in this equivalence are advised to scan this chapter in order to get accustomed to the style of presentation and notation used in the remainder of this book. Chapter 2 is the largest chapter in the book and consists of three parts. In the first part it is proven that latent growth curve models and factor models are nested under latent simplex models. Nested means that latent growth curve models and factor models can be obtained by fixation of parameters in latent simplex models (and not the other way around). In the second part it is proven that the random weights in latent growth curve models, the latent factors in factor models, and the latent simplex in quasi-simplex models can be removed, yielding equivalent models without these latent variables. The tool for obtaining this result is a suitably generalized theorem from time series analysis, which then is applied to latent simplex models. Because latent growth curve models and factor models are nested under the latent simplex model, this generalized theorem also applies to these latent growth curve and factor models. Part 2 of chapter 2 is, I think, the most difficult part of this book, although I have tried to keep the discussion as transparent as possible (some of my students are using this material in their own research now). Also, all derivations are illustrated with numerical examples. In part 3 of chapter 2, it is shown that the removal of latent variables from structural equation models is sanctioned by theorems from realization theory.

Chapter 3 is devoted to the relationship (or lack thereof) between results obtained in analyses of between-subject variation (like in standard multivariate statistical analysis in the social sciences) and analysis of within-subject variation (I use the term variation in its dictionary definition: the degree to which something differs, for example, from a former state or value, from others of the same type, or from a standard). The importance of classical ergodic theorems for understanding the (lack of) relationship between analyses of N=many versus N=1 will be explained in a heuristic way. The often-surprising effects of heterogeneity in a population of subjects will be discussed, in particular regarding the foundations of classical test theory. Finally, in chapter 4 some unfinished business is taken care of (equivalence between regression estimator of longitudinal factor scores and Kalman smoother; state-space models versus state-space representations) and some conclusions and implications for further research are mentioned.

The central chapters of this book are chapters 2 and 3. It should be possible to read each chapter independently of each other (after getting acquainted with the style of presentation and notation). As I mentioned before, parts of chapter 2 and 3 are used by my students in their research. But in general this is not a standard textbook. It is more a display of (relatively) new ideas with possibly profound consequences for the field of structural equation modeling. These ideas should be considered as tentative conjectures; their definite treatment has to await (a lot of) further elaboration and research.

It is my pleasure to thank several persons for being instrumental in writing this book. My friends in the Methodology SuperCenter: John Nesselroade, Mike Rovine and Alex von Eye. My friends and colleagues in the Psychological Department at the University of Amsterdam, in particular Don Mellenbergh, Conor Dolan, Denny Borsboom, Han van der Maas and Ineke van Osch. A special thanks is due to the board of my department, for providing the excellent facilities to carry out an endeavor like this. Last, but in reality first, I thank my wife Madeleine and my two daughters, Charlotte and Françoise, for their love and comfort.

1. An introduction to the relationships between state-space modeling and structural equation modeling

In this chapter a first encounter is presented with the relationship between state-space models and structural equation models. This will provide the opportunity to introduce these types of models in a leisurely fashion. To wit, many details will have to be neglected in this way of presentation, but these will be introduced at appropriate places later on.

It is noted that some aspects of this relationship have been addressed in the published literature. MacCallum & Ashby (1986) discuss the relationships between linear systems theory and covariance structure modeling. Perhaps the first paper is due to Priestley & Subba Rao (1975), who rewrite the regression estimator (predictor) of factor scores in a standard factor model as a special instance of the Kalman filter (to be defined shortly) associated with a linear state-space model. In what follows I will present the result originally obtained by Priestley & Subba Rao, using a different derivation which, I think, is a bit more straightforward. Bold-face lower-case letters denote column vectors, while bold-face upper-case letters denote matrices. Transposition is denoted by the superscript '. Expectation is denoted by E[.] and covariance by cov(., .).

1.1 Ensembles

To facilitate comparison between state-space models and structural equation models, the concept of ensemble is introduced. For most readers this concept may not be familiar, so I will start with giving some additional background information.

The concept of ensemble arose in statistical mechanics, in particular in the innovative work of Gibbs (cf. Dorfman, 1999, chapter 5). It resembles the concept of the 3-way data box introduced in the psychometrical literature by Cattell (1946). The conceptual dimensions of Cattell's data box refer to, respectively, variables, persons, and times. In a similar vein an ensemble is a collection of time-dependent trajectories describing the dynamic behavior of a set of systems. The elements of this set are mutually interchangeable, i.e., identical in all relevant aspects.

Although the 3-way data box bears a resemblance to an ensemble, there are also important differences. An ensemble allows for interpretations that do not apply to the data box. For instance, an ensemble can be interpreted as the collection of all possible realizations of a stochastic function characterizing the behavior of a single system. This shows that an ensemble allows for a kind of intensional interpretations, whereas the data box always has extensional meaning (namely the 3-way lay-out of actual data). Also, an ensemble is intrinsically stochastic, whereas the typical 3-way data box is considered to be fixed (but see Bentler & Lee, 1979). Finally, the concept of ensemble is standard in the mathematical-statistical theory which will be used later in this book and therefore it is helpful to use it consistently.

Consider a single system S, the time-dependent behavior of which is represented by the p-variate vector-valued function $\mathbf{y}_S(t)$, t=0,±1,±2,..., where the dimension p is finite ($1 \leq p < \infty$). We can picture $\mathbf{y}_S(t)$ as a trajectory in (p+1)-

dimensional space (where t is the extra dimension). Following Brillinger (1975, section 2.11) we also can conceive of $\mathbf{y}_S(t)$ as a particular realization of a random function $\mathbf{y}_S(\omega,t)$, $\omega\in\Omega$, where Ω is an arbitrary space or set of "outcomes". After defining a probability measure on Ω, we obtain an ensemble of possible trajectories of which the actual realization $\mathbf{y}_S(t)$ constitutes a sample. It will explained shortly that this ensemble of possible trajectories is the basic domain for state-space modeling as well as structural equation modeling.

We can regard the ensemble of possible trajectories as a covering of a (p+1)-dimensional space, the density of which is obtained from the probability measure defined on Ω. From another point of view, this ensemble can be regarded as the collection of trajectories of copies of system S which are interchangeable with S in all relevant aspects. These two alternative points of view (cf. Caines, 1988, for a clear pictorial presentation) bear a resemblance to the distinction between the stochastic subject formulation and the random sampling formulation of latent variable models (e.g., Ellis & Junker, 1997). It then makes sense to denote the ensemble of trajectories by $\{\mathbf{y}_i(t), t=0,\pm1,...; i=1,2,\ldots\}$, under the obvious condition that different systems i do not interact (because they have to generate possible trajectories under the alternative stochastic subject point of view).

1.2 Standard factor models

For the moment the ensemble $\{\mathbf{y}_i(t), t=0,\pm1,...; i=1,2,\ldots\}$ is regarded as the common domain of both state-space modeling and structural equation modeling. The structural model which will be considered is the standard factor model at a fixed time t_1:

$$(1.1) \qquad \mathbf{y}_i(t_1) = \boldsymbol{\Lambda}_{t1}\boldsymbol{\eta}_i(t_1) + \boldsymbol{\varepsilon}_i(t_1), \text{ or } \mathbf{y}_i = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i, \; i=1,2,\ldots$$

where in the second expression the redundant fixed time index has been omitted. It is understood that the expectation of $\mathbf{y}_i$ with respect to the probability measure over systems (subjects) i at time t_1 is zero: $E[\mathbf{y}_i] = 0$. $\boldsymbol{\eta}_i$ denotes a q-variate common factor and $\boldsymbol{\varepsilon}_i$ p-variate measurement error. $\boldsymbol{\Lambda}$ is the (p,q)-dimensional matrix of factor loadings. To complete the description of the standard factor model, simple distributional assumptions are introduced for $\boldsymbol{\eta}_i$ and $\boldsymbol{\varepsilon}_i$: $\boldsymbol{\eta}_i \sim\aleph(\mathbf{0}, \boldsymbol{\Phi})$ and $\boldsymbol{\varepsilon}_i \sim\aleph(\mathbf{0}, \boldsymbol{\Theta})$, where $\aleph(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. It is assumed that $\boldsymbol{\Theta}$ is diagonal (mutually uncorrelated measurement errors). It follows from these distributional assumptions and the assumption that $\text{cov}[\boldsymbol{\eta}_i, \boldsymbol{\varepsilon}_i] = \mathbf{0}$ that $\mathbf{y}_i \sim\aleph(\mathbf{0}, \boldsymbol{\Sigma}_y)$, where $\boldsymbol{\Sigma}_y = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$.

1.3 Linear state-space models

The state-space model which will be considered is the linear state-space model for a fixed system i_1:

(1.2) $\mathbf{y}_{i1}(t) = \Lambda_{i1}\boldsymbol{\eta}_{i1}(t) + \boldsymbol{\varepsilon}_{i1}(t)$, or $\mathbf{y}(t) = \Lambda\boldsymbol{\eta}(t) + \boldsymbol{\varepsilon}(t)$, $t=0,\pm1,\ldots$

where in the second expression the redundant system index i_1 has been omitted. For the moment it is understood that the expectation of $\mathbf{y}(t)$ with respect to the probability measure over times t for system i_1 is zero: $E[\mathbf{y}(t)] = 0$. $\boldsymbol{\eta}(t)$ denotes a q-variate latent factor series, $\boldsymbol{\varepsilon}(t)$ a p-variate measurement error series, which will be defined shortly. Λ is a (p,q)-dimensional matrix of factor loadings. In contrast to the standard factor model (1.1) where the index i refers to an unordered set of systems (i.e., the standard factor model is invariant under permutations of the set of systems), the time index t in (1.2) refers to an ordered set of realizations (i.e., permutation of this set is not allowed). The ordering of t allows for the representation of sequential dependencies, i.e., the introduction of (stochastic) difference equations expressing local dynamic relationships. Presently, only the simplest possible dynamic relationship will be considered, namely a stochastic difference equation relating $\boldsymbol{\eta}(t)$ to $\boldsymbol{\eta}(t-1)$:

(1.3) $\boldsymbol{\eta}(t) = \mathbf{B}\boldsymbol{\eta}(t-1) + \boldsymbol{\zeta}(t)$, $t=0,\pm1,\ldots$

where $\mathbf{B}$ is a (q,q)-dimensional matrix of regression coefficients and $\boldsymbol{\zeta}(t)$ is called a q-variate innovations process which lacks any sequential dependency.

In the definition of (1.2), $\boldsymbol{\eta}(t)$ and $\boldsymbol{\varepsilon}(t)$ are referred to as (time) series, while $\boldsymbol{\zeta}(t)$ in (1.3) is called a (random or stochastic) process. Also $\mathbf{y}(\omega,t) = \mathbf{y}(t)$ is called a stochastic or random function. I will use the terms (time series, stochastic process, random function) interchangeably. It is acknowledged that these terms may have slightly different connotations in more technical contexts, but within the context of this book such differences are not important. The terms refer to a time-dependent probabilistic structure which can be summarized as follows.

A multivariate stochastic process $\mathbf{x}(t)$ in discrete time t is characterized by a set (a so-called cylinder set; cf. Billingsley, 1995) of finite-dimensional distributions $P(\mathbf{x};t) =\text{Prob}[\mathbf{x}(t)<\mathbf{x}]$, $P(\mathbf{x}_1,\mathbf{x}_2;t_1,t_2) = \text{Prob}[\mathbf{x}(t_1)<\mathbf{x}_1;\mathbf{x}(t_2)<\mathbf{x}_2]$, etc. Accordingly, we can consider the first-order moment function (of time) of $\mathbf{x}(t)$, the second-order moment function of $\mathbf{x}(t)$, etc. In general these moment functions can be time-varying, i.e., they can depend upon the time t (first-order moment function), the pair of times t_1, t_2 (second-order moment function), etc. If, however, the first-order moment function is a constant, $E[\mathbf{x}(t)]=\mathbf{c}_x$, then $\mathbf{x}(t)$ is called first-order stationary. If its second-order central moment function only depends upon the lag between t_1 and t_2, $E[(\mathbf{x}(t_1)-\mathbf{c}_x(t_1)),(\mathbf{x}(t_2)-\mathbf{c}_x(t_2))'] = \text{cov}[\mathbf{x}(t_1),\mathbf{x}(t_2)'] = \mathbf{C}_x(u)$, $u = t_2-t_1$, then $\mathbf{x}(t)$ is called second-order stationary. If $\mathbf{x}(t)$ is both first- and second-order stationary then it is called weakly stationary.

The definition of the probability structure of weakly stationary time series is rather limited in that only the first-order and second-order moment functions are involved and because these are considered to be invariant under time shifts. Yet for the present purposes the assumption of weak stationarity suffices (in later sections much more general probability structures will be considered). It allows for the specification of the distributional assumptions associated with (1.2) and (1.3). Starting with the latter, it is sufficient to specify the probability structure of $\boldsymbol{\zeta}(t)$: $\boldsymbol{\zeta}(t) \sim\aleph(\mathbf{0}, \boldsymbol{\Psi})$ and $\text{cov}[\boldsymbol{\zeta}(t), \boldsymbol{\zeta}(t+u)] = \mathbf{C}_\zeta(u) = \delta(u)\boldsymbol{\Psi}$, $u=0,\pm1,\ldots$, where the Kronecker delta $\delta(u)$ equals 1 if u=0 and equals zero otherwise. The covariance function $\mathbf{C}_\zeta(u)$ of $\boldsymbol{\zeta}(t)$ is only nonzero if the lag u is zero: $\mathbf{C}_\zeta(0) = \boldsymbol{\Psi}$. Hence $\boldsymbol{\zeta}(t)$ lacks any sequential dependency and can be regarded as a series of independent draws from $\aleph(\mathbf{0}, \boldsymbol{\Psi})$. I will refer to this kind of weakly stationary series lacking sequential dependency as white noise. Hence $\boldsymbol{\zeta}(t)$ is called Gaussian q-variate white noise. The denotation white noise arose in the engineering sciences in view of the fact that the spectrum (the Fourier transform of the covariance function) of such a series is constant across all frequencies, like the spectrum of white light.

Using (1.3) the probability structure of $\boldsymbol{\eta}(t)$ can be derived from the distributional assumptions about $\boldsymbol{\zeta}(t)$ and the additional assumption that cov[$\boldsymbol{\zeta}(t)$, $\boldsymbol{\eta}(t')$] = $\mathbf{0}$ for t $\geq$ t'. Hence there is no need to introduce distributional assumptions about $\boldsymbol{\eta}(t)$. To give a simple illustration: suppose q = 1. Then (1.3) reduces to: $\eta(t) = \beta\eta(t-1) + \zeta(t)$, where $|\beta| < 1$ to guarantee weak stationarity. Then cov[$\eta(t)$, $\eta(t)$] = $c_\eta(0) = E[\eta(t)^2] = E[\{\beta\eta(t-1) + \zeta(t)\}^2]$. Expanding the final expression, using that cov[$\eta(t-1)$, $\zeta(t)$] = 0, one obtains: $E[\{\beta\eta(t-1) + \zeta(t)\}^2] = \beta^2 c_\eta(0) + \psi$. This expression equals $E[\eta(t)^2] = c_\eta(0)$. Rearranging yields: $c_\eta(0) = \psi / (1 - \beta^2)$. For positive lags u unequal to zero, u > 0, one obtains using (1.3) again: cov[$\eta(t)$, $\eta(t+u)$] = $c_\eta(u)$ = $E[\eta(t), \beta\eta(t+u-1) + \zeta(t+u)]$. Expanding the final expression and using cov[$\eta(t)$, $\zeta(t+u)$] = 0 and cov[$\eta(t+u-1)$, $\zeta(t+u)$] = 0, one gets: $c_\eta(u) = \beta c_\eta(u-1)$. This recursive equation, with $c_\eta(0) = \psi / (1 - \beta^2)$ as initial condition, yields the covariance function of $\eta(t)$ for all lags u $\geq$ 0. An easy argument shows that the covariance function of any univariate weakly stationary time series, and hence $c_\eta(u)$ too, is symmetrical about lag zero.

The final distributional assumption concerns the measurement error $\boldsymbol{\varepsilon}(t)$ in (1.2): $\boldsymbol{\varepsilon}(t) \sim \aleph(\mathbf{0}, \boldsymbol{\Theta})$ and cov[$\boldsymbol{\varepsilon}(t)$, $\boldsymbol{\varepsilon}(t+u)$] = $\mathbf{C}_\varepsilon(u) = \delta(u)\boldsymbol{\Theta}$, u=0,±1,… Hence $\boldsymbol{\varepsilon}(t)$ is p-variate Gaussian white noise. It is assumed that $\boldsymbol{\Theta}$ is diagonal. It then follows, using the additional assumption that cov[$\boldsymbol{\varepsilon}(t)$, $\boldsymbol{\zeta}(t')$] = $\mathbf{0}$ for all t and t', that $\mathbf{y}(t)$ is a p-variate Gaussian time series with mean zero and covariance function: cov[$\mathbf{y}(t)$, $\mathbf{y}(t+u)$] = $\mathbf{C}_y(u) = \boldsymbol{\Lambda}\mathbf{C}_\eta(u)\boldsymbol{\Lambda}' + \delta(u)\boldsymbol{\Theta}$, u=0,±1,…

1.4 Estimation of latent states and factors

First a matter of notation. Let $\mathbf{x}(t)$, t=0,±1,... be a time series, i.e., a probability structure as defined in the previous section. A sample from this time series constitutes a stretch of values $\mathbf{x}(1)=\mathbf{x}_1$, $\mathbf{x}(2)=\mathbf{x}_2$, ..., $\mathbf{x}(T)=\mathbf{x}_T$. Such a sample will be denoted by: {$\mathbf{x}(t)$, t=1,...,T}. Of course this latter notational convention is not correct (a finite part of one trajectory or realization of $\mathbf{x}(t)$ is denoted by $\mathbf{x}(t)$ itself), but is common practice in the time series literature and can be used without causing ambiguity.

Priestley & Subba Rao (1975) present a derivation of the Kalman filter, associated with the linear state-space model, from the regression estimator of factor scores associated with the standard factor model. The Kalman filter is an estimator of the state process $\boldsymbol{\eta}(t)$ given a set of observations {$\mathbf{y}(k)$, k=1,…,t} and given the model equations (1.2) and (1.3). That is, it is assumed that $\boldsymbol{\Lambda}$, $\boldsymbol{\Theta}$, $\mathbf{B}$, and $\boldsymbol{\Psi}$ are known. Then the Kalman filter constitutes a recursive estimator of {$\boldsymbol{\eta}(t)$, t=1,…,T}.

A recursive estimator makes use of the ordering of the time axis and can be schematically represented as: estimate at time t = F[estimate at time t-1 and the observation at time t], where F[.] denotes some appropriate function. Take for instance the simple example of the estimator of the mean of {$\mathbf{y}$(t), t=1,…,T}. The standard estimator is: $\mathbf{m_y} = T^{-1}\Sigma_{t=1,T}\mathbf{y}(t)$. Following Koopmans (1995) I will call this kind of estimator a batch estimator: it is an estimator using at once the complete batch of observations {$\mathbf{y}$(t), t=1,…,T}. Denote the equivalent recursive estimator at time t by: $\mathbf{m}_y(t \mid t)$. This notation expresses that the recursive estimator at time t is conditional on the information obtained up to and including time t:

$$(1.4) \qquad \mathbf{m}_y(t \mid t) =: \mathbf{m}_y(t \mid \mathbf{y}(1), \mathbf{y}(2), ..., \mathbf{y}(t)),$$

where =: denotes an equality by convention. It is an easy exercise to determine that $\mathbf{m}_y(t \mid t) = (t-1)/t\ \mathbf{m}_y(t-1 \mid t-1) + \mathbf{y}(t)/t$, t=1,2,...,T; $\mathbf{m}_y(0 \mid 0) = \mathbf{0}$.

Using the notation of (1.4), the Kalman filter for the state-space model (1.2)-(1.3) is given by:

$$\eta(t \mid t) = \mathbf{B}\eta(t-1 \mid t-1) + \mathbf{K}(t)[\mathbf{y}(t) - \mathbf{\Lambda B}\eta(t-1 \mid t-1)]$$

$$\mathbf{K}(t) = \mathbf{V}(t \mid t-1)\Lambda'[\Lambda\mathbf{V}(t \mid t-1)\Lambda' + \Theta]^{-1}$$

$$(1.5) \qquad \mathbf{V}(t+1 \mid t) = \mathbf{BV}(t \mid t)\mathbf{B}' + \Psi$$

$$\mathbf{V}(t \mid t) = [\mathbf{I_q} - \mathbf{K}(t)\Lambda]\mathbf{V}(t \mid t-1)$$

$$\eta(0 \mid 0) = \mu_0,\ \mathbf{V}(0 \mid 0) = \mathbf{V_0}$$

where $\mathbf{I_q}$ denotes the (q,q)-dimensional unity matrix. The set of recursive equations (1.5) yields the estimated trajectory $\eta(t \mid t)$, t=1,...,T of the state process $\eta(t\ )$ in (1.2)-(1.3) associated with the observations {$\mathbf{y}$(t), t=1,...,T}. It works rather simple. At time t=1, first $\mathbf{V}(1 \mid 0)$ is computed according to the third equation in (1.5): $\mathbf{V}(1 \mid 0) = \mathbf{BV_0B}' + \Psi$. Next $\mathbf{K}(1)$ is computed according to the second equation in (1.5), after which $\eta(1 \mid 1)$ can be determined according to the first equation in (1.5) using $\eta(0 \mid 0) = \mu_0$ and the observed value of $\mathbf{y}(1)$. Finally, $\mathbf{V}(1 \mid 1)$ is determined according to the fourth equation in (1.5). The values thus obtained at t=1 for $\eta(1 \mid 1)$ and $\mathbf{V}(1 \mid 1)$ constitute the starting values for the computations at t=2, etc.

The estimate $\eta(t \mid t)$ is composed of two parts: $\mathbf{B}\eta(t-1 \mid t-1)$, which is the predicted value based on the information available at time t-1, and $\mathbf{K}(t)[\mathbf{y}(t) - \mathbf{\Lambda B}\eta(t-1 \mid t-1)]$, which is the correction based on the new information $\mathbf{y}(t)$ entering at time t. In the latter correction component the term $\mathbf{\Lambda B}\eta(t-1 \mid t-1)$ is the predicted value $\mathbf{y}(t \mid t-1)$ of $\mathbf{y}(t)$ based on the information available at time t-1. The (q,p)-dimensional weight matrix $\mathbf{K}(t)$ in the correction component is the famous Kalman gain.

The (q,q)-dimensional matrix $\mathbf{V}(t \mid t)$ is $E[(\eta(t \mid t) - \eta(t)), (\eta(t \mid t) - \eta(t))' \mid \mathbf{Y}^t]$, the covariance matrix of $\eta(t \mid t)$ conditional on $\mathbf{Y}^t$, the information available up to time t. The (q,q)-dimensional matrix $\mathbf{V}(t \mid t-1)$ is $E[(\eta(t \mid t-1) - \eta(t)), (\eta(t \mid t-1) - \eta(t))' \mid \mathbf{Y}^{t-1}]$, the covariance of $\eta(t \mid t-1)$ conditional on $\mathbf{Y}^{t-1}$, the information available up to time t-1. The reader is referred to Caines (1988, p. 161) for the delicate difference between conditional and unconditional covariance matrices (see also Anderson & Moore, 1979). In view of the Gaussian distribution assumptions with respect to (1.2)-(1.3) we can for the time being neglect this difference. The initial conditions $\eta(0 \mid 0) = \boldsymbol{\mu_0}$ and $\mathbf{V}(0 \mid 0) = \mathbf{V_0}$ are assumed to be given (but see the final section of this chapter).

Before closing this initial encounter with the Kalman filter, it has to be determined what kind of estimator it is. Note that any type of estimator, whether derived according to (generalized) least-squares techniques, maximum likelihood techniques or the Bayesian approach, can be (and has been) casted in recursive form. From this point of view the recursive form merely is a computational aspect. Hence the recursive form of the Kalman filter does not provide a clue to what kind of criterion or discrepancy function is optimized by it. Jazwinsky (1970, p. 201, Theorem 7.2) shows that the Kalman filter associated with our state-space system (1.2)-(1,3) constitutes the minimum variance filter for the state process $\eta(t)$. He also presents the derivation of the Kalman filter according to the maximum likelihood technique (op. cit., pp. 207-208). Sage & Melsa (1971, pp. 272-283) derive the Kalman filter according to the Bayesian approach. Caines (1988, p. 158, Theorem 4.1) shows that the Kalman filter associated with our state-space system, requiring only that $\zeta(t)$ and $\varepsilon(t)$ are weakly stationary white noise series (hence dropping the assumptions of Gaussianity) constitutes the linear least squares filter for the state process $\eta(t)$.

We now turn to the estimation of the realizations of η_i in the standard factor model (1.1) given a set of observations $\{\mathbf{y}_i$, i=1,...,N} and given Λ, Φ and Θ. It is sometimes argued that calling this "estimation" of factor scores involves incorrect terminology and that "prediction" of factor scores should be used instead (e.g., Bartholomev, 1987). I take no position on this issue, but will use factor score "estimation" only because it appears to be the more common denotation. The regression estimator for η_i is given by (e.g., Lawley & Maxwell, 1971, p. 109):

$$\eta(i \mid \mathbf{y}_i) = \Phi\Lambda'\Sigma^{-1}\mathbf{y}_i \text{ , } i=1,...,N$$

(1.6)

$$E[(\eta(i \mid \mathbf{y}_i) - \eta_i), (\eta(i \mid \mathbf{y}_i) - \eta_i)'] = \Phi(\mathbf{I}_q + \Lambda'\Theta^{-1}\Lambda\Phi)^{-1}$$

Apart from the regression estimator there are several other kinds of factor score estimators. These will not be considered here, as Priestley & Subba Rao (1975) only discuss the relationship between (1.6) and (1.5).

1.5 The regression estimator as Kalman filter

I will now show that the regression estimator for factor scores constitutes a special case of the Kalman filter. Specifically, it will be proven that (1.6) is a special case of (1.5). But first I will outline the proof of Priestley & Subba Rao (1975).

In fact, Priestley & Subba Rao present a proof in the reverse direction, namely that the Kalman filter can be rewritten as the regression estimator. Their proof consists of three steps. In the first step the state-space model is rewritten as a standard factor model (1.1). This is accomplished by means of the following transformations:

$$\mathbf{y}^*(t) = \mathbf{y}(t) - E[\mathbf{y}(t) \mid \mathbf{Y}^{t-1}],\ \boldsymbol{\eta}^*(t) = \boldsymbol{\eta}(t) - E[\boldsymbol{\eta}(t) \mid \mathbf{Y}^{t-1}]$$

where $\mathbf{Y}^{t-1}$ denotes the set of variables {$\mathbf{y}$(t-1), $\mathbf{y}$(t-2), ...}. Given (1.2)-(1.3) it follows that $\mathbf{y}^*(t)$ and $\boldsymbol{\eta}^*(t)$ are Gaussian white noise processes lacking any sequential dependencies and that substitution of $\mathbf{y}^*(t)$ and $\boldsymbol{\eta}^*(t)$ in (1.2)-(1.3) transforms the state-space model into a standard factor model. Hence for this transformed state-space model the regression estimator is a valid estimator of $\boldsymbol{\eta}^*(t)$. In their second step Priestley & Subba Rao apply the inverse of their original transformation (yielding $\mathbf{y}(t)$ and $\boldsymbol{\eta}(t)$ again) to the regression estimator, which then is shown to become identical to the Kalman filter.

In what follows I will show that the standard factor model constitutes a special instance of the state-space model, apply the Kalman filter to the standard factor model, and then show that it reduces to the regression estimator. This approach precludes the need to introduce the transformations and their inverses which is necessary in the approach taken by Priestley & Subba Rao (1975). I will also extend the proof of Priestley & Subba Rao and show the equivalence of the covariance matrices of the Kalman filter estimate and the regression estimate. The latter equivalence is not discussed in Priestley & Subba Rao's paper.

To understand that the standard factor model (1.1) is a special instance of the state-space model (1.2)-(1.3), it may be helpful to remember the definition of ensemble given in section 1.1. An ensemble is defined as a set of trajectories {$\mathbf{y}_i$(t), t=0,±1,...; i=1,2,…}, where tractories of different systems i and i', $i \neq i'$, are mutually independent (do not influence each other). The standard factor model is defined with respect to this ensemble by taking a fixed time point t_1 and then describe the variation between different realizations (values of different trajectories) i=1,2,... . This can be interpreted as defining the same state-space model for each realization $\mathbf{y}_i(t_1)$, i=1,2,..., where each trajectory is observed at the same single fixed time point t_1. It then is obvious that (1.3) reduces to: $\boldsymbol{\eta}_i(t_1) = \boldsymbol{\zeta}_i(t_1)$, i=1,2,…, because observations only are available at a single time point t_1 and therefore it makes no sense to consider relationships between consecutive time points. Consequently, $\mathbf{B}$ is fixed at $\mathbf{B} = 0$ in (1.3).

It follows from this reasoning (which is also implicitly used in Priestley & Subba Rao's (1975) paper) that application of the Kalman filter to the standard factor model cum state-space model reduces to a single recursion step for each i=1,2,..., given the model parameters and initial values. To ease the presentation it is assumed that $t_1 = 1$. Obviously, this is an inconsequential notational convention. The initial values are: $\boldsymbol{\eta}(0 \mid 0) = \mathbf{0}$ (because $E[\boldsymbol{\eta}_i] = 0$) and $\mathbf{V}(0 \mid 0) = \boldsymbol{\Phi}$ (because $\mathrm{cov}[\boldsymbol{\eta}_i, \boldsymbol{\eta}_i] = \boldsymbol{\Phi}$). Together with the restriction that $\mathbf{B} = \mathbf{0}$, it follows that (1.5) can be rewritten for each i=1,2,... as:

$$\boldsymbol{\eta}_i(1 \mid 1) = \mathbf{K}(1)\mathbf{y}_i(1)$$

$$\mathbf{K}(1) = \boldsymbol{\Phi\Lambda}'[\boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{\Theta}]^{-1}$$

(1.7)

$$\mathbf{V}(1 \mid 0) = \Phi$$

$$\mathbf{V}(1 \mid 1) = [\mathbf{I}_q - \mathbf{K}(1)\Lambda]\Phi$$

It is immediately clear from the second equation in (1.7) that the Kalman gain $\mathbf{K}(1)$ equals the regression estimator (1.6): $\mathbf{K}(1) = \Phi\Lambda'[\Lambda\Phi\Lambda' + \Theta]^{-1} = \Phi\Lambda'\Sigma^{-1}$. Hence $\eta_i(1 \mid 1) = \eta(i \mid \mathbf{y}_i)$.

This derivation has been accomplished in a rather effortless way, circumventing the transformations needed in the derivation given by Priestley & Subba Rao (1975). Moreover, it allows for the simple execution of an additional step, not considered by Priestley & Subba Rao, namely a proof of the equivalence of $\mathbf{V}(1| 1)$ in (1.7) and $E[(\eta(i \mid \mathbf{y}_i) - \eta_i), (\eta(i \mid \mathbf{y}_i) - \eta_i)']$ in (1.6). Again this proof is obtained quite effortless. First, using $\mathbf{K}(1) = \Phi\Lambda'\Sigma^{-1}$, $\mathbf{V}(1 \mid 1)$ in (1.7) is rewritten as: $\mathbf{V}(1 \mid 1) = \Phi - \Phi\Lambda'\Sigma^{-1}\Lambda\Phi$. Second, Lawley & Maxwell (1971, p.109) show that $E[\eta(i \mid \mathbf{y}_i), \eta(i \mid \mathbf{y}_i)] = \Phi\Lambda'\Sigma^{-1}\Lambda\Phi = \Phi - \Phi(\mathbf{I}_q + \Lambda'\Theta^{-1}\Lambda\Phi)^{-1}$.Hence, $E[(\eta(i \mid \mathbf{y}_i) - \eta_i), (\eta(i \mid \mathbf{y}_i) - \eta_i)'] = \Phi(\mathbf{I}_q + \Lambda'\Theta^{-1}\Lambda\Phi)^{-1} = \Phi - E[\eta(i \mid \mathbf{y}_i), \eta(i \mid \mathbf{y}_i)]$. Substitution of the equality $E[\eta(i \mid \mathbf{y}_i), \eta(i \mid \mathbf{y}_i)] = \Phi\Lambda'\Sigma^{-1}\Lambda\Phi$ in the latter expression yields: $E[(\eta(i \mid \mathbf{y}_i) - \eta_i), (\eta(i \mid \mathbf{y}_i) - \eta_i)'] = \Phi - \Phi\Lambda'\Sigma^{-1}\Lambda\Phi$. Hence $\mathbf{V}(1 \mid 1) = E[(\eta(i \mid \mathbf{y}_i) - \eta_i), (\eta(i \mid \mathbf{y}_i) - \eta_i)']$. QED.

1.6 Discussion

The result obtained in this chapter, namely a proof that the regression estimator (1.6) of factor scores in the standard factor model (1.1) constitutes a special case of the Kalman filter (1.5) associated with the state-space model (1.2)-(1.3), should be qualified in a number of ways. Firstly, it is only a variation of the innovative work of Priestley & Subba Rao (1975). Using a remark made in Jazwinsky (1970, p. 209): "It is easy to rederive a known result". Secondly, both the standard factor model and the state-space model considered in this chapter have rather special properties, such as Gaussian distributional assumptions and known model parameters. In chapter 4 I will consider the same kind of relationship under more general formulations of the models, in particular allowing for uncertain (i.e., estimated) model parameters. Thirdly, it would be interesting to consider other kinds of factor models, in particular longitudinal factor models. In chapter 4 it will also be shown that this leads to a more intricate kind of relationship between the regression estimator for longitudinal factor scores and a variant of the Kalman filter.

Apart from the opportunity to consider a particular kind of relationship between structural equation models and state-space models, this chapter also served as a (hopefully) smooth introduction of notational conventions and technicalities which will be used in the remainder of this book. For this reason I did not provide any numerical illustrations in this chapter. The concepts of an ensemble, the linear Gaussian state-space model and the Kalman filter will recur at several places in what follows. We should now be prepared enough to successfully tackle our first main problem in the next chapter.

2. A stepwise proof that factor models and latent growth curve models are nested under quasi-simplex models, followed by a general scheme to rewrite latent variable models as equivalent models with only observed variables

During a large part of the last decade of the previous century I had a part-time assignment at the Pennsylvania State University. I was, and still am, very grateful for getting the opportunity to learn about, and participate in, academic life typical of a renown university in the US. One recurring aspect of that life, which still is relatively uncommon at Dutch universities, involved the proceedings of selecting a candidate for a vacant position. For each position, several candidates were invited to spend a few days at the department concerned and present a lecture. Because my assignment was at a methodology department, I witnessed many lectures of candidates for one or the other vacant position related to methodology. I became increasingly surprised by the number of lectures devoted to hierarchical linear modeling, in particular latent growth curve modeling of longitudinal data (this model will be specified below). For some positions almost all candidates presented a lecture using some innovative variant of the latent growth curve model or a particularly interesting application of the latent growth curve model. Yet the descriptions of the vacant positions concerned did not restrict the expertise of candidates in this severe way. What, then, could explain the popularity of this particular type of model?

In the beginning of the nineties I could not find an answer to this question. To make my question somewhat more explicit, I explained in my courses on structural equation modeling that the latent growth curve model constitutes a special instance of a confirmatory structural equation model. I also indicated that general hierarchical modeling constitutes a special instance of the confirmatory multi-group structural equation model (cf. Lee & Poon, 1993, for a more recent overview). By showing that these popular models are special instances of a general model which includes numerous alternative variants as special cases, I hoped to convey more clearly my questions about the seemingly arbitrary focus on one particular kind of variant. In retrospect, these efforts constituted the starting point of the contents of the present chapter.

I still do not have a satisfactory explanation for the apparent popularity of hierarchical linear models and latent growth curve models. But this is immaterial for the contents of this chapter, because it will not play any further role. In this rather long chapter I will first prove that the standard factor model and the latent growth curve model both are special instances of the quasi-simplex model. This part builds on a recent paper by Rovine & Molenaar (2001). To some readers it may come as a surprise that the standard factor model and the latent growth curve model are nested under the quasi-simplex model, because until now it appears to have been commonly accepted that at least the latent growth curve model and the quasi-simplex model constitute rather different alternative models for longitudinal data (cf. Rogosa & Willett, 1985). Yet the proof is surprisingly simple and only requires a suitable generalization of a technical theorem in time series analysis. The rest of the proof is straightforward matrix algebra. I will also give several illustrations using simulated data, and occasionally present an application to empirical data. Moreover, I will pay due attention to somewhat older work in psychometrics in which serious doubts have been raised about the appropriateness of the standard factor model for the analysis of longitudinal data. For some unjustified reason, this work seems to have been forgotten.

In the next major part of this chapter I will use the same technical theorem in time series analysis to rewrite a large class of latent variable models as models involving only manifest variables. To the best of my knowledge, this result is new in the methodology of behavioral and social sciences (excluding perhaps econometrics). Again the proof is rather straightforward. I will not strive to maximum generality, but mainly show the way in which proofs of this kind can be given. Again illustrations using simulated data will be given. Also some implications of this result will be discussed, in particular regarding ongoing discussions about the reality, identifiability and estimability of latent variables.

In the final major part of this chapter it will be shown that the results obtained in the previous parts may be considered to be new in the behavioral and social sciences, but can be derived from general theorems in mathematical systems theory and hence are common knowledge in the engineering sciences. This part is necessarily more technical, but I will try to convey the gist of the argumentation as clearly as possible. A preliminary remark about terminology. The quasi-simplex model to be introduced soon will also be referred to as the univariate latent simplex. The latter denotation does not appear to be common in the literature, but it easily generalizes to higher dimensions, like the latent bivariate simplex, etc.

2.1 The standard factor and latent growth curve models as special instances of the quasi-simplex model

Our starting point is the definition of the quasi-simplex model. Then it is shown that the standard 1-factor model constitutes a special instance of the quasi-simplex model. This result is generalized to the standard multi-factor model. Next it is made explicit that the latent growth curve model is itself a special instance of the confirmatory q-factor model (where q usually, but not always, is taken to be q=2). Because of the nesting of factor models under the quasi-simplex model, it follows that also the latent growth curve model, regarded as a special instance of the factor model, is nested under the quasi-simplex model.

2.1.1 The quasi-simplex model

The simplex model has a rather long history in psychometrics, which I will neglect here in order to keep the discussion focussed. It was Jöreskog (1970) who put the quasi-simplex as we currently know it on the map. The reader is referred to Jöreskog & Sörbom (1989) for a detailed description of this model. In this section the quasi-simplex in its simplest form is defined, after which a few elaborations are given which will be used in later sections.

Consider an ensemble of univariate trajectories $\{y_i(t), t=0,\pm1,...; i=1,2,\dots\}$ and choose a set of T fixed time points $t_1 < t_2 < ... < t_T$. To avoid clumsy notation these fixed times will be denoted by t=1,2,...,T. It should be kept in mind, however, that in this section time points are treated as fixed indices (time is not considered to be a domain of generalization), whereas this is not the case when dealing with time series.

Consequently, although the present notation resembles the notation used for time series and the state-space model, I hope that the benefits of easy notation will outweigh the possibility of ambiguity. The standard quasi-simplex model defined for $\{y_i(t), t=1,...,T; i=1,2,\ldots\}$, assuming that $E[y_i(t)] = 0$ for t=1,...,T, is defined by

$$y_i(t) = \eta_i(t) + \varepsilon_i(t), \ t=1,...,T$$

(2.1)

$$\eta_i(t) = \beta_{t,t-1}\eta_i(t-1) + \varsigma_i(t), \ t=2,...,T$$

The first equation in (2.1) describes at each time point t a decomposition of the observed variable $y_i(t)$ into measurement error $\varepsilon_i(t)$ and an error-free component $\eta_i(t)$. The second equation in (2.1) describes the regression of $\eta_i(t)$ on $\eta_i(t-1)$. Notice that no regression of $\eta_i(1)$ on $\eta_i(0)$ is possible, because $y_i(0)$ is not available. This can be expressed by the "empty" equation: $\eta_i(1) = \varsigma_i(1)$. Another way to express the unavailability of $y_i(0)$, and hence of $\eta_i(0)$, is to regard the second equation in (2.1) as a stochastic difference equation, the solution of which requires the independent specification of initial conditions (following up on the latter interpretation, one could conceive of the solution of this equation in terms of stochastic Green's functions; e.g., Keilson, 1965). A heuristic interpretation, which for instance has been given within the context of longitudinal behavior genetics in which $\eta_i(t)$ refers to additive genetical influences (Boomsma & Molenaar, 1987), is to regard $\beta_{t,t-1}\eta_i(t-1)$ as the effects of those genes which are "turned on" on both time t-1 and time t. Following this heuristic interpretation, $\varsigma_i(t)$ then can be regarded as the effects of genes which are "turned on" for the first time at time t. This explains the technical term for $\varsigma_i(t)$: the innovation at time t. Hence the complete second equation in (2.1) expresses $\eta_i(t)$ as composed of a locally stable part $\beta_{t,t-1}\eta_i(t-1)$ and an innovative part $\varsigma_i(t)$.

Before proceeding I have to voice again a word of caution. The form of (2.1) closely resembles the form of the state-space model (1.2)-(1.3). If p = q = 1 and Λ is taken to be the identity, then this results in a representation that is akin to (2.1). Yet in (2.1) the time points are fixed and the model describes the variation between systems (subjects) i=1,2,… . In contrast, in (1.2)-(1.3) systems (subjects) are fixed and the model describes the variation within each system over time (within-subjects variation). As will become evident in the next chapter, these differences can have far-reaching consequences and hence these distinct implications of the two representations should be kept in mind, despite their formal resemblance.

Introducing the vectors $\mathbf{y}_i = [y_i(1),...,y_i(T)]'$, $\boldsymbol{\varepsilon}_i = [\varepsilon_i(1),...,\varepsilon_i(T)]'$, $\boldsymbol{\eta}_i = [\eta_i(1),...,\eta_i(T)]'$ and $\boldsymbol{\varsigma}_i = [\varsigma_i(1),...,\varsigma_i(T)]'$, the model (2.1) can be rewritten in its usual form:

$$\mathbf{y}_i = \boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i$$

$$\boldsymbol{\eta}_i = \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\varsigma}_i,$$

(2.2)

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ \beta_{2,1} & 0 & \cdots & 0 & 0 \\ . & . & & . & \\ . & . & & . & \\ 0 & 0 & & \beta_{T,T-1} & 0 \end{bmatrix}$$

The distributional assumptions for (2.2) resemble those for the standard factor model (1.1): $\varsigma_i \sim \aleph(\mathbf{0}, \mathbf{\Psi})$, $\varepsilon_i \sim \aleph(\mathbf{0}, \Theta)$, where $\mathbf{\Psi}$ and Θ are taken to be diagonal (uncorrelated innovations and measurement errors) and cov[$\boldsymbol{\eta}_i$, $\boldsymbol{\varepsilon}_i$] = cov[$\boldsymbol{\eta}_i$, ς_i] = cov[ε_i, ς_i] = **0**. It then follows from (2.2) that cov[$\mathbf{y}_i$, $\mathbf{y}_i$] = $\mathbf{\Sigma}_y = \mathbf{\Phi} + \Theta$ = $(\mathbf{I}_T - \mathrm{B})^{-1}\mathbf{\Psi}(\mathbf{I}_T - \mathrm{B}')^{-1} + \Theta$, $\mathbf{I}_T$ being the (T,T)-dimensional identity matrix.

2.1.2 The standard 1-factor model as quasi-simplex model

To show that the standard factor model (1.1) is nested under the quasi-simplex model (2.2), it has to be specified which parameters in (2.2) have to be restricted in order to obtain (1.1). To ease the discussion, attention will first be restricted to the standard 1-factor model.

A salient difference between the 1-factor model and the quasi-simplex model concerns the number of latent η-variables. For T longitudinal measurement occasions the 1-factor model has one latent η-variable, whereas the quasi-simplex model has T η-variables, one for each measurement occasion. Hence it is obvious that the quasi-simplex model only can yield a 1-factor model if its T η-variables are reduced to single η-variable. This is accomplished by restricting all innovation variances along the diagonal of $\mathbf{\Psi}$ to zero, expect at time t = 1. Because $\varsigma_i \sim \aleph(\mathbf{0}, \mathbf{\Psi})$ in (2.2), these restrictions imply that the second equation in (2.1) reduces to: $\eta_i(t) = \beta_{t,t-1}\eta_i(t-1)$, t=2,...,T. Moreover, $\mathbf{\Phi}$ = $(\mathbf{I}_T - \mathrm{B})^{-1}\mathbf{\Psi}(\mathbf{I}_T - \mathrm{B}')^{-1}$ in (2,2) is a (T,T)-dimensional covariance matrix of rank 1, which implies that the initial T η-variables in (2,2) reduce to a single η-variable.

To illustrate, consider a quasi-simplex model where T = 4. Let $\mathbf{\Psi}$ = diag[φ_{11}, 0, 0, 0], where diag[.] specifies the diagonal of a covariance matrix in which all off-diagonal elements are zero. Then $\mathbf{\Phi}$ is the (4,4)-dimensional covariance matrix, of which the (i,j)-th element ϕ_{ij}, i=1,...,4, j=i,...,4, is given by:

$\phi_{11} = \varphi_{11}$

$\phi_{21} = \beta_{21}\varphi_{11}$ $\phi_{22} = (\beta_{21})^2 \varphi_{11}$

$\phi_{31} = \beta_{32}\beta_{21}\varphi_{11}$ $\phi_{32} = \beta_{32}(\beta_{21})^2\varphi_{11}$ $\phi_{33} = (\beta_{32}\beta_{21})^2\varphi_{11}$

$\phi_{41} = \beta_{43}\beta_{32}\beta_{21}\varphi_{11}$ $\phi_{42} = \beta_{43}\beta_{32}(\beta_{21})^2\varphi_{11}$ $\phi_{43} = \beta_{43}(\beta_{32}\beta_{21})^2\varphi_{11}$

$$\phi_{44} = (\beta_{43}\beta_{32}\beta_{21})^2\varphi_{11}$$

$\mathbf{\Phi}$ has 3 zero eigenvalues, while the single nonzero eigenvalue χ equals the sum of the diagonal elements of $\mathbf{\Phi}$: $\chi = \varphi_{11}(1 + (\beta_{21})^2 + (\beta_{32}\beta_{21})^2 + (\beta_{43}\beta_{32}\beta_{21})^2)$. This implies that $\mathbf{\Phi}$ has rank 1. The eigenvector $\mathbf{e}_\chi$ associated with the single nonzero eigenvalue χ is:

$$\mathbf{e}_\chi = (\surd\chi)^{-1}[1, \beta_{21}, \beta_{32}\beta_{21}, \beta_{43}\beta_{32}\beta_{21}]'$$

Hence if it is assumed that $\mathbf{\Psi} = \text{diag}[\varphi_{11}, 0, 0, 0]$ in the quasi-simplex, then it reduces to a 1-factor model in which the variance of the univariate factor equals $\phi_{11} = \varphi_{11}$ and the vector of factor loadings equals $\boldsymbol{\lambda} = \surd\chi\mathbf{e}_\chi$. Consequently, the 1-factor model is nested under the quasi-simplex model.

To give an arbitrary numerical illustration, take $\beta_{21} = 1$, $\beta_{32} = 2$, $\beta_{43} = .5$, $\varphi_{11} = 1$, and $\Theta = \text{diag}[1, 2, 3, 4]$. Then the true covariance matrix associated with this simplex model is:

$$\begin{array}{c|cccc|}
 & y(1) & y(2) & y(3) & y(4) \\
y(1) & 2 & & & \\
y(2) & 1 & 3 & & \\
y(3) & 2 & 2 & 7 & \\
y(4) & 1 & 1 & 2 & 5
\end{array}$$

The 1-factor model yields a perfect fit to this covariance matrix. If the univariate factor is scaled by fixing λ_{11} at $\lambda_{11} = 1$, the following parameter estimates are obtained:

$$\boldsymbol{\lambda} = [1, 1, 2, 1]'$$

$$\phi_{11} = 1$$

$$\Theta = \text{diag}[1, 2, 3, 4]$$

The parameter estimates for $\boldsymbol{\lambda}$ and ϕ_{11} conform to, respectively, $\surd\chi\mathbf{e}_\chi$ and ϕ_{11}, after substitution of the quasi-simplex parameter values for $\beta_{t,t-1}$, t=2,...,T, and φ_{11} used in generating the covariance matrix.

The demonstration given above shows that leaving out the genuine innovations in a quasi-simplex model, i.e., constraining the innovation variances $\varphi_{t,t}$ at $\varphi_{t,t} = 0$ for t=2,...,T, yields a standard 1-factor model. The reverse demonstration, namely that each 1-factor model is an instance of the quasi-simplex model in which $\varphi_{t,t} = 0$ for t=2,...,T, proceeds along the same lines and will not be spelled out here. For given dimension p of $\mathbf{y}_i$ the number of free parameters in the 1-factor model equals the number of free parameters in the quasi-simplex model in which $\varphi_{t,t} = 0$ for t=2,...,T. For instance, in the numerical example p = T = 4, and the number of free parameters in both the 1-factor model and the constrained quasi-simplex model is 8. This implies that the degrees of freedom of the likelihood ratio test of each model is also equal (namely 2 in the numerical example).

One immediate consequence of the finding that the 1-factor model is nested under the quasi-simplex model concerns the way in which comparisons between these two models can be carried out. A model which is nested under another model constitutes a constrained version of the model under which it is nested. It therefore makes sense to denote the nested model as a constrained submodel of the model under which it is nested. Hence such a constrained submodel can be compared with the unconstrained model, or less severely constrained submodels, by means of the difference between the likelihood ratio of the unconstrained model (or a less severely constrained submodel) and the likelihood ratio of the constrained model under scrutination. It is not necessary to use alternative criteria for model comparison or model selection, such as Akaike's information criterion (Akaike, 1987).

It is noted that the nesting of the 1-factor model under the quasi-simplex model can also be conceived of as a purely formal affair. What I mean by this is that the equivalence between the 1-factor model and the constrained quasi-simplex model can be regarded as an algebraic mapping that does not depend upon the interpretations of both models involved. From this formal point of view the interpretation of the quasi-simplex model as a longitudinal factor model is immaterial to the equivalence between a constrained version of it with the 1-factor model. It then follows that any 1-factor model, irrespective of the content of the observations making up $\mathbf{y}_i$, can be rewritten as a constrained quasi-simplex model. Of course this may imply that some of the parameters in **B** in (2.2) are negative, but that would be immaterial to the validity of the formal equivalence concerned (see Jöreskog, 1970, for a discussion of negative parameters in **B** in the context of longitudinal data).

2.1.3 The orthogonal q-factor model as a quasi-simplex model

The extension of the approach given in the previous section to prove the equivalence between standard q-factor models, q > 1, and constrained quasi-simplex models would appear to be straightforward. As will be shown in the present section, this is indeed the case in that a standard q-factor model turns out to be a constrained latent q-variate simplex model. But I want to point out from the outset that this straightforward extension is not the only one which is possible. There exists another, more subtle and rigorous equivalence between standard q-factor models and constrained univariate quasi-simplex models. The latter equivalence can only be proven after some technical results from time series analysis have been introduced, and hence its presentation has to wait until a later section.

To make the generalization of the 1-factor model to q-factor models, q > 1, it suffices to explain the generalization to 2-factor models. Extensions to q-factor models, q > 2, then follow by straightforward induction. Hence in the present section the discussion is limited to 2-factor models. First the orthogonal 2-factor model is considered, followed by a discussion of the oblique 2-factor model.

The general (explorative) orthogonal 2-factor model for the p-variate manifest variable $\mathbf{y}_i$ can be represented as:

$$\mathbf{y}_i = \Lambda\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i,\ \boldsymbol{\eta}_i = [\eta_{1i}, \eta_{2i}]'$$

(2.3)

$$\Sigma_y = \Lambda\Lambda' + \Theta,\ \Theta = \mathrm{diag}[\theta_{11}, \ldots, \theta_{pp}]$$

where Λ denotes a (p,2)-dimensional matrix of factor loadings. The univariate latent factors η_{1i} and η_{2i} have been scaled by constraining their variances at $\phi_{11} = \phi_{22} = 1$.

Because (2.3) is a representation of a so-called explorative 2-factor model, it is presumed that there are no a prori constraints on Λ. That is, Λ is considered to be a (p,2)-dimensional matrix which consists of 2p free parameters (factor loadings). It is well-known that this leaves the explorative model (2.3) only identified up to orthonormal rotation. Post-multiplication of Λ by a (2,2)-dimensional orthonormal matrix **R** transforms a solution of (2.3) into an equivalent solution. The orthonormal rotation matrix **R** can be represented as:

$$\mathbf{R}(\alpha) = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix}$$

where α is some angle in the range $0 \le \alpha \le 2\pi$. Because α can take on infinitely many values, there are infinitely many equivalent solutions of (2.3). The minimum constraint to arrive at a uniquely identified solution of (2.3) is to fix one element of Λ. For instance, this is accomplished by fixing λ_{12} at $\lambda_{12} = 0$. The constraint that $\lambda_{12} = 0$ can always be realized by a unique choice of α. Let λ_{11} and λ_{12} denote the loadings associated with y_{1i} in Λ. What we want is a unique rotation (within the range $0 \le \alpha \le 2\pi$) that transforms λ_{12} into $*\lambda_{12} = -\lambda_{11}\sin(\alpha) + \lambda_{12}\cos(\alpha) = 0$. This implies that the angle realizing this constraint equals $\alpha = \tan^{-1}(\lambda_{12} / \lambda_{11})$. Hence the minimum identifiability constraint that $\lambda_{12} = 0$ can be interpreted as the choice of a particular rotation orientation of the solution of (2.3).

The uniquely identifiable explorative 2-factor model will be chosen to be the one in which $\lambda_{12} = 0$. It is a simple exercise to show that this factor model is nested under a generalization of (2.2) involving two independent latent simplex processes. Let $y_i(t)$ denote the univariate observation at occasion t for subject i. Let $\mathbf{y}_i = [y_i(1),...,y_i(T)]'$ be the associated vector of longitudinal observations of subject i across the T fixed measurement occasions and $\boldsymbol{\varepsilon}_i = [\varepsilon_i(1),...,\varepsilon_i(T)]'$ the vector of measurement errors. Let $\eta_{1i}(t)$, t=1,...,T, and $\eta_{2i}(t)$, t=2,...,T, denote two latent simplex processes (note that the second simplex process $\eta_{2i}(t)$ starts at t=2). Define the (2T-1)-dimensional vector $\boldsymbol{\eta}_i = [\eta_{1i}(1), ..., \eta_{1i}(T), \eta_{2i}(2), ..., \eta_{2i}(T)]'$ and the (2T-1)-dimensional vector $\boldsymbol{\varsigma}_i = [\varsigma_{1i}(1), ..., \varsigma_{1i}(T), \varsigma_{2i}(2), ..., \varsigma_{2i}(T)]'$. With these definitions the generalization of (2.2) is obtained as:

$$\mathbf{y}_i = \Lambda\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i$$

$$\boldsymbol{\eta}_i = \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\varsigma}_i$$

$$\Lambda = [\mathbf{I}_T, \Lambda_2]$$

(2.4)

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_{22} \end{bmatrix}$$

$$\mathbf{B}_{11} = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ \beta_{2,1} & 0 & \cdots & 0 & 0 \\ . & . & & . & \\ . & . & & . & \\ 0 & 0 & & \beta_{T,T-1} & 0 \end{bmatrix}$$

$$\mathbf{B}_{22} = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ \delta_{3,2} & 0 & \cdots & 0 & 0 \\ . & . & & . & \\ . & . & & . & \\ 0 & 0 & & \delta_{T,T-1} & 0 \end{bmatrix}$$

The (T, 2T-1)-dimensional matrix Λ is composed of two parts: the (T,T)-dimensional identity matrix $\mathbf{I}_T$ associated with $\eta_{1i}(t)$, t=1,...,T, and the (T, T-1)-dimensional matrix Λ_2 associated with $\eta_{2i}(t)$, t=2,...,T. The first row of Λ_2 consists of T-1 zeroes, while its remaining part is $\mathbf{I}_{T-1}$. The (2T-1, 2T-1)-dimensional matrix **B** has two nonzero components: the (T,T)-dimensional submatrix $\mathbf{B}_{11}$ associated with $\eta_{1i}(t)$ together with the (T-1, T-1)-dimensional submatrix $\mathbf{B}_{22}$ associated with $\eta_{2i}(t)$. The distributional assumptions for (2.4) are similar to those for (2.2).

The bivariate latent simplex process in (2.4) can be written as:

$$\eta_{1i}(t) = \beta_{t,t-1}\eta_{1i}(t-1) + \varsigma_{1i}(t), \ t=2,...,T; \ \eta_{1i}(1) = \varsigma_{1i}(1)$$

$$\eta_{2i}(t) = \delta_{t,t-1}\eta_{2i}(t-1) + \varsigma_{2i}(t), \ t=3,...,T; \ \eta_{2i}(2) = \varsigma_{2i}(2)$$

where it is understood that $\mathrm{cov}[\varsigma_{1i}(t), \varsigma_{2i}(t')] = 0$ for all times t and t' at which this covariance is defined. Consequently,
$\mathrm{cov}[\eta_{1i}(t), \eta_{2i}(t')] = 0$ for all times t and t' at which this covariance is defined. In addition, it follows from (2.4) that:

$$y_i(1) = \eta_{1i}(1) + \varepsilon_i(1)$$

$$y_i(t) = \eta_{1i}(t) + \eta_{2i}(t) + \varepsilon_i(t), \ t=2,...,T$$

It now only requires a straightforward generalization of the discussion given in the previous section to show that the following constrained instance of (2.4) reduces to the explorative orthogonal 2-factor model:

$$\begin{aligned}
&\eta_{1i}(t) = \beta_{t,t-1}\eta_{1i}(t-1),\ t=2,...,T;\ \eta_{1i}(1) = \varsigma_{1i}(1)\\
&\eta_{2i}(t) = \delta_{t,t-1}\eta_{2i}(t-1),\ t=3,...,T;\ \eta_{2i}(2) = \varsigma_{2i}(2)\\
&y_i(1) = \eta_{1i}(1) + \varepsilon_i(1)\\
&y_i(t) = \eta_{1i}(t) + \eta_{2i}(t) + \varepsilon_i(t),\ t=2,...,T
\end{aligned} \tag{2.5}$$

Note that the latent simplex process $\eta_{1i}(t)$ has no innovations after the first time point t=1 at which it is defined, while the latent simplex process $\eta_{2i}(t)$ has no innovations after the first time point t=2 at which it is defined. It follows that the (2T-1, 2T-1)-dimensional covariance matrix $\mathbf{\Phi} = (\mathbf{I}_{2T-1} - \mathbf{B})^{-1}\mathbf{\Psi}(\mathbf{I}_{2T-1} - \mathbf{B}')^{-1}$ of $\boldsymbol{\eta}_i$ associated with the constrained latent bivariate simplex model (2.5) has rank 2.

To illustrate the nesting of the orthogonal 2-factor model (2.3) under the latent 2-variate simplex model (2.5), let the dimension of the manifest vector $\mathbf{y}_i$ be T = p = 5. In addition, for the first univariate simplex component process $\eta_{1i}(t)$ the following parameter values are chosen: $\beta_{t,t-1} = 0.9$, t=2,3,4,5, and var[$\varsigma_{1i}(1)$] = 1. For the second univariate simplex component $\eta_{2i}(t)$ the following parameter values are chosen: $\delta_{3,2} = 1$, $\delta_{4,3} = 2$, $\delta_{5,4} = .5$, and var[$\varsigma_{2i}(2)$] =1. Note that these parameter values for $\eta_{2i}(t)$ are the same as for the numerical illustration given in the previous section. Finally, the covariance matrix of the measurement errors is $\mathbf{\Theta}$ = diag[5, 4, 3, 2, 1]. The true covariance matrix associated with these parameter values for (2.5) is:

$$\begin{array}{c|ccccc|}
 & y(1) & y(2) & y(3) & y(4) & y(5) \\
y(1) & 6.00 & & & & \\
y(2) & 0.90 & 5.81 & & & \\
y(3) & 0.81 & 1.73 & 4.66 & & \\
y(4) & 0.73 & 2.66 & 2.59 & 6.53 & \\
y(5) & 0.66 & 1.59 & 1.53 & 2.48 & 2.43
\end{array}$$

The 2-factor model (2.3) yields a perfect fit to this covariance matrix. The obtained parameter estimates are:

$$\mathbf{\Lambda} = \begin{bmatrix} 1.00 & 0.00 \\ 0.90 & 1.00 \\ 0.81 & 1.00 \\ 0.73 & 2.00 \\ 0.66 & 1.00 \end{bmatrix}$$

Θ = diag[5, 4, 3, 2, 1]

The first column of Λ in the factor solution has the pattern [1, β_{21}, $\beta_{32}\beta_{21}$, $\beta_{43}\beta_{32}\beta_{21}$, $\beta_{54}\beta_{43}\beta_{32}\beta_{21}$]', for the values $\beta_{t,t-1}$ = 0.9, t=2,3,4,5. Hence the first univariate latent factor η_{1i} has loadings that equal the scaled eigenvector associated with the covariance matrix of the first constrained univariate latent simplex $\eta_{1i}(t)$. This pattern was analytically derived in the previous section. Similarly, the nonzero entries in the second column of Λ in the factor solution has the pattern [1, δ_{32}, $\delta_{43}\delta_{32}$, $\delta_{54}\delta_{43}\delta_{32}$]', for the values $\delta_{3,2}$ = 1, $\delta_{4,3}$ = 2, $\delta_{5,4}$ = .5. Hence the second univariate latent factor η_{2i} has loadings that equal the scaled eigenvector associated with the covariance matrix of the second constrained univariate latent simplex $\eta_{2i}(t)$. Note again that these factor loadings in the second column of Λ equal those obtained in the numerical illustration given in the previous section.

The argumentation and numerical illustration given above bear a close resemblance to the argumentation and numerical illustration given in the previous section. Apart from details related to minimum identifiability constraints, the argumentation given in the present section decomposes into a twofold proof, one for each univariate latent simplex component process, where each part of the proof repeats the steps in the proof given for the nesting of the 1-factor model under the latent univariate simplex as given in the previous section. This implies that analogous remarks can be made to the ones given at the end of the previous section concerning the implications of this nesting relation for model selection and with respect to the interpretation of this nesting relation as a formal relationship not tied up with longitudinal data. And what holds for the proof of the nesting of the explorative orthogonal 2-factor model under the latent 2-variate simplex model can be generalized straightforwardly to proofs of the nesting of explorative orthogonal q-factor models, q > 2, under the latent q-variate simplex model. I leave the specification and elaboration of these implications and generalizations to the reader, and instead proceed with a discussion of the relationship between the oblique q-factor model and the latent q-variate simplex model.

2.1.4 The exploratory oblique q-factor model as a quasi-simplex model

In discussing the nesting of oblique q-factor models under latent q-variate simplex models, it is convenient to distinguish between exploratory oblique q-factor models on the one hand, and confirmative oblique q-factor models on the other hand. It is not my intention to suggest that this distinction is fundamental; it is merely used in order to ease the argumentation. In addition, for the same reasons why only the exploratory orthogonal 2-factor model has been considered, attention will be restricted in what follows to a consideration of exploratory and confirmatory oblique 2-factor models.

The exploratory oblique 2-factor model is given by:

$\mathbf{y}_i = \Lambda\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i$, $\boldsymbol{\eta}_i = [\eta_{1i}, \eta_{2i}]'$

(2.6)

$$\Sigma_y = \Lambda\Phi\Lambda' + \Theta, \Phi = \text{cov}[\eta_i, \eta_i'], \Theta = \text{diag}[\theta_{11}, \ldots, \theta_{pp}]$$

where Λ is again a (p,2)-dimensional matrix of factor loadings. The distributional assumptions are the same as for the exploratory orthogonal 2-factor model (2.3). Because (2.6) is considered to be an exploratory model, there are no a priori constraints on Λ. Hence, like for the exploratory orthogonal model (2.3), it only remains to introduce minimal constraints on Λ to guarantee unique identifiability. In contrast to exploratory orthogonal factor models, however, there does not appear to be a simple rule to arrive at uniquely identifiable exploratory oblique factor models. Of course it is always possible to test for identfiability by means of alpha-numerical software (e.g., Maple, cf. Heck, 1996), or by means of simplifications of such alpha-numerical checks (Bekker, Merckens, & Wansbeek, 1994). But such a test, involving a check of the rank of the first-order derivative of any likelihood-like function (cf. Wooldridge, 1994) with respect to the free parameters, does not by itself yield the minimum number of constraints to arrive at unique identifiability in a direct way. It is required to carry out a search over all plausible alternatives to detect what may be the minimum number of constraints, where each alternative is subjected to the alpha-numerical test.

With respect to the unique identifiability of exploratory oblique factor models a different, more pragmatic, approach will followed. This approach is based on Jöreskog (1969) and consists in the transformation of an exploratory oblique q-factor models to an equivalent exploratory orthogonal q-factor model. Such a transformation is always well-defined. In particular, the oblique model (2.6) is transformed to the oblique model (2.3). This can be accomplished by considering the so-called spectral decomposition of the covariance matrix $\mathbf{\Phi}$ of η_i: $\mathbf{\Phi} = \mathbf{EDE}'$, where the columns of $\mathbf{E}$ are the eigenvectors of $\mathbf{\Phi}$ and where $\mathbf{D}$ is a diagonal matrix containing the eigenvalues of $\mathbf{\Phi}$ along the diagonal. Because $\mathbf{\Phi}$ is at least nonnegative-definite, the eigenvalues of $\mathbf{\Phi}$ are nonnegative. Hence the square roots of the diagonal elements of $\mathbf{D}$ are well defined and $\mathbf{D}$ can be written as $\mathbf{D} = \mathbf{D}^{1/2}\mathbf{D}^{1/2}$, where $\mathbf{D}^{1/2}$ is a diagonal matrix containing the square roots of the eigenalues of $\mathbf{\Phi}$ along the diagonal. Now define $\Lambda^* = \mathbf{\Lambda ED}^{1/2}$. Then in terms of the transformed matrix of factor loadings Λ^*, (2.6) transforms into the equivalent model: $\mathbf{y}_i = \mathbf{\Lambda}^*\eta_i + \varepsilon_i$, $\Sigma_y = \mathbf{\Lambda}^*\mathbf{\Lambda}^{*\prime} + \mathbf{\Theta}$. The latter model is an instance of (2.3), the exploratory orthogonal 2-factor model. Because it has already been shown that (2.3) is nested under the latent 2-variate simplex model (2.4), and because (2.6) has been transformed to (2.3), the nesting of (2.6) under (2.4) follows.

The line of argumentation showing that the explanatory oblique 2-factor model is nested under the latent bivariate simplex model can be detailed somewhat further by using the Choleski decomposition instead of the spectral decomposition of $\mathbf{\Phi}$. The Choleski decomposition of the covariance matrix $\mathbf{\Phi}$ in (2.6) is defined as $\mathbf{\Phi} = \mathbf{LU}$, where $\mathbf{L}$ is a lower-triangular matrix and $\mathbf{U} = \mathbf{L}'$. For instance, if $\mathbf{\Phi}$ is a (2,2)-dimensional correlation matrix, $\mathbf{\Phi}$ and its associated Choleski component $\mathbf{L}$ are:

(2.7) $$\mathbf{\Phi} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \qquad \mathbf{L} = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{(1-\rho^2)} \end{bmatrix}$$

Suppose that the covariance matrix of $\boldsymbol{\eta}_i$ in (2.6) is given by the correlation matrix $\mathbf{\Phi}$ specified in (2.7) Define $\mathbf{\Lambda}^* = \mathbf{\Lambda L}$, where $\mathbf{L}$ is the lower-triangular Choleski component given in (2.7). Then the oblique 2-factor model (2.6) transforms into the equivalent model: $\mathbf{y}_i = \mathbf{\Lambda}^*\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i$, $\mathbf{\Sigma}_y = \mathbf{\Lambda}^*\mathbf{\Lambda}^{*\prime} + \mathbf{\Theta}$, which is again an instance of the orthogonal 2-factor model (2.3). Using this transformation of (2.6) to (2.3), one can further specify the relationships between the factor models and latent simplex models, as will be shown in the following example.

To illustrate the discussion in this section, the covariance matrix $\mathbf{\Sigma}_y$ associated with a restricted bivariate latent simplex with correlated innovations is determined. This covariance matrix is subjected to an orthogonal 2-factor analysis, yielding a perfect fit. It is shown how the factor loadings in this orthogonal factor model are related to the parameters in the correlated bivariate latent simplex. In the final step it is shown that an orthogonal bivariate latent simplex also yields a perfect fit to $\mathbf{\Sigma}_y$, and it is shown how the parameters in this orthogonal latent simplex model are related to the parameters in the correlated latent simplex model used to generate $\mathbf{\Sigma}_y$.

The structural equations defining the bivariate latent simplex model with correlated innovations are given by (2.5) in the previous section, and are repeated below:

$$\eta_{1i}(t) = \beta_{t,t-1}\eta_{1i}(t-1),\ t=2,...,T;\ \eta_{1i}(1) = \varsigma_{1i}(1)$$

$$\eta_{2i}(t) = \delta_{t,t-1}\eta_{2i}(t-1),\ t=3,...,T;\ \eta_{2i}(2) = \varsigma_{2i}(2)$$

(2.5)

$$y_i(1) = \eta_{1i}(1) + \varepsilon_i(1)$$

$$y_i(t) = \eta_{1i}(t) + \eta_{2i}(t) + \varepsilon_i(t),\ t=2,...,T$$

Also the parameter values are the same as in the example described in the previous section: the dimension of the manifest vector $\mathbf{y}_i$ is T = p = 5 and for the first univariate simplex component process $\eta_{1i}(t)$ the following parameter values are chosen: $\beta_{t,t-1}$ = 0.9, t=2,3,4,5, and var[$\varsigma_{1i}(1)$] = 1. For the second univariate simplex component $\eta_{2i}(t)$ the following parameter values are chosen: $\delta_{3,2} = 1$, $\delta_{4,3} = 2$, $\delta_{5,4} = .5$, and var[$\varsigma_{2i}(2)$] =1. The only difference with the example in the previous section is that now cov[$\varsigma_{1i}(1)$, $\varsigma_{2i}(2)$] = 0.7. Like in the example in the previous section, the covariance matrix of the measurement errors is $\mathbf{\Theta}$ = diag[5, 4, 3, 2, 1]. The true covariance matrix associated with these parameter values for the restricted latent simplex with correlated innovations is:

$$
\begin{array}{c|ccccc}
 & y(1) & y(2) & y(3) & y(4) & y(5) \\
\hline
y(1) & 6.00 & & & & \\
y(2) & 1.60 & 7.07 & & & \\
y(3) & 1.51 & 2.93 & 5.79 & & \\
y(4) & 2.13 & 4.43 & 4.23 & 8.57 & \\
y(5) & 1.36 & 2.68 & 2.56 & 3.91 & 3.35
\end{array}
$$

The orthogonal 2-factor model (2.3) yields a perfect fit to this covariance matrix. The obtained parameter estimates are:

$$
\boldsymbol{\Lambda}* = \begin{bmatrix} 1.00 & 0.00 \\ 1.60 & 0.71 \\ 1.51 & 0.71 \\ 2.13 & 1.43 \\ 1.36 & 0.71 \end{bmatrix}
$$

$\boldsymbol{\Theta}$ = diag[5, 4, 3, 2, 1]

Notice that the matrix of factor loadings is denoted by $\boldsymbol{\Lambda}*$. The reason is that it is the transformed matrix of factor loadings $\boldsymbol{\Lambda}* = \boldsymbol{\Lambda}\mathbf{L}$, where $\mathbf{L}$ is the Choleski component given by (2.7). If the actual value of ρ, $\rho = 0.7$, used in generating $\boldsymbol{\Sigma}_y$ is substituted in $\mathbf{L}$, then it follows that its inverse $\mathbf{L}^{-1}$ is given by:

$$
\mathbf{L}^{-1} = \begin{bmatrix} 1 & 0 \\ -.98 & 1.40 \end{bmatrix}
$$

Postmultiplication of $\boldsymbol{\Lambda}* = \boldsymbol{\Lambda}\mathbf{L}$ by $\mathbf{L}^{-1}$ yields $\boldsymbol{\Lambda}$, with numerical values:

$$
\boldsymbol{\Lambda} = \begin{bmatrix} 1.00 & 0.00 \\ 0.90 & 1.00 \\ 0.81 & 1.00 \\ 0.73 & 2.00 \\ 0.66 & 1.00 \end{bmatrix}
$$

The numerical values of $\boldsymbol{\Lambda}$ are exactly the same as the factor loadings obtained in the example in the previous section.

Let us pause for a moment and consider what has been established. A covariance matrix has been generated according to a restricted bivariate latent simplex with correlated innovations, where this correlation is $\rho = \text{cor}[\varsigma_{1i}(1), \varsigma_{2i}(2)] = 0.7$. Denote this model by RLSC and denote the orthogonal 2-factor model which is equivalent to the RLSC by OFMC. All remaining parameters in the RLSC have the same numerical values as the latent simplex model used in the previous section to illustrate the nesting of the orthogonal 2-factor model under the bivariate latent simplex with independent innovations. Denote the latter model by RLSI and denote the orthogonal 2-factor model which is equivalent to the RLSI by OFMI. RSLC and RLSI only differ in the correlation between the innovations ($\rho = 0.7$ and $\rho = 0.0$, respectively). The matrix of factor loadings $\mathbf{\Lambda}^*$ in the OFMC turns out to have the form $\mathbf{\Lambda}^* = \mathbf{\Lambda L}$, where **L** is the Choleski component given by (2.7) for $\rho = 0.7$ and where $\mathbf{\Lambda}$ is the matrix of factor loadings in the OFMI. Hence OFMC and OFMI have a simple relationship: the matrix of factor loadings $\mathbf{\Lambda}^*$ in the OFMC equals the matrix of factor loadings $\mathbf{\Lambda}$ in the OFMI postmultiplied by the Choleski component **L** of the correlation matrix of the innovations in the RLSC. The relationships concerned can be written schematically as follows:

a) RLSC given by (2.5) with $\rho = \text{cor}[\varsigma_{1i}(1), \varsigma_{2i}(2)] = 0.7$

b) RLSC → $\mathbf{\Sigma}^{c}_{y}$: generate $\mathbf{\Sigma}^{c}_{y}$ by RLSC

c) RLSC ↔ OFMC: equivalence of RLSC and OFMC

d) Substitute $\rho = 0.7$ in (2.7) → **L**

e) RLSI given by (2.5) with $\rho = \text{cor}[\varsigma_{1i}(1), \varsigma_{2i}(2)] = 0.0$

f) RLSI → $\mathbf{\Sigma}^{i}_{y}$: generate $\mathbf{\Sigma}^{i}{}_{y}$ by RLSI

g) RLSI ↔ OFMI: equivalence of RLSI and OFMI (see previous section)

h) $\mathbf{\Lambda}^* = \mathbf{\Lambda L}$: $\mathbf{\Lambda}^*$ in OFMC equals $\mathbf{\Lambda}$ in OFMI postmultiplied by **L**

This schema shows that the nesting of the exploratory oblique 2-factor model under the latent bivariate simplex with correlated innovations obeys the same rules as the nesting of the orthogonal 2-factor model under the latent bivariate simplex with independent innovations as described in the previous section. The only new feature is the Choleski component **L**, whose operation is described by h). The schema also shows that both the restricted latent bivariate simplexes with correlated innovations (RLSC) and independent innovations (RLSI) have been linked up with orthogonal 2-factor models. According to c) the RLSC yields the OFMC and according to g) the RLSI yields the OFMI. This mapping of the RLSC on the OFMC is the basic feature of Jöreskog's (1969) approach to exploratory oblique factor analysis. It has the merit of sidestepping the identifiability issue with oblique factor models. Using this aspect of the Jöreskog (1969) approach precludes the need to specify the minimal identifiability constraints for the exploratory oblique 2-factor model (2.6). Instead, the well-known minimal identifiability constraints for orthogonal factor models suffice.

There remains one particular point that needs to be clarified. It concerns the distinction between the (2,2)-dimensional covariance matrix of the common factors in the factor model (2.6), $\text{cov}[\eta_i, \eta_i'] = \mathbf{\Phi}$, on the one hand, and the covariance matrix of the innovations in the restricted latent simplex with correlated innovations (RLSC) on the other hand. The RLSC is defined by (2.5) with the additional specification that $\text{cov}[\varsigma_{1i}(1), \varsigma_{2i}(2)] = 0.7$. The matrix equations underlying (2.5) are given by (2.4). In (2.4) the covariance matrix of the innovations is the (2T-1, 2T-1)-dimensional matrix

$\mathrm{cov}[\varsigma_i, \varsigma_i'] = \boldsymbol{\Psi}$. In the illustration given in this section, T = p = 5, consequently $\boldsymbol{\Psi}$ is a (9,9)-dimensional covariance matrix. The restrictions associated with the RLSC imply that $\boldsymbol{\Psi}$ has rank 2: all elements of $\boldsymbol{\Psi}$ are zero, save for $\psi_{11} = \mathrm{var}[\varsigma_1] = 1.0$, $\psi_{77} = \mathrm{var}[\varsigma_7] =1$, and $\psi_{71} = \mathrm{cov}[\varsigma_7, \varsigma_1] = 0.7$. Obviously $\boldsymbol{\Psi}$ differs from $\boldsymbol{\Phi}$ in that the respective dimensions are different. Yet, knowledge of $\boldsymbol{\Psi}$ enables one to reconstruct $\boldsymbol{\Phi}$, and vice versa. Hence the two covariance matrices are unambiguously related to each other. Notwithstanding this relationship, the Choleski decomposition (2.7) is defined with respect to $\boldsymbol{\Phi}$, not $\boldsymbol{\Psi}$. This raises the question whether it is possible to define a Choleski decomposition of $\boldsymbol{\Psi}$ and construct a schema based on the Choleski component thus obtained which specifies the same relationships as the schema given above.

In the remainder of this section this question and some additional implications of the schema given above will be further elaborated. The following discussion may be skipped on first reading, as it is only tangentially related to the main line of argument in this chapter.

It is of course possible to define a Choleski decomposition for the (2T-1, 2T-1)-dimensional innovations covariance matrix $\boldsymbol{\Psi}$ associated with the RLSC. Denote the lower-triangular Choleski component thus obtained by $\boldsymbol{\Delta}$. Premultiplication of the structural matrix equation $\boldsymbol{\eta}_i = \mathbf{B}\boldsymbol{\eta}_i + \varsigma_i$ in (2.4) by $\boldsymbol{\Delta}^{-1}$ then yields: $\boldsymbol{\Delta}^{-1}\boldsymbol{\eta}_i = \boldsymbol{\Delta}^{-1}\mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\Delta}^{-1}\varsigma_i$. Let $\mathbf{B}^* = \boldsymbol{\Delta}^{-1}\mathbf{B}\boldsymbol{\Delta}$, $\boldsymbol{\eta}_i^* = \boldsymbol{\Delta}^{-1}\boldsymbol{\eta}_i$, and $\varsigma_i^* = \boldsymbol{\Delta}^{-1}\varsigma_i$. Accordingly, the transformed structural equation becomes: $\boldsymbol{\eta}_i^* = \mathbf{B}^*\boldsymbol{\eta}_i^* + \varsigma_i^*$, where $\mathrm{cov}[\varsigma_i^*, \varsigma_i^{*\prime}]$ is diagonal. However, $\mathbf{B}^*$ does not have the required structure as specified by (2.4) in that its (T+2, 1)-th element is nonzero. Consequently, this straightforward approach to transformation in the latent simplex model does not work properly.

The approach underlying the schema given above consists in relegating transformation in restricted latent simplex models to transformation in equivalent factor models. That is, the (2T-1, 2T-1)-dimensional innovations covariance matrix $\boldsymbol{\Psi}$ of rank 2 is collapsed into the (2,2)-dimensional factor covariance matrix $\boldsymbol{\Phi}$. Choleski decomposition of $\boldsymbol{\Phi}$ then yields the desired transformation. To complete this approach, it is required to translate the results obtained in terms of factor models back to latent simplex models. This inverse relationship can be defined straightforwardly, as I will show now.

To start with, we go back to the discussion of the numerical illustration in the previous section. The illustration concerned the relationship between the RLSI (2.5) and the OFMI (2.3). More specifically, it was shown that the first column of $\boldsymbol{\Lambda}$ in the OFMI has the pattern $[1, \beta_{21}, \beta_{32}\beta_{21}, \beta_{43}\beta_{32}\beta_{21}, \beta_{54}\beta_{43}\beta_{32}\beta_{21}]'$, for the values $\beta_{t,t-1} = 0.9$, t=2,3,4,5, used in generating the manifest covariance matrix in that illustration. Hence the first univariate latent factor η_{1i} has loadings that equal the scaled eigenvector associated with the covariance matrix of the first constrained univariate latent simplex $\eta_{1i}(t)$ in (2.5). Similarly, the nonzero entries in the second column of $\boldsymbol{\Lambda}$ in the OFMI has the pattern $[1, \delta_{32}, \delta_{43}\delta_{32}, \delta_{54}\delta_{43}\delta_{32}]'$, for the values $\delta_{3,2} = 1$, $\delta_{4,3} = 2$, $\delta_{5,4} = .5$, used in generating the manifest covariance matrix in that illustration. Hence the second univariate latent factor η_{2i} has loadings that equal the scaled eigenvector associated with the covariance matrix of the second constrained univariate latent simplex $\eta_{2i}(t)$ in the RLSI. This specific patterning of the columns of the matrix of factor loadings in the OFMI enables the transformation from factor solutions back to latent simplexes.

Consider first the matrix of factor loadings in the illustration given in the present section concerning the relationship between the RLSC and the OFMC. This matrix is repeated below for convenience:

$$\Lambda* = \begin{bmatrix} 1.00 & 0.00 \\ 1.60 & 0.71 \\ 1.51 & 0.71 \\ 2.13 & 1.43 \\ 1.36 & 0.71 \end{bmatrix}$$

$\Lambda*$ is the matrix of factor loadings in an orthogonal 2-factor model where cov[η_i, η_i'] = $\mathbf{I}_2$. While the manifest covariance matrix $\Sigma^c{}_y$ to which it is fitted has been generated by the RLSC, the common factors in the OFMC are independent. Hence the pattern of each column in $\Lambda*$ corresponds to a restricted latent bivariate simplex with independent innovations. It is emphasized that the latter restricted latent bivariate simplex with independent innovations is <u>not</u> the RLSI used in the illustration in the previous section and mentioned in the schema given above. In contrast, it is the rotated RLSC, where the rotation has removed the correlation between the innovations. This rotated RLSC has not yet been considered in the discussion, nor has it been fitted to $\Sigma^c{}_y$. Its parameter values, however, can be recovered from $\Lambda*$.

Consider the first column in $\Lambda*$: [1.00, 1.60, 1.51, 2.13, 1.36]'. The first element has already the required value of 1.0. The second element equals β_{21} = 1.60. The third element equals $\beta_{32}\beta_{21}$ = 1.51, hence β_{32} = 1.51/1.60 = 0.94. Etc. In fact, all β-parameters can be recovered from the first column by the same algorithm:

$$\beta_{t,t-1} = \lambda_t / \lambda_{t-1}, \text{ t=2,3,4,5,}$$

where λ_t denotes the t-th element of the first column of $\Lambda*$. Accordingly, the first factor in the OFMI corresponds to a restricted latent simplex with parameters: β_{21} = 1.60, β_{32} = 0.94, β_{43} = 1.41, and β_{54} = 0.64. The variance of the innovation at t = 1 is var[ς_{1i}(1)] = 1.0. This completes the description of the first latent univariate simplex.

The second column in $\Lambda*$ is: [0.0, 0.71, 0.71, 1.43, 0.71]'. Application of the same algorithm as before to the nonzero elements (i.e., for t=3,4,5) in this second column yields: $\delta_{3,2}$ = 1, $\delta_{4,3}$ = 2, $\delta_{5,4}$ = .5. But now the first nonzero element of the second column does not have the required value of 1.0. This is accommodated by taking the variance of the innovation at t = 2 to be var[ς_{2i}(2)] = $(0.71)^2$. This completes the description of the second latent univariate simplex. Furthermore, cov[ς_{1i}(1), ς_{2i}(2)] = 0.0, implying that the complete latent bivariate simplex has independent innovations.

2.1.5 The confirmatory oblique q-factor model as a quasi-simplex model

In contrast to the preceding sections, where we had to deal with minimal identifiability issues associated with exploratory multifactor models, it is assumed in the present section that a confirmatory oblique q-factor model always is identifiable. I consider this identifiability criterion to be a necessary part of the definition of a confirmatory model. It then becomes a straightforward exercise to show that the confirmatory oblique q-factor model is nested under the latent q-variate simplex with correlated innovations. Again, for the same reasons as before, only the confirmatory oblique 2-factor model will be considered.

Using the same argumentation as in the previous sections, it can be shown that the confirmatory oblique 2-factor model given by:

$$\mathbf{y}_i = \Lambda\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i,\ \boldsymbol{\eta}_i = [\eta_{1i}, \eta_{2i}]'$$

$$\Sigma_y = \Lambda\Phi\Lambda' + \Theta,\ \Phi = \mathrm{cov}[\boldsymbol{\eta}_i, \boldsymbol{\eta}_i'],\ \Theta = \mathrm{diag}[\theta_{11}, \ldots, \theta_{pp}]$$

is nested under the latent bivariate simplex with correlated innovations. Let $I_k \subseteq \{1,...,p\}$ denote the set of indices of free factor loadings in the k-th column of Λ, k=1,2. In addition, let $i_k(1)$ be the first element of I_k and let $J_k = I_k - i_k(1)$, i.e., J_k is I_k without its first element $i_k(1)$. Then the confirmatory oblique 2-factor model is equivalent to the following restricted latent bivariate simplex with correlated innovations:

$$\eta_{1i}(t) = \beta_{t,t-1}\eta_{1i}(t-1),\ t \in J_1;\ \eta_{1i}(i_1(1)) = \varsigma_{1i}(i_1(1))$$

$$\eta_{2i}(t) = \delta_{t,t-1}\eta_{2i}(t-1),\ t \in J_2;\ \eta_{2i}(i_2(1)) = \varsigma_{2i}(i_2(1))$$

$$y_i(t) = L_1\eta_{1i}(t) + L_2\eta_{2i}(t) + \varepsilon_i(t),$$

where $L_k = 1$ if $t \in I_k$, k=1,2, and $L_k = 0$ otherwise. Moreover, $\mathrm{var}[\varsigma_{1i}(i_k(1))] = \mathrm{var}[\eta_{ki}]$, k=1,2, and $\mathrm{cov}[\varsigma_{1i}(i_1(1)), \varsigma_{2i}(i_2(1))] = \mathrm{cov}[\eta_{1i}, \eta_{2i}]$.

To illustrate, consider the confirmatory oblique 2-factor model given by:

$$\Lambda = \begin{bmatrix} 1.00 & 0.00 \\ 0.90 & 1.00 \\ 0.81 & 1.00 \\ 0.73 & 2.00 \\ 0.00 & 1.00 \end{bmatrix}$$

$$\Phi = \begin{bmatrix} 1.00 & \\ 0.70 & 1.00 \end{bmatrix}$$

$$\Theta = \mathrm{diag}[5, 4, 3, 2, 1]$$

In this particular model I_1 = {1,2,3,4}, the set of indices of free factor loadings in the first column of Λ; $i_1(1) = 1$, the first element of I_1; J_1 = {2,3,4}, which is I_1 minus $i_1(1)$. In addition, I_2 = {2,3,4,5}, the set of indices of free factor loadings in the second column of Λ, $i_2(1) = 2$, and J_2 = {3,4,5}. It then follows that this factor model is equivalent to the restricted latent bivariate simplex with correlated innovations in which $\text{var}[\varsigma_{1i}(i_1(1))] = \text{var}[\varsigma_{1i}(1)] = \text{var}[\eta_{1i}] = 1$, $\text{var}[\varsigma_{2i}(i_2(1))] = \text{var}[\varsigma_{2i}(2)] = \text{var}[\eta_{2i}] = 1$, and $\text{cov}[\varsigma_{1i}(i_1(1)), \varsigma_{2i}(i_2(1))] = \text{cov}[\varsigma_{1i}(1), \varsigma_{2i}(2)] = \text{cov}[\eta_{1i}, \eta_{2i}] = 0.7$. In addition, from the specifications of I_1, etc., and using the computational rules given in the previous section it follows that, for the component simplex process $\eta_{1i}(t)$, $\beta_{t,t\text{-}1} = 0.9$, t=2,3,4, and for the simplex component $\eta_{2i}(t)$, $\delta_{3,2} = 1$, $\delta_{4,3} = 2$, $\delta_{5,4} = .5$. Both models yield the same covariance matrix:

$$\begin{array}{c|ccccc|}
 & y(1) & y(2) & y(3) & y(4) & y(5) \\
y(1) & 6.00 & & & & \\
y(2) & 1.60 & 7.07 & & & \\
y(3) & 1.51 & 2.93 & 5.79 & & \\
y(4) & 2.13 & 4.43 & 4.23 & 8.57 & \\
y(5) & 0.70 & 1.63 & 1.57 & 2.51 & 2.00
\end{array}$$

The relationship between confirmatory oblique q-factor models and restricted latent q-simplex models turns out to be much more straightforward than the relationship between their respective exploratory analogues. This is due to the minimum identifiability constraints which have to be introduced for exploratory models, but are assumed to be already taken care of in confirmatory models.

2.1.6 Reflections about factor models for longitudinal data

In the foregoing sections it was shown that factor models are nested under latent simplex models. This completes the first step in the long argument to which this chapter is devoted. Before going on with the remaining steps in this argument, I would like to pause and present in this section some reflections about factor models for longitudinal data.

As was indicated before, the equivalence between factor models and restricted latent simplex models can be regarded as a formal equivalence that holds whether or not the data are longitudinal. In the present section, however, it will be understood that the data are longitudinal. Hence the focus is on factor models for longitudinal data. I will present some considerations implying that factor models may not be appropriate models for longitudinal data. First, however, we have to make some conceptual distinctions.

The simple quasi-simplex model defined by (2.1) is an instance of the longitudinal factor model of Jöreskog (1979). Let $\mathbf{y}_i(t)$ denote the p-variate vector of observations for subject i at time t; t=1,2,...,T. Then the longitudinal factor model is defined by:

$$\mathbf{y}_i(t) = \Lambda_t\boldsymbol{\eta}_i(t) + \boldsymbol{\varepsilon}_i(t), \ t=1,...,T; \ i=1,2,...$$

$$\boldsymbol{\eta}_i(t) = \mathbf{B}_{t,t-1}\boldsymbol{\eta}_i(t-1) + \boldsymbol{\zeta}_i(t), \ t=2,...,T$$

where Λ_t is a (p,q)-dimensional matrix of factor loadings at time t, $\boldsymbol{\eta}_i(t)$ is a q-variate latent factor at time t, $\boldsymbol{\varepsilon}_i(t)$ is p-variate measurement error at time t, $\mathbf{B}_{t,t-1}$ is the (q,q)-dimensional matrix of regression weights linking $\boldsymbol{\eta}_i(t)$ to $\boldsymbol{\eta}_i(t-1)$, and $\boldsymbol{\zeta}_i(t)$ is q-variate innovation at time t. The quasi-simplex model (2.1) is obtained by taking p = q = 1 in the longitudinal factor model. Hence the quasi-simplex model is the longitudinal 1-factor model for univariate time-dependent observations. Stated more succinctly, the quasi-simplex is a univariate longitudinal 1-factor model.

The factor models considered in the previous sections are of a different nature than the longitudinal factor model. These factor models are instances of Cattell's T-technique (e.g., Cattell, 1946). According to Wohlwill (1973, p. 269): "... T-technique is not, properly speaking, a multivariate technique at all, at least in the sense of providing information concerning covariation among different *response* variables. In compensation, it allows for the examination of temporal patterns, for a single response measure, but for a sample of individuals." (italics in the original text). I do not entirely agree with Wohlwill's characterization of T-technique (I consider T-technique to be a genuine multivariate technique which also can accommodate multiple response variables). But it nicely captures the essence of the technique. In what follows I will concentrate on factor models used in T-technique applied to univariate longitudinal data.

Wohlwill criticizes the use of T-technique in the following way: "The number of factors that would result is clearly very much determined by the spacing of the occasions ... In a developmental context, however, a different issue arises, relating to the effect of the dimension of temporal proximity or distance represented by the set of occasions. For such a set will inevitably form a simplex, in Guttman's sense, that is, the correlations will be maximal nearest the diagonal (that is, among adjoining occasions) and fall away systematically as the distance between the variables along the time (or any other) dimension increases." (Wohlwill, 1973, p.269). He then points out that: "The more basic point, however, ... is the fact that the factor-analytic model is fundamentally unsuited to data conforming to a simplex, because of the determination of the correlations by the single dimension of proximity (i.e., decreasing as an inverse function of the temporal interval separating them)." (Wohlwill, 1973, pp. 270-271).

I think that Wohlwill's criticism of T-technique is based on an interesting intuition, namely the possible unsuitability of factor models to accommodate the ordered time dimension, but his arguments do not appear to be definitive. Below I will present mathematical-statistical reasons underpinning aspects of Wohlwill's criticism. But first, a more heuristic argument is considered which would seem to corroborate Wohlwill's point of view.

The heuristic argument concerned is based on the recognition that factor models imply perfect predictability at the latent level. Take for instance the 1-factor model considered in section 2.1.2. It can be rewritten as a restricted quasi-simplex model in which only the initial condition at the latent level is random. More specifically, the 1-factor model is equivalent to the restricted latent univariate simplex given by:

$$y_i(t) = \eta_i(t) + \varepsilon_i(t),\ t=1,...,T$$

$$\eta_i(1) = \varsigma_i(1);\ \eta_i(t) = \beta_{t,t-1}\eta_i(t-1),\ t=2,...,T$$

Only $\eta_i(1) = \varsigma_i(1)$ is random; the relationships $\eta_i(t) = \beta_{t,t-1}\eta_i(t-1)$, t=2,...,T, merely involve consecutive affine transformations of $\eta_i(1)$. Consequently, the (T,T)-dimensional covariance matrix cov[$\boldsymbol{\eta}_i$, $\boldsymbol{\eta}_i$], where $\boldsymbol{\eta}_i = [\eta_i(1),..., \eta_i(T)]'$, has rank 1. Given this restricted quasi-simplex model, if $\eta_i(1) = \varsigma_i(1)$ is known for an arbitrary subject i, then $\eta_i(t)$ is known for all other times t at which the model is defined. Stated otherwise, the 1-factor model for longitudinal data implies that within each subject i there exists at the latent level complete stability and perfect predictability across all time points. Such a state of affairs could be obtained if a genuine trait variable is repeatedly measured within a limited time span {t=1,...,T}. But many empirical processes do not obey this strict invariance criterion and will show less than perfect predictability. For the latter processes there will be random innovations at later time points t > 1, leading to an increase in the rank of cov[$\boldsymbol{\eta}_i$, $\boldsymbol{\eta}_i$], i.e., the rank of cov[$\boldsymbol{\eta}_i$, $\boldsymbol{\eta}_i$] > 1. If this rank is r, $r \leq T$, then it would seem that r factors are required to describe cov[$\boldsymbol{\eta}_i$, $\boldsymbol{\eta}_i$]. This simple deduction holds if one could restrict attention to the latent level, that is, if measurement error $\varepsilon_i(t)$, t=1,...,T, is absent. The presence of measurement error, however, complicates the relationship between the rank of cov[$\boldsymbol{\eta}_i$, $\boldsymbol{\eta}_i$] in the quasi-simplex model and the required number of factors in equivalent factor models. But the general form of this relationship stays the same: if the rank of cov[$\boldsymbol{\eta}_i$, $\boldsymbol{\eta}_i$] in the quasi-simplex increases then the number of factors in equivalent factor models with additive measurement error also increases, though at a slower rate.

These considerations suggest that a T-technique 1-factor model for univariate longitudinal data requires perfect predictability at the latent level. The factor loadings in such a 1-factor model can be interpreted as a trend function: $y_i(t) = \lambda_t\eta_i + \varepsilon_i(t)$, t=1,...,T, where λ_t denotes the factor loading at time t. The ordered set of factor loadings {λ_t, t=1,...,T} can be depicted as a time-dependent curve or trend function. Given the factor model concerned, and given the factor score η_a for an arbitrary fixed subject a, the latent trajectory $\lambda_t\eta_a$, t=1,...,T is a deterministic function. In case there is less than perfect predictability at the latent level at a sufficient number of measurement occasions, additional factors are required in T-technique in order to describe the additional random sources causing this decreased predictability (where the latter random sources are represented by the innovations in the equivalent quasi-simplex model). The factor loadings associated with each additional T-technique factor can again be interpreted as another trend function. According to this scenario, a single quasi-simplex model defined at a an indefinitely increasing number of time points gives rise to an equivalent T-technique factor model with an indefinitely increasing number of factors. Consequently we have on the one hand a simple univariate latent simplex model and, on the other hand, an increasingly complex T-technique factor model, both of which yield equivalent fits to longitudinal data with less than perfect predictability at the latent level. In this scenario I do not consider the additional factors in the T-technique factor model, required to accommodate the lack of predictability in the latent univariate simplex process, to be spurious in a strict sense, as has sometimes been suggested in the literature (cf. the discussion of Cronbach's (1968) criticism of T-technique in Wohlwill, 1973, pp. 270-273). But it is obvious that there are important differences of some kind between both models. In order to try to explain what these differences amount to, I will have to make a short digression to the field of statistical time series analysis.

Consider a weakly stationary time series, for instance the simple autoregression described by: $x(t) = \rho x(t-1) + e(t)$, $t=0,\pm1,...$, where $|\rho| < 1$ and $e(t)$ is white noise. In chapter 1 it was shown that the autocorrelation function of x(t) is: $c_x(u) = \rho^{|u|}$, $u=0,\pm1,...$ Now consider x(t) at the time interval t=1,...,T. Then the (T,T)-dimensional covariance matrix of x(t), t=1,...,T, is given by $\mathbf{C}_x = \{c_x(i-j) , i,j=1,...,T\}$. That is, the (i,j)-th element of $\mathbf{C}_x$ is given by $c_x(i-j) = \rho^{|i-j|}$. $\mathbf{C}_x$ is called a Toeplitz matrix and is a prime example of a simplex in the sense of Guttman (1965). Of course, $\mathbf{C}_x$ is a covariance matrix of the same type as considered in T-technique. Letting T increase indefinitely, it can be shown that the eigenvectors of $\mathbf{C}_x$ become purely sinusoidal. A proof of this remarkable result can be found in Brillinger (1975, sections 3.7 and 4.7). Hence for large T a principal component analysis of a weakly stationary time series x(t), i.e., an eigenvalue decomposition of $\mathbf{C}_x$, converges to a spectral analysis of x(t), $t=0,\pm1,\dots$.

The implications of the asymptotic equivalence (in some appropriate sense) of the time series analogue of T-technique and spectral analysis of a weakly stationary time series are manifold. Some of these implications will be discussed in the next chapter, but most of them (e.g., the asymptotic independence of loadings at different frequencies in spectral analysis) cannot be considered further here as it would lead us too far away from the main themes of this book. Presently I will focus on one noteworthy implication, namely that the factor loadings in T-technique analysis of weakly stationary time series coverge (again, in some appropriate sense) to purely sinusoidal form and therefore become, in Wohlwill's words cited above, determined "... by the single dimension of proximity (i.e., decreasing as an inverse function of the temporal interval separating them)". The sinusoidal form concerned is $\sin[2\pi f(t-1) + \phi_f]$, t=1,2,...,T; where the frequency f = k/T, k=0,1,...,T-1, and hence is a function of the equidistant measurement occasions only (the restriction to equidistant measurement occasions is for convenience; cf., e.g., Papoulis, 1985, chapter 5, for irregularly spaced intervals, and Parzen, 1984, for random intervals). The phase ϕ_f is not important for our present concerns. Hence we have here one clear sense in which Wohwill's criticism of T-technique holds: the factor loadings in a T-technique analysis of weakly stationary time series become asymptotically a deterministic (oscillatory) function of frequency f and time t only.

According to the citation given earlier, Wohlwill considers a set of factor loadings of the form $\{\sin[2\pi f(t-1) + \phi_f], t=1,2,...,T; f = k/T, k=0,1,...,T-1\}$ to be uninformative because these loadings do not provide "... information concerning covariation among different *response* variables". On first reading this may appear to be a simple statement of fact, because x(t) is a univariate time series and hence realizations at different measurement occasions t_1 and t_2, yielding univariate realizations $x(t_1)$ and $x(t_2)$, would seem to involve repeated measurements of the same response variable (namely the response variable represented by x). But on closer scrutiny Wohwill's point of view may raise interesting questions. Suppose that the response variable x is the electrocortical potential (EEG) repeatedly registrated at a fixed location on the head. Applications of T-technique and spectral analysis to such a time series are standard in psychophysiology, and the resulting loadings associated with each T-technique factor are considered to describe the output of basic components of cortical information processing. Hence in this case such loadings are considered to be of prime importance and informative, in contrast to Wohlwill's conjecture. Or suppose instead that the response variable x is the total score on a psychological test and that at different times equivalent (e.g., parallel, congeneric, etc.) test forms are used (cf. Holtzman, 1963, for an application). Are the factor loadings obtained in an application of T-technique to such a time series still uninformative in the sense intended by Wohlwill? I doubt it, because that would seem to have serious implications for classical test theory. As a final comment on this issue, I would like to reiterate a point made earlier, namely that T-technique also applies to multivariate or multidimensional time series (Molenaar, 1981). Furthermore spectral analysis, i.e., the asymptotic form of T-technique, of multivariate or multidimensional time series is a standard tool in all kinds of scientific research. Again this state of affairs undermines Wohlwill's conjecture concerning the limitations of T-technique.

I conclude from these considerations that only part of Wohlwill's criticism of T-technique can be accepted, namely that the factor loadings obtained in a T-technique analysis of a weakly stationary time series depend upon the time intervals between repeated measurements. Asymptotically this dependence becomes the set of deterministic sinusoidal functions underlying spectral analysis. But even in the latter extreme case (T-technique applied to the (T,T)-dimensional Toeplitz matrix $\mathbf{C}_x$ associated with a weakly stationary time series x(t), where $T \rightarrow \infty$), the set of sinusoidal factor loadings thus obtained is informative about the autocorrelation function $c_x(u)$, u=0, ±1,..., making up $\mathbf{C}_x$. This autocorrelation function does not depend upon time t, but only on the lag u: $c_x(u) = \text{cor}[x(t), x(t+u)]$ for all t. Shifts along the time axis do not affect $c_x(u)$, and it is a fundamental result in abstract Fourier analysis that the set of sinusoidal functions constitutes the eigenfunctions of the operator defining such shifts (cf. Hannan, 1970, chapter 2, section 10). Although this is not the place to elaborate this beautiful result, it shows that even in the extreme case under consideration T-technique yields factor loadings that correctly characterize the special nature of a shift-invariant autocorrelation structure. Hence these factor loadings are not spurious or uninformative, despite their deterministic functional dependence upon the time intervals between repeated measurements. The picture becomes more complex in case measurement error is present and T is finite, but I am confident that this will not affect the informativeness of results of T-technique (although it will complicate the interpretation of factor loadings). The case of nonstationary time series is partly considered in Hannan (1970, chapter 2, section 6).

So T-technique cannot be dismissed on the grounds given by Wohlwill and it still remains to answer our initial question about the distinction between T-technique and the latent simplex model. Again a lead can be found in the literature about time series analysis. Jenkins & Watts (1968) characterize spectral analysis of weakly stationary time series as a multi-parametric approach, an approach that they consider to be akin to nonparametric approaches. This characterization makes explicit that standard spectral analysis is not based on a restrictive parametric model, but instead involves the application of a model with the maximum number of identifiable parameters, i.e., a saturated model. In contrast, the simple autoregression $x(t) = \rho x(t-$

1) + e(t) considered earlier is an instance of a restrictive parametric model. Following this lead, the difference between T-technique and latent simplex analysis could be characterized as the difference between a multi-parametric and a restrictive parametric approach.

Returning again to the setting of longitudinal models, what can be made of a characterization of T-technique factor analysis as a multi-parametric approach and latent simplex modeling as a restrictive parametric approach? Does this characterization indeed capture the differences between these two kinds of longitudinal models which were noted earlier in this section? I think it does, as can be shown by a simple argument. Consider again the general expression (2.2) for the covariance structure associated with the latent univariate simplex model: cov[$\mathbf{y}_i$, $\mathbf{y}_i$] = $(\mathbf{I}_T - \mathbf{B})^{-1}\Psi(\mathbf{I}_T - \mathbf{B}')^{-1} + \Theta$, where $\mathbf{y}_i = [y_i(1),...,y_i(T)]'$. Compare this with the general expression for the T-technique factor model for $\mathbf{y}_i$: cov[$\mathbf{y}_i$, $\mathbf{y}_i$] = $\Lambda\Phi\Lambda' + \Theta$. If we define: $\Lambda^* = (\mathbf{I}_T - \mathbf{B})^{-1}$, then the latent univariate simplex model can be represented as an orthogonal T-technique factor model: cov[$\mathbf{y}_i$, $\mathbf{y}_i$] = $\Lambda^*\Psi\Lambda^{*\prime} + \Theta$, where Ψ is diagonal.

To illustrate, let T = 4 in (2.2). Then the (4,4)-dimensional matrix $\Lambda^* = (\mathbf{I}_T - \mathbf{B})^{-1}$ of factor loadings in the T-technique model representing this latent univariate simplex is:

$$\Lambda^* = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \beta_{2,1} & 1 & 0 & 0 \\ \beta_{2,1}\beta_{3,2} & \beta_{3,2} & 1 & 0 \\ \beta_{2,1}\beta_{3,2}\beta_{4,3} & \beta_{3,2}\beta_{4,3} & \beta_{4,3} & 1 \end{bmatrix}$$

Clearly, Λ^* has a very restrictive parametric structure; its 16 elements depend upon only 3 free parameters $\beta_{2,1}$, $\beta_{3,2}$ and $\beta_{4,3}$. In contrast, the general T-technique model for T = 4 repeated measurements has no such restrictive parametric structure. Letting $\Phi = \mathbf{I}_4$, the (4,4)-dimensional matrix of factor loadings Λ is:

$$\Lambda = \begin{bmatrix} \lambda_{11} & 0 & 0 & 0 \\ \lambda_{21} & \lambda_{22} & 0 & 0 \\ \lambda_{31} & \lambda_{32} & \lambda_{33} & 0 \\ \lambda_{41} & \lambda_{42} & \lambda_{43} & \lambda_{44} \end{bmatrix}$$

The lower triangular pattern of Λ solely expresses the identifiability conditions for orthogonal explorative factor models considered earlier. Therefore Λ is seen to contain the maximum number of 10 identifiable free parameters. Comparison of Λ and Λ^* nicely illustrates the difference between the multi-parametric and the restrictive parametric approach.

In conclusion, it appears that the distinction between the longitudinal T-technique factor model and the longitudinal quasi-simplex model can be characterized as one between a multi-parametric approach and a restrictive parametric approach, respectively. Before returning to the main line of argument, however, I would like to address a possible misunderstanding to which this conclusion might give rise.

Consider again the example just given in which for T = 4 the longitudinal latent univariate simplex is rewritten as a restrictive instance of the orthogonal T-technique model. One might be tempted to interpret this restrictive T-technique analogue of the quasi-simplex model as indicative of a nesting relationship. That is, it might be concluded that the quasi-simplex model is nested under the T-technique factor model. Of course, this conclusion would be the exact inverse of the point of view defended in this chapter! Yet it should be immediately evident that the restrictive T-technique analogue of the quasi-simplex model is NOT nested under the general T-technique model. For T = 4 the general orthogonal T-technique factor model is cov[$\mathbf{y}_i$, $\mathbf{y}_i$] = $\mathbf{\Lambda\Lambda}'$ + $\mathbf{\Theta}$, where $\mathbf{\Lambda}$ is the (4,4)-dimensional matrix of factor loadings given above. This model has 14 free parameters, and therefore not identifiable (it beats the Ledermann bound; cf. Bekker, Merckens & Wansbeek, 1994, p. 84). Restricting the number of factors in this general T-technique model, in order to arrive at an identifiable model, never can lead to an analogue of the quasi-simplex model for T = 4 because the latter analogue factor model has 4 factors. Removing the measurement error variances in the general T-technique factor model, again to arrive at an identifiable model, also cannot lead to an analogue of the quasi-simplex model because the latter has nonzero measurement errors. There does not exist an identifiable T-technique factor model from which an analogue of the quasi-simplex model can be obtained by setting parameters at zero. Hence the quasi-simplex model is not nested under the factor model. The illustration given above should not be misinterpreted in this way; it was only given to characterize the difference between multi-parametric and restrictive parametric approaches. For the same reasons, the remark by Meredith & Horn (2001, p. 220) that their equation (7.13), in which a factor model analogue is given for the quasi-simplex model, "... clearly represents a factor analytic model" also should not be misinterpreted as indicating such a nesting relationship.

The only reason to stress the distinction between a relationship between models in which one model is represented as an instance of another type of model on the one hand, and on the other hand a relationship in which one model is nested under another model, is that the nesting relationship is much stronger (more restrictive) than the representation relationship. Nesting relationships constitute only a small subset of all representation relationships. This formal distinction is important (although to the best of my knowledge it has not been worked out in all its aspects) and its importance is not affected by the difficulties that beset the use of the nesting relationship in comparative model testing (cf. Titterington, Smith, & Makov, 1985).

2.1.7 The latent growth curve model as quasi-simplex model

Instead of considering hierarchical linear models in general, I will restrict attention in this section to latent growth curve models. Latent growth curve models are simple, yet interesting and popular instances of hierarchical linear models. The simplicity of latent growth curve models will ease the argumentation given in this section. I expect that the same line of argumentation, showing that latent growth curve models are nested under the quasi-simplex model, will carry over to hierarchical linear models in general, although at present this has not been worked out in detail. A second reason to focus on latent growth curve models is related to a much cited paper of Rogosa & Willett (1985) in which it is claimed that the latent growth curve model and the quasi-simplex model are quite different models of longitudinal data. The present section can be regarded as a refutation of this claim in that the latent growth curve model is shown to be a special instance of the latent simplex model. Another claim which is made in the paper of Rogosa & Willett (1985), namely that the quasi-simplex model yields inappropriately good fits to data generated according latent growth curve models, is shown to be incorrect by Mandys, Dolan, & Molenaar (1994).

The line of reasoning in this section will be simple: it is shown that latent growth curve models are special instances of confirmatory factor models. Because it has already been shown in the previous sections that factor models are nested under the latent simplex model, it is concluded that latent growth curve models are nested under quasi-simplex models (nesting is a transitive relationship). The main part of this section will be devoted to analyses of an empirical data set, illustrating the nesting relationships concerned.

For an up to date description of latent growth curve models, including their representation as factor models and references to the published literature, the reader is referred to Bijleveld, & van der Kamp (1998, chapter 5). The focus in the present section is on simple latent growth curve models for univariate longitudinal data $y_i(t)$, t=1,2,...,T; i=1,2,... (the main results thus obtained can easily be generalized to multivariate longitudinal data). Following Bijleveld & van der Kamp (1998), the latent growth curve model for univariate longitudinal data can be represented as the following hierarchical two-level model:

$$\text{(2.8)} \quad \begin{aligned} y_i(t) &= \Sigma_{k=1,q} f_k(t)\beta_{ik} + \varepsilon_i(t), \ t=1,\dots,T;\ i=1,2,\dots; \\ \beta_{ik} &= \mu_k + \zeta_{ik},\ k=1,\dots,q. \end{aligned}$$

The first equation of (2.8) describes $y_i(t)$ as the sum of weighted trend functions $f_k(t)\beta_{ik}$, k=1,...,q, and error $\varepsilon_i(t)$. To ease the presentation, the error $\varepsilon_i(t)$ will be considered to be independently normally distributed (Rovine & Molenaar, 2000, describe the use of several alternative error structures): $\boldsymbol{\varepsilon}_i \sim \aleph(\mathbf{0}, \boldsymbol{\Theta})$, where $\boldsymbol{\varepsilon}_i = [\varepsilon_i(1), \dots, \varepsilon_i(T)]'$ and $\boldsymbol{\Theta}$ is its (T,T)-dimensional diagonal covariance matrix. The second equation of (2.8) describes the weight β_{ik} of the k-th trend function $f_k(t)$ as a linear combination of a mean μ_k and a residual ζ_{ik}. It is assumed that the residuals ζ_{ik} obey a q-variate normal distribution: $\boldsymbol{\zeta}_i \sim \aleph(\mathbf{0}, \boldsymbol{\Phi})$, where $\boldsymbol{\zeta}_i = [\zeta_{i1}, \dots, \zeta_{iq}]'$ and $\boldsymbol{\Phi}$ is its (q,q)-dimensional covariance matrix. Consequently, $\boldsymbol{\beta}_i \sim \aleph(\boldsymbol{\mu}, \boldsymbol{\Phi})$, where $\boldsymbol{\beta}_i = [\beta_{i1}, \dots, \beta_{iq}]'$ and $\boldsymbol{\mu} = [\mu_1, \dots, \mu_q]'$.

The latent growth curve model (2.8) can readily be transformed to a confirmatory oblique q-factor model by defining the (T,q)-dimensional matrix of fixed loadings:

$$\text{(2.9)} \quad \Lambda = \begin{bmatrix} f_1(1) & f_2(1) & \cdots & f_q(1) \\ f_1(2) & f_2(2) & \cdots & f_q(2) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ f_1(T) & f_2(T) & \cdots & f_q(T) \end{bmatrix}$$

Letting $\mathbf{y}_i = [y_i(1), ..., y_i(T)]'$, (2.8) can be represented as:

(2.10) $\mathbf{y}_i = \Lambda\beta_i + \varepsilon_i$, $E[\mathbf{y}_i] = \Lambda\mu$, $\text{cov}[\mathbf{y}_i, \mathbf{y}_i] = \Lambda\Phi\Lambda' + \Theta$.

Clearly, (2.10) is a confirmatory oblique q-factor model and the latter model can be rewritten as a restricted latent simplex model, as is shown in section 2.1.5. Consequently, the latent growth model is nested under the latent simplex model. In fact, the latent growth curve model can be regarded as a confirmatory T-technique longitudinal factor model in which the factor loadings are fixed in such a way as to describe trend functions $f_k(t)$. This implies that the previous discussion of T-technique modeling carries over straightforwardly to the latent growth curve model. In particular, for a given subject there also exists perfect predictability at the latent level of the growth curve model.

The remainder of this section is devoted to illustrative analyses of an empirical longitudinal data set. The data have been obtained with 75 young adolescents who were administered a test measuring ego-strength at 6 repeated measurement occasions. The data were given to me several years ago by dr. Aline Sayers, then at Pennsylvania State University, with the request to fit a quasi-simplex model. Since then this data set has been regularly used as example in my annual structural equation modeling classes. The longitudinal means and covariance matrix are given below.

Longitudinal Covariance Matrix Ego-Strength
N=75

$$\begin{array}{c|llllll|}
t_1 & 361.14 & & & & & \\
t_2 & 202.33 & 307.88 & & & & \\
t_3 & 265.77 & 226.00 & 404.93 & & & \\
t_4 & 227.27 & 214.95 & 265.56 & 367.55 & & \\
t_5 & 227.64 & 186.76 & 289.48 & 239.37 & 477.63 & \\
t_6 & 109.31 & 105.10 & 111.55 & 134.50 & 196.30 & 287.15
\end{array}$$

Longitudinal Means Ego-Strength

t_1	t_2	t_3	t_4	t_5	t_6
141.44	142.65	144.73	147.83	162.13	172.85

To start with, only the longitudinal covariance matrix will be considered; later on the longitudinal means will be included. The simplest latent growth curve model is obtained by defining in (2.8) a constant function $f_1(t) = c$ and a linear trend $f_2(t) = a + bt$, $t=1,...,6$. It is customary (although not necessary) to choose $c = 1$ for the constant function $f_1(t)$. The values of a and b for the linear trend $f_2(t)$ also are arbitrary, hence in order to proceed we choose $a = 0$ and $b = 1$. This implies the following fixed pattern for Λ according to (2.9):

$$\Lambda' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{bmatrix}$$

It appears that this simple latent growth curve model does not yield a satisfactory fit to the observed longitudinal covariance matrix: the likelihood ratio against a general positive-definite alternative yields a chi-square of 26.69 with 12 degrees of freedom. Under the usual assumptions the postulated model has a probability of P = .0086.

It should be stressed that the choice of the real-valued constants a, b and c in, respectively, $f_2(t)$ and $f_1(t)$ is arbitrary (save for the special values b = 0 and c = 0, of course). For instance, taking c = 2 in $f_1(t)$ and b =2 in $f_2(t)$ yields the same chi-square of 26.69 with 12 degrees of freedom. One could say that the latent growth curve model is identified up to the values of these constants, in a similar way as the exploratory multi-factor model is identified up to rotation. Another way to formulate the same state of affairs is that $f_1(t)$ and $f_2(t)$ are comparable to contrasts in analysis of variance: affine transformation of a contrast function leaves the outcomes of an analysis of variance invariant, like affine transformation of $f_1(t)$ and $f_2(t)$ leaves the fit of the latent growth curve invariant.

A variable x that is specified up to affine transformation, $x^* = a + bx$, defines an interval scale (Suppes & Zinnes, 1963). The setting we are working in, the general linear model with normally distributed variables, is meant for interval scales of measurement and therefore should be invariant in the relevant sense under affine transformation. This is not the proper place to spell this out in detail, but in my opinion there is an unwarranted lack of interest in transformation theory in the social sciences (in contrast to, e.g., quantitative genetics; cf. Gianola, Im, Fernando, & Foulley, 1990; see also the impressive critique of Bookstein, 1990, on tests of gene-environment interaction effects). Regarding the simple latent growth curve model under consideration, it turns out that is not invariant in all relevant aspects under affine transformation. Rovine & Molenaar (1998) show that affine transformation of the linear trend $f_2(t)$ affects the correlation cor$[\beta_1, \beta_2]$ between the random weights associated with $f_1(t)$ and $f_2(t)$.

More specifically, let $a \neq a^*$ and define $f_2(t) = a + bt$ and $f_2(t)^* = a^* + bt$. Then the fit of the model with $\Lambda = [f_1(t), f_2(t)]$ will yield the same chi-square as the model with $\Lambda* = [f_1(t), f_2(t)^*]$, but cor$[\beta_1, \beta_2]$ will be different between these two model variants. For instance in the model already fitted to the longitudinal ego-strength covariance matrix, a = 0 and b = 1, yielding $f_2(t) = t$, t=1,...,6. For this model the estimated cor$[\beta_1, \beta_2] = -.63$. Taking $a^* = -5$ and b = 1 yields $f_2(t)^* = t - 5$, t=1,...,6. This model also yields a chi-square of 26.69 with 12 degrees of freedom, but the estimated cor$[\beta_1, \beta_2] = -.04$. In the first model the estimated negative correlation cor$[\beta_1, \beta_2] = -.63$ might be interpreted as the result of some law of initial value, in that a subject i having a relatively high level $\beta_{i1}f_1(t)$ is expected to show a relatively slow linear growth $\beta_{i2} f_2(t)$, and vice versa. But for the second model with affine transformed $f_2(t)^*$, the estimate cor$[\beta_1, \beta_2] = -.04$ and this can no longer be interpreted as the effect of some law of initial value.

In general, the correlation between normally distributed random variables is invariant under affine transformation of one or both of the variables. Yet this turns out not to be the case for the correlation between the normally distributed random weights in the latent growth curve model. In fact, the effects of affine transformation of the trend contrasts $f_k(t)$ in (2.8) can be more wide-spread. For instance, the regression of the random weights β_k on explanatory variables also is affected by affine transformation of the $f_k(t)$. All this raises important questions about the ways in which latent growth curve models behave under affine transformation. In particular, it raises doubts about the interpretability of the correlation between random weights in such models.

Returning to the example, a more elaborated latent growth curve model will be considered in an attempt to arrive at an acceptable goodness-of-fit to the observed longitudinal ego-strength covariance matrix. In addition to the constant level function $f_1(t)$ and the linear trend $f_2(t)$, a quadratic trend function $f_3(t)$ will be added to Λ. A general expression for a quadratic trend is $f_3(t) = d + et + gt^2$, where d, e and g > 0 are arbitrary real-valued constants. Presently, however, only the quadratic trend $f_3(t) = (t - h)^2$ will be considered, which is easily seen to be a special instance of the general quadratic expression. For the moment, let h = 0 , yielding $f_3(t) = t^2$. Hence Λ now has the following fixed pattern:

$$\Lambda' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 4 & 9 & 16 & 25 & 36 \end{bmatrix}$$

The fit of this model is acceptable: chi-square = 14.01 with 9 degrees of freedom, P = .12. Compared with the previous model fit, the difference in chi-square is 26.69 – 14.01 = 12.68 with 3 degrees of freedom, which appears to be a significant improvement (P = .0056).

The estimated (3,3)-dimensional covariance matrix Φ of the random weights has entries with very high estimated standard errors (given in parentheses):

$$\text{est-}\Phi = \begin{bmatrix} 82.02 & & \\ (129.70) & & \\ & & \\ 38.45 & 42.78 & \\ (69.76) & (46.47) & \\ & & \\ -5.36 & -8.52 & 1.70 \\ (9.12) & (6.57) & (.99) \end{bmatrix}$$

Also the estimated correlation between β_2 and β_3 is close to –1, implying that est-Φ is close to being nonsingular (but apparently not exactly nonsingular, because the Lisrel-8 program does not give a warning). All this is reason for concern about the model fit, despite the satisfactory value of the chi-square goodness-of-fit test.

There are various ways in which the questionable aspects of the obtained model fit can be tackled. Below I will discuss the possibility of removing the linear trend $f_2(t)$ and fit a model with only $f_1(t)$ and $f_3(t)$. But first some variants of the original model with all three trend functions will be considered. Our (conveniently restricted) definition of the quadratic trend is: $f_3(t) = (t - h)^2$, in which the value of h was chosen as h = 0. We now consider alternative values of h, for instance h = 5, yielding $f_3(t) = (t - 5)^2$. The remaining columns are kept the same as before: $f_1(t) = 1$ and $f_2(t) = t$, t=1,...,6. The chi-square is the same as before: 14.01 with 9 degrees of freedom, P = .12. But now the estimated covariance matrix of the random weights is:

$$\text{est-}\Phi = \begin{bmatrix} 1411.33 & & \\ (403.32) & & \\ & & \\ -226.73 & 42.23 & \\ (78.35) & (15.60) & \\ & & \\ -47.01 & 8.46 & 1.70 \\ (9.12) & (3.67) & (.99) \end{bmatrix}$$

It appears that now the estimated standard errors (within parentheses) of the elements of est-Φ are on the whole still larger than before, but the t-ratios of estimated (co-)variance and standard error have much larger absolute values. This means that most simple univariate 95% confidence intervals do not include the value zero. However, the estimated correlation between β_2 and β_3 is still extreme: it is close to 1 and therefore the condition of est-Φ still is worrisome.

Of course one could carry out a systematic search over all possible affine transformations of $f_2(t)$ and $f_3(t)$, and select the model that improves the quality of est-Φ the most. I deliberately leave the definition(s) of what constitutes this quality unspecified for further theoretical scrutiny. But the condition of est-Φ and the magnitudes of simple t-ratios are obvious candidates. I will not follow this approach any further here, because its rationale will be clear. Perhaps one can prove that this search will be in vain when it comes to removing certain aspects of the lack of quality of est-Φ. This constitutes a nice problem for mathematical statisticians, as part of the necessary work on the effects of transformations of variables to which I alluded earlier.

Another approach to arrive at a more satisfactorily fitting model is to consider latent growth curve models with only a subset of the trend functions $\{f_k(t), k=1,2,3\}$. In particular, I will focus on models involving only the two latent growth curves $f_1(t) = 1$ and $f_3(t) = (t – h)^2$, h = 0, .. ., 6. The obtained 7 model fits yield the following results:

h	chi-square	df	P
0	22.92	12	.028
1	21.26	12	.047
2	17.82	12	.120
3	19.94	12	.068
4	30.11	12	.003
5	31.40	12	.002
6	30.17	12	.003

Clearly, the models are no longer equivalent in the sense of yielding the same chi-square values. The model in which h = 2 and $f_3(t) = (t – 2)^2$ yields an acceptable chi-square value, whereas the model in which h = 5 and $f_3(t) = (t – 5)^2$ yields the worst chi square value. This raises another question that to the best of my knowledge has not yet been worked out in the published literature: in latent growth curve models involving only a constant curve and a quadratic curve, the chi-square goodness-of-fit test is not invariant under a shift of the zero point of the time axis.

Stated in a more general way, it appears that some nonlinear growth curve models are not invariant under affine transformation. This finding raises additional issues, in particular about the proper choice of the range of affine transformations of the time variable in nonlinear polynomials in time. For instance, in the quadratic trend curve $f_3(t) = (t - h)^2$, only seven integer-valued shifts h = 0, ..., 6 of the zero point of the time axis have been considered. This range of h values is quite arbitrary in that one could consider negative values for h, as well as noninteger values of h. Because the latter possibility of noninteger values for h involves a search over an uncountable infinity of model variants, it would seem to be solvable only by incorporating the shift h of the zero time point as an additional free parameter in the model fit. But then it should be assumed that the likelihood function (or any other criterion function) has a unique minimum as function of h. I do not know whether this assumption is always unproblematic.

In the context of this illustrative application we proceed with the solution for h = 2 (P = .12). The relevant details of this solution are:

$$\Lambda' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 4 & 9 & 16 \end{bmatrix}$$

$$\text{est-}\Phi = \begin{bmatrix} 261.39 & \\ (49.33) & \\ & \\ -8.45 & .86 \\ (2.77) & (.30) \end{bmatrix}$$

The elements of the estimate of diag-Θ are 88.00 (27.16), 123.96 (27.36), 117.82 (25.36), 133.36 (25.96), 219.54 (40.43), and 93.83 (50.03). The estimated correlation between the random weights associated with the constant level and the quadratic trend is est-cor[β_1, β_3] = -.56.

This latent growth curve model for the longitudinal ego-strength data, with a constant level and quadratic trend, will be rewritten as a restricted latent simplex model according to the rules set out previously. This will provide a nontrivial illustration of my claim that latent growth curve models are nested under latent simplex models. It is nontrivial in that a) there is a quadratic latent growth curve involved, b) this quadratic latent growth curve has random weight that is correlated with the random weight of the constant level function, and c) the quadratic latent growth curve is a polynomial in time t – h, where h = 2, t=1, ..., 6 (hence the time axis has undergone an affine transformation). The steps in the construction of the restricted latent simplex analogue will be made explicit for the benefit of the reader, although this involves some degree of repetition. Within the context of this illustration, the latent growth curve concerned will be referred to as the LGC, while its restricted latent simplex analogue will be referred to as the RLS.

The LGC is a confirmatory oblique 2-factor model. As explained in section 2.1.5, each factor (random weight of latent growth curve) is described by a univariate latent simplex having only innovation at the initial time point. First the weighted constant level in the LGC is considered: $\beta_{i1} f_1(t)$, where $f_1(t) = 1$, t=1, ..., 6. For this constant level component the following univariate latent simplex is defined:

*) $\eta_{1i}(t) = \gamma_{t,t-1}\eta_{1i}(t-1)$, t=2,...,6; $\eta_{1i}(1) = \varsigma_{1i}(1)$.

The coefficients $\gamma_{t,t-1}$ in *) will be derived from the fixed values of $f_1(t)$. But first another univariate latent simplex is defined for the weighted quadratic trend curve in the LGC:

**) $\eta_{2i}(t) = \delta_{t,t-1}\eta_{2i}(t-1)$, t=2,...,6; $\eta_{2i}(1) = \varsigma_{2i}(1)$.

Again, the coefficients $\delta_{t,t-1}$ in **) will be derived from the fixed values of $f_3(t) = (t - h)^2$. Together, *} and **) define a 12-variate vector

$$\boldsymbol{\eta}_i' = [\eta_{1i}(1), ..., \eta_{1i}(6), \eta_{2i}(1), ..., \eta_{2i}(6)].$$

Because the observed longitudinal ego-strength scores of subject i define the 6-variate vector

$$\mathbf{y}_i = [y_i(1), ..., y_i(6)]'$$

it follows that the matrix $\boldsymbol{\Lambda}$ in the RLS is the (6,12)-dimensional matrix

$$\boldsymbol{\Lambda} = [\mathbf{I}_6, \mathbf{I}_6]$$

Hence $\boldsymbol{\Lambda}$ consists of a concatenation of two (6,6)-dimensional unity matrices. Hence the relationship between $\mathbf{y}_i$ and $\boldsymbol{\eta}_i$ is given by:

$$\mathbf{y}_i = [\mathbf{I}_6, \mathbf{I}_6]\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i,$$

where $\boldsymbol{\varepsilon}_i$‘= $[\varepsilon_{1i}, ..., \varepsilon_{8i}]$ and cov$[\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_i$‘$] = \boldsymbol{\Theta}$ is a diagonal (6,6)-dimensional covariance matrix.

The coefficients $\gamma_{t,t-1}$ in *) have to be obtained from the fixed values of $f_1(t)$ = 1, t = 1, ..., 6. Notice that $f_1(t)$ constitutes the first column in the matrix of factor loadings of the LGC. It was shown in section 2.1.3 that factor loadings are a function of the coefficients of a restricted univariate latent simplex: Accordingly, the first column of factor loadings in the LGC are related to the coefficients $\gamma_{t,t-1}$ in *) in the following way:

$$f_1(1) = 1 = 1$$
$$f_1(2) = 1 = \gamma_{2,1}$$
$$f_1(3) = 1 = \gamma_{3,2}\gamma_{2,1}$$
$$f_1(4) = 1 = \gamma_{4,3}\gamma_{3,2}\gamma_{2,1}$$
$$f_1(5) = 1 = \gamma_{5,4}\gamma_{4,3}\gamma_{3,2}\gamma_{2,1}$$
$$f_1(6) = 1 = \gamma_{6,5}\gamma_{5,4}\gamma_{4,3}\gamma_{3,2}\gamma_{2,1}$$

Hence the coefficients $\gamma_{t,t-1}$ can be obtained form the fixed values of $f_1(t)$ according to the following rule:

$$\gamma_{t,t-1} = f_1(t) / f_1(t-1), t=2, ..., 6.$$

It follows easily that $\gamma_{t,t-1} = 1$ for t = 2, ..., 6.

In a similar vein the coefficients $\delta_{t,t-1}$ in **) are derived from the fixed values of $f_3(t) = (t - h)^2$. Given that h = 2, we have that $f_3(2) = 0$. This zero loading can be taken care of in the computation of the $\delta_{t,t-1}$ coefficients in various ways. For instance, $\eta_{2i}(2)$ can be skipped in the latent simplex **) associated with $f_3(t)$. According to this approach $\eta_{2i}(3)$ is directly related to $\eta_{2i}(1)$ by

$$\eta_{2i}(3) = \delta_{3,1}\eta_{2i}(1)$$

while for $\eta_{2i}(t)$, $t \geq 3$, the equations as specified by **) still hold. This leads to the following set of equations:

$$\begin{aligned} f_3(1) &= \ \ 1 = 1 \\ f_3(3) &= \ \ 1 = \delta_{3,1} \\ f_3(4) &= \ \ 4 = \delta_{4,3}\delta_{3,1} \\ f_3(5) &= \ \ 9 = \delta_{5,4}\delta_{4,3}\delta_{3,1} \\ f_3(6) &= 16 = \delta_{6,5}\delta_{5,4}\delta_{4,3}\delta_{3,1} \end{aligned}$$

It then follows easily that $\delta_{3,1} = f_3(3) / f_3(1) = 1$, $\delta_{4,3} = f_3(4) / f_3(3) = 4$, $\delta_{5,4} = f_3(5) / f_3(4) = 2.25$, and $\delta_{6,5} = f_3(6) / f_3(5) = 1.778$.

Finally, the pattern of the (12,12)-dimensional covariance matrix Ψ of the innovations in the RLS has to be specified. According to *) there is only an innovation $\varsigma_{1i}(1)$ associated with $\eta_{1i}(1)$, while according to **) there is only an innovation $\varsigma_{2i}(1)$ associated with $\eta_{2i}(1)$. Because $\eta_{1i}(1)$ is the 1-st element of $\boldsymbol{\eta}_i$, and $\eta_{2i}(1)$ is the 7-th element of $\boldsymbol{\eta}_i$, only the diagonal elements $\psi_{11} = \text{var}[\varsigma_{1i}(1)]$ and $\psi_{77} = \text{var}[\varsigma_{2i}(1)]$ are free parameters. Moreover, $\psi_{71} = \text{cov}[\varsigma_{1i}(1), \varsigma_{21}(1)]$ is a free parameter, expressing the covariance between the random weights in the LGC. This completes the specification of the RLS representing the LGC.

The fit of the RLS to the longitudinal covariance matrix of the ego-strength data is in all details the same as the fit of the LGC. The RLS chi-square is 17.82 with 12 degrees of freedom (P = .12). In addition, est-ψ_{11} = 261.39 (49.33) in the RLS equals est-ϕ_{11} in the LGC, est-ψ_{77} = .86 (.30) equals est-ϕ_{22}, and est-ψ_{71} = -8.45 (2.77) in the RLS equals est-ϕ_{21} in the LCG. Also the estimated diagonal elements of $\boldsymbol{\Theta}$ in the RLS are the same as for the GLC.

If we return to the GLC and now fit (2.8) to the longitudinal mean trend and covariance matrix, the chi-square is 25.96 with 16 degrees of freedom (P = .055). In comparison with the fit of the LGC to only the longitudinal covariance matrix, the difference in chi-square values is 8.14 with 4 degrees of freedom (P = .085). This difference can be neglected, hence we conclude that the LCG yields a satisfactory fit to both the longitudinal trend and covariance matrix of the ego-strength data. The estimated mean of β_{1i} associated with the constant level est-μ_1 = 141.56 (2.01). The estimated mean of β_{2i} associated with the quadratic trend is est-μ_2 = 1.98 (.14). The remaining parameter estimates are about the same as in the LCG for the longitudinal covariance matrix given earlier. The fit of the RLS to the longitudinal ego-strength means and covariances yields exactly the same chi-square value, estimated parameters and standard errors as the LCG. In particular, the estimated means of $\varsigma_{1i}(1)$ and $\varsigma_{2i}(1)$ are equal to est-μ_1 and est-μ_2 , respectively.

I would like to close the discussion of this example with a presentation of what might be one of the best fitting models to the longitudinal ego-strength means and covariance matrix. This is the following quasi-simplex (latent univariate simplex; cf. (2.1)):

$$\eta_i(t) = \beta_{t,t-1}\eta_i(t-1) + \varsigma_i(t), t=2,...,6,$$

est-$\beta_{2,1}$ = 1.01 (.001)
est-$\beta_{3,2}$ = 1.02 (.001)
est-$\beta_{4,3}$ = 1.02 (.001)
est-$\beta_{5,4}$ = 1.10 (.002)
est-$\beta_{6,5}$ = 0.47 (.100)

var[$\varsigma_i(1)$] = 219.76 (40.65)
var[$\varsigma_i(5)$] = 164.83 (79.84)
var[$\varsigma_i(6)$] = 124.23 (91.22)

E[$\varsigma_i(1)$] = 141.46 (2.18)
E[$\varsigma_i(6)$] = 96.25 (17.01)

var[ε_{1i}] = 132.79 (18.46)
var[ε_{2i}] = 132.79 (18.46)
var[ε_{3i}] = 118.36 (24.66)
var[ε_{4i}] = 126.07 (26.11)
var[ε_{5i}] = 68.71 (73.55)
var[ε_{6i}] = 68.71 (73.55)

This quasi-simplex model yields a chi-square of 11.43 with 13 degrees of freedom (P = .57). In fact, most of the β-parameters could be restricted to be equal to each other, while the measurement error variances at time points 5 and 6 also could be restricted to equal zero. This would yield an even more flattering chi-square value for this model. Notice that the innovation variances at time points 2, 3 and 4 have been fixed at zero. Also the means of the innovations at time points 2, 3, 4 and 5 have been fixed at zero. Finally, note that the innovations variance at time point 6 has a rather small t-ratio. I find this latent univariate simplex quite appealing, not only because of its excellent fit to the data, but also because it allows for an interesting interpretation of the process underlying ego-strength development. The development of ego-strength appears to be quite stable until time point 5 in that there are no genuine innovations at time points 2, 3 and 4 (remember that the initial time point 1 is special because $\eta_i(1)$

cannot be explained: $\eta_i(1) = \varsigma_i(1)$). Also the means of the innovations at these time points 2, 3 and 4 equals zero. Then, at time point 5, zero-mean random innovations are injected into the developmental process: var[$\varsigma_i(5)$] = 164.83 and E[$\varsigma_i(5)$] = 0. At time point 6 this situation is more or less inverted in that the innovations have much less variance, but nonzero mean: var[$\varsigma_i(6)$] = 96.25 and E[$\varsigma_i(6)$] = 96.25. Hence the change in mean ego-strength (at time point 6) is preceded by a change in inter-individual variation in ego-strength (at time point 5). Such a dynamic pattern is not uncommon for processes undergoing some kind of phase transition (van der Maas & Molenaar, 1992) or entering a sensitive period. Potentially interesting aspects of the development of ego-strength indeed!

We have come at the close of part one of this chapter, in which it has been shown that factor models and latent growth curve models are nested under latent simplex models. In the present section an elaborate illustration of this nesting relationship was shown for a latent growth curve model and a restricted bivariate latent simplex. In addition, an excellently fitting latent univariate simplex (quasi-simplex) was considered for the same data. Perhaps the reader wonders what might possibly be the relationship between the bivariate latent growth curve model cum restricted bivariate latent simplex on the one hand, and the univariate latent simplex on the other hand. In the next part of this chapter, some new light will be casted on the latter question. There, it will be shown that there is a much more strict sense in which factor models, latent growth curve models and their restricted simplex analogues are related to the quasi-simplex (univariate latent simplex). An interesting offspin of this discussion is a proof of how all standard latent variable models can be rewritten as models having no latent variables.

2.2 A general scheme to manipulate latent variables, in particular to remove latent variables from structural equation models

The centrality of latent simplex models, discussed in the previous part of this chapter, opens up the possibility to focus attention on these models. Results obtained for latent simplex models immediately carry over to factor models and latent growth curve models, because the latter models are special instances of the former. This state of affairs has some intellectual appeal, as well as some consequences for model selection procedures like likelihood ratio tests, but it certainly is not my intention to recommend the rather elaborate restricted latent simplex analogues of factor models and latent growth curve models in applied data analysis. There is, however, an additional aspect of the centrality of latent simplex models that has much more far-reaching consequences. A latent simplex has a model structure that resembles a well-known class of time series models, the class of autoregressive moving-average (ARMA) models. This implies that theoretical results obtained for ARMA models may carry over to latent simplex models, and hence also to factor and latent growth curve models. In what follows I will present one such theoretical result for ARMA models and show how a suitably adapted version of this result also applies to latent simplex models. I then will elaborate some of the implications of this result, in particular it will be shown how it is possible to transform latent simplex (factor, latent growth curve) models into models without latent variables. Moreover, and this I consider to be an equally intellectually inspiring result, it will be shown how latent q-variate simplex models (q-factor models, q-variate latent growth curve models; q > 1) can be rewritten as latent univariate simplex models. This opens up some new perspectives. For instance, I will present a structural model involving a latent simplex as well as a latent factor and rewrite that model as a latent univariate simplex. In the original presentation of this model, the latent simplex has been associated with a state-like process having less than perfect stability, whereas the latent factor has been associated with a stable trait-like process (e.g., Hewitt, Eaves, Neal, & Meyers, 1988). Given that the state-like simplex and the trait-like factor collapse into a single latent simplex, it would appear that the distinction between state-like and trait-like influences may not involve a fundamental qualitative difference.

To the best of my knowledge, much of what is presented below is new in the field of structural equation modeling. Although the prerequisite ingredients from time series analysis and state-space modeling are in themselves simple, they will be described in the next section 2.2.1 in sufficient detail to make the argumentation readily accessible to structural equation modelers. At the end of that section, a theorem on the addition of ARMA models will be presented. This theorem will turn out to have important consequences for structural equation models.

I realize that the contents of what follows contains much that is new to many structural equation modelers. It therefore may be worthwhile to give some guidelines that may be helpful in a first reading of this material. To start with, the definition of ARMA models and their variants (like the NARMA models introduced later on) should be taken at face value. These time series models are used as empirical models with which (sequential) dependencies are described with as few parameters as possible. It is not advised to try to interpret these models (although such interpretations can be given): they are only used as convenient model structures enabling the removal of latent variables from structural equation models. Only at the close of this chapter some interpretations of what has been accomplished in terms of these time series models will be given. Secondly, in the course of the discussion some rather involved algebraic manipulations will be presented. It should be kept in mind that these algebraic manipulations serve a particular purpose, namely to prove the equivalence of models with and without latent variables. Apart from this, all models under consideration (with and without latent variables, ARMA, NARMA, etc.) are in themselves linear models that can straightforwardly be fitted by means of the Lisrel program (or equivalent programs). It is only the proof of their equivalence that requires some more involved calculations. Thirdly, and lastly, the basic tenet of the following discussion is to show that latent variables in factor models, latent growth curve models and latent simplex models can be transformed away. This has important implications for the status of these latent variables. An additional result that will be proven is that different latent variables can be added (e.g., two factors in a factor model can be added). Again, this has interesting implications for our understanding of what latent variables are. In the final sections of this chapter it will be shown that such manipulations of latent variables (addition, removal) constitute the beginnings of a general transformation theory for structural equation models that has its roots in mathematical systems theory.

2.2.1 Preliminary results from time series analysis

In this section all time series are univariate series. Some definitions of time series models will be given in which mainly the occurrence of certain polynomials in a formal operator is emphasized. Manipulation of these polynomials characterizing time series models will enable a simple proof of the main result of this section, namely a theorem given by Granger & Morris (1976).We start with the definition of autoregressions, an instance of which was already discussed in chapter 1. The general definition of an autoregression of order p, denoted by AR(p), where $p \geq 0$ is an integer, is:

$$y(t) + a_1y(t-1) + a_2y(t-2) + ... + a_py(t-p) = e(t), t=0, \pm1, ...$$

The term e(t) denotes white noise (cf. chapter 1): $E[e(t)] = 0$ and $cov[e(t), e(t+u)] = \delta(u)\sigma^2$, u=0, ±1, ..., where $\delta(u)$ is Kronecker's delta. Hence e(t) lacks sequential dependence.

No attempt will be made to interpret AR(p) models; for this the reader is referred to the many excellent textbooks (e.g., Box & Jenkins, 1970; Anderson, 1971; Shumway & Stoffer, 2000). Here we only consider a particular representation of the AR(p). Let *B* denote the so-called backward shift operator whose action is defined by:

$$By(t) = y(t-1)$$

It then follows immediately that $B^v y(t) = y(t-v)$, where $v \geq 0$ is an integer. An AR(p) can therefore be represented as:

$$[1 + a_1B + a_2B^2 + ... + a_pB^p]y(t) = e(t).$$

A convenient notation for $[1 + a_1B + a_2B^2 + ... + a_pB^p]$, the p-th order polynomial in the backshift operator *B*, is a[*B*,p], resulting in the following representation of an AR(p):

$$a[B,p]y(t) = e(t).$$

The presumed stationarity of y(t) implies that the absolute values of all roots of a[*B*,p] = 0 are greater than 1 (notice that in general the roots are complex-valued). Substituting $B^v = \exp[i2\pi v]$, v=1, ..., p, where $i = \sqrt{-1}$ is the imaginary unit, yields the discrete Fourier transform of the AR(p) underlying spectral analysis of y(t). Also, the autocovariance function cov[y(t), y(t+u)], u = 0, ±1, ... is a function of a[*B*,p].

The definition of a moving-average of order q, denoted by MA(q), where $q \geq 0$ is an integer, is:

$$y(t) = e(t) + b_1e(t-1) + b_2e(t-2) + ... + b_qe(t-q),\ t=0, \pm1, ...$$

where e(t) is again white noise with mean zero and variance σ^2. Employing the backward shift operator *B* again, the MA(q) can be represented as:

$$y(t) = b[B,q]e(t)$$

where $b[B,q] = [1 + b_1B + b_2B^2 + ... + b_qB^q]$ denotes the q-th order polynomial in the backshift operator *B*. For finite order q, an MA(q) is always stationary. But special care has to be taken about the choice of the roots of b[*B*,q] = 0 so that the inverse $b[B,q]^{-1}$ exists (cf. Molenaar, 1999). Like for an AR(p), the spectrum and the autocovariance function of an MA(q) can be derived from b[*B*,q].

The definition of an autoregressive moving-average of order p and q, denoted by ARMA(p,q), is:

$$a[B,p]y(t) = b[B,q]e(t),\ t=0, \pm1, ...$$

where a[*B*,p] is defined in the same way as for an AR(p), b[*B*,q] is defined in the same way as for an MA(q), and e(t) is white noise with mean zero and variance σ^2. The spectrum and autocovariance function of an ARMA(p,q) are a function of a[*B*,p] and b[*B*,q]. Notice that an ARMA(p,q) equals an MA(∞), that is a moving average of infinite order The polynomial in the backward shift operator *B* associated with this ARMA(p,q) = MA(∞) is given by the ratio $b[B,q]a[B,p]^{-1}$. Hence one possible interpretation of an ARMA(p,q) is that it yields an economical model involving p + q + 1 parameters (including the variance σ^2 of e(t)) for an MA(∞) involving an unbounded number of parameters. In a similar vein, an ARMA(p,q) equals an AR(∞), where the latter AR(∞) has a polynomial in the backward shift operator *B* given by the ratio $a[B,p]b[B,q]^{-1}$. Hence another possible interpretation of an ARMA(p,q) is that it yields an economical model for an AR(∞) involving an unbounded number of parameters. Other possible interpretations of an ARMA(p,q) are given in the textbooks alluded to earlier as well as by Granger & Morris (1976).

The definition of an ARMA(p,q), including the definitions of an AR(p) = ARMA(p,0) and an MA(q) = ARMA(0,q) as special cases, is all we need for proving a theorem of Granger & Morris (1976). The theorem concerned characterizes the sum of two independent ARMA processes, where each ARMA process is considered to be weakly stationary. This theorem will turn out to have important consequences for structural equation models. First the theorem will be proved, using only simple aspects of the formal definition of an ARMA(p,q) given above. In fact, only the orders of products of polynomials in the backward shift operator *B* will figure in the proof. This should enable readers unacquainted with time series analysis to follow the proof.

Consider two ARMA processes, x(t) and y(t), which each are weakly stationary. Specifically, E[x(t)] = c_x, cov[x(t), x(t+u)] = c_x(u), E[y(t)] = c_y, and cov[y(t), y(t+u)] = c_y(u). Without loss of generality, the means of x(t) and y(t) are taken to be zero: $c_x = c_y = 0$. In addition, importantly, it is assumed that the crosscovariance function between x(t) and y(t) is zero at all lags u: cov[x(t), y(t+u)] = 0 for all u = 0, ±1, ... The process models for x(t) and y(t) are, respectively, ARMA(p,q) and ARMA(m,n):

$$a_x[B,p]x(t) = b_x[B,q]e_x(t)$$

$$a_y[B,m]y(t) = b_y[B,n]e_y(t)$$

where e_x(t) and e_y(t) are white noise series with variance σ_x^2 and σ_y^2, respectively.

Now consider the sum of x(t) and y(t): z(t) = x(t) + y(t). Hence z(t) is the sum of an ARMA(p,q) and an ARMA(m,n). Then the following theorem about z(t) can be proved:

Theorem (Granger & Morris, 1976; Box & Jenkins, 1970)

Let x(t) be a weakly stationary zero mean ARMA(p,q) and y(t) a weakly stationary zero mean ARMA(m,n). Let x(t) and y(t) be weakly orthogonal, i.e, the crosscovariance function of x(t) and y(t) is zero at all lags. Then z(t) = x(t) + y(t) is a weakly stationary zero mean ARMA(r,s), where r ≤ p + m and s ≤ max[p + n, q + m]

Schematic proof.

As z(t) = x(t) + y(t), it follows that multiplication of this equality by $a_x[B,p]a_y[B,m]$ yields:

*) $a_x[B,p]a_y[B,m]z(t) = a_y[B,m]a_x[B,p]x(t) + a_x[B,p]a_y[B,m]y(t)$

But $a_x[B,p]x(t) = b_x[B,q]e_x(t)$ and $a_y[B,m]y(t) = b_y[B,n]e_y(t)$. Substitution of the latter equalities in the right-hand side of *) yields:

**) $a_x[B,p]a_y[B,m]z(t) = a_y[B,m]b_x[B,q]e_x(t) + a_x[B,p]b_y[B,n]e_y(t)$

The order r of the polynomial product $a_x[B,p]a_y[B,m]$ in the left-hand side of **) is not larger than p + m. It can be smaller than p + m in case $a_x[B,p]$ and $a_y[B,m]$ have common roots, which can be removed from $a_y[B,m]$ before the multiplication by $a_x[B,p]a_y[B,m]$ to arrive at *). Hence r ≤ p + m. The assumption that x(t) and y(t) are weakly orthogonal implies that the crosscovariance function between e_x(t) and e_y(t) is zero at all lags. In the right-hand side of **), the order of $a_y[B,m]b_x[B,q]$ is not larger than q + m (smaller than q + m in case common roots have been removed from $a_y[B,m]$) and the order of $a_x[B,p]b_y[B,n]$ equals p + n. Hence the order s of the sum of polynomial products $a_y[B,m]b_x[B,q]$ and $a_x[B,p]b_y[B,n]$ cannot be larger than the maximum order of the summands: s ≤ max[p + n, q + m].

The proof follows Granger & Morris (1976), but it is schematic in that it does not address a technicality concerning the addition of the two moving average components of the right-hand side of **). This technicality involves again the choice of the roots of the polynomial in the backward shift operator *B* associated with moving-averages (cf. Molenaar, 1999), the details of which can be found in Granger & Morris (1976). On the other hand, the proof has been elaborated a bit in order to convey as clearly as possible its basic structure. This will provide a convenient starting point for our investigation in the next section of the possibility to remove some of the restrictive conditions under which the proof has been obtained, in order to make the theorem applicable to structural equation models.

The theorem implies that the sum of two weakly orthogonal as well as stationary ARMA processes is again a weakly stationary ARMA process, where the order of the latter ARMA process bears a simple relationship to the orders of the summands. To give an elementary example, let x(t) be a first-order autoregressive process, i.e., x(t) is an ARMA(1,0). In addition, let y(t) be a white noise process, i.e., y(t) is an ARMA(0,0). Assuming that x(t) and y(t) are weakly stationary and weakly orthogonal, the theorem implies that their sum z(t) = x(t) + y(t) is an ARMA(1,1). It will be shown in the next section that this ARMA(1,1) is closely related to the quasi-simplex model, defined as the sum of a latent simplex and measurement error.

2.2.2 The addition of nonstationary simplex processes

In this section the theorem of Granger & Morris, henceforth referred to as TGM, will be generalized so as to accommodate the addition of nonstationary simplex processes occurring in a structural equation model. This will require the removal of some of the restrictive assumptions underlying the TGM. But we also need to transfer the generalized TGM thus obtained from its original context within time series analysis to the new context of structural equation modeling. The latter change of context involves the shift from an analysis of within-subject covariation to an analysis of between-subject covariation, which will be discussed in the next section.

TGM applies to weakly stationary ARMA processes. I will first elaborate somewhat further the concept of stationarity and the various ways in which a process can be nonstationary. It then can be specified unambiguously in which sense simplex processes occurring in structural equation models are allowed to be nonstationary without affecting TGM’s applicability. Throughout this section, all time series are assumed to have zero mean function.

Consider one of the simplest ARMA processes, namely a zero mean first-order autoregression, AR(1) = ARMA(1,0), given by:

$$a[B,1]y(t) = y(t) + a_1 y(t-1) = e(t),\ t = 0, \pm 1, \ldots$$

This model has to obey the following restrictions in order to describe a weakly stationary process:

- the order of the polynomial a[*B*,1] has to be constant in time
- the coefficient of the polynomial a[*B*,1] has to be constant in time
- the absolute value of the root of a[*B*,1] = 0 has to be larger than 1

- the variance of the white noise e(t) has to be constant in time

The first requirement stipulates that y(t) is an AR(1) for all times t. It rules out the possibility of a structural change at some time point t_0, after which $y(t \geq t_0)$ would become another type of process (e.g., an ARMA(p,q) with $p \neq 1$ and/or $q > 0$). The second requirement restricts a_1 to be a time-invariant coefficient, while the third requirement implies that $| a_1 | < 1$. The fourth requirement stipulates that var[e(t)] = σ^2 is time-invariant. A model obeying these four restrictions is called a weakly stationary AR(1). In a similar vein, a general ARMA(p,q) model obeying suitably adapted variants of these four restrictions is called a weakly stationary ARMA(p,q).

In case an ARMA process does not obey all four requirements, it is nonstationary. Hence there are at least $2^4 - 1 = 15$ different kinds of nonstationary ARMA processes (there are additional kinds of nonstationarity which are not tied up with the four requirements given here; e.g., Cohen, 1995). To keep things a bit manageable, it will be assumed that the first requirement is always met. That is, it is assumed that ARMA processes do not undergo structural changes in the sense indicated above. We therefore restrict attention to nonstationary ARMA(p,q) processes with constant order p and q. The general expression for such a nonstationary ARMA process, referred to as NARMA(p,q), is:

$$a_t[B,p]y(t) = b_t[B,q]e(t), \ t=0, \pm 1, ...$$

$$\text{var}[e(t)] = \sigma(t)^2 \qquad (2.11)$$

$$a_t[B,p] = [1 + a_{1,t}B + a_{2,t}B^2 + ... + a_{p,t}B^p]$$

$$b_t[B,q] = [1 + b_{1,t}B + b_{2,t}B^2 + ... + b_{q,t}B^q]$$

The polynomials $a_t[B,p]$ and $b_t[B,q]$ in (2.11) have time-varying coefficients, while the variance $\sigma(t)^2$ of e(t) also is time-varying. Of course, not all these parameters of (2.11) have to be time-varying simultaneously. For instance, a simple instance of a NARMA(1,0) is one in which only the variance of e(t) is time-varying: $y(t) + a_1 y(t-1) = e(t)$, var[e(t)] = $\sigma(t)^2$.

At the close of this section, I will discuss further, more heuristic, aspects of NARMA processes. But first we proceed to the main result of this section: I claim that nothing in the proof of the TGM precludes its application to NARMA processes. The built-up of this proof was exceedingly simple:

- multiply z(t) = x(t) + y(t) by $a_x[B,p]a_y[B,m]$
- substitute for the ARMA expressions for x(t) and y(t)
- count the orders of the products of polynomials in *B* thus obtained

Nothing in these three steps of the proof hinges on the stationarity of x(t) and y(t).

Stated more specifically, nothing in this proof exploits either the time-invariance of the polynomials in the backward shift operator *B* making up the ARMA models for x(t) and y(t), or the constancy of the variances of the innovations $e_x(t)$ and $e_y(t)$. Hence my conjecture that the TGM generalizes straightforwardly to the addition of NARMA processes:

Conjecture

Let x(t) and y(t) be two weakly orthogonal zero mean NARMA(p,q) and NARMA(m,n) processes, respectively:

$$a_{x,t}[B,p]x(t) = b_{x,t}[B,q]e_x(t)$$

$$a_{y,t}[B,m]y(t) = b_{y,t}[B,n]e_y(t)$$

where $e_x(t)$ and $e_y(t)$ are white noise series with variance $\sigma_x(t)^2$ and $\sigma_y(t)^2$, respectively. Let z(t) be the sum of x(t) and y(t): z(t) = x(t) + y(t). Then z(t) is a zero mean NARMA(r,s) process, where $r \leq p + m$ and $s \leq \max[p + n, q + m]$

Proof.

As z(t) = x(t) + y(t), it follows that multiplication of this equality by $a_{x,t}[B,p]a_{y,t}[B,m]$ yields:

*) $$a_{x,t}[B,p]a_{y,t}[B,m]z(t) = a_{y,t}[B,m]a_{x,t}[B,p]x(t) + a_{x,t}[B,p]a_{y,t}[B,m]y(t)$$

But $a_{x,t}[B,p]x(t) = b_{x,t}[B,q]e_x(t)$ and $a_{y,t}[B,m]y(t) = b_{y,t}[B,n]e_y(t)$. Substitution of the latter equalities in the right-hand side of *) yields:

**) $$a_{x,t}[B,p]a_{y,t}[B,m]z(t) = a_{y,t}[B,m]b_{x,t}[B,q]e_x(t) + a_{x,t}[B,p]b_{y,t}[B,n]e_y(t)$$

The order r of the polynomial product $a_{x,t}[B,p]a_{y,t}[B,m]$ in the left-hand side of **) is not larger than p + m. It can be smaller than p + m in case $a_{x,t}[B,p]$ and $a_{y,t}[B,m]$ have common roots, which can be removed from $a_{y,t}[B,m]$ before the multiplication by $a_{x,t}[B,p]a_{y,t}[B,m]$ to arrive at *). Hence $r \leq p + m$. The assumption that x(t) and y(t) are weakly orthogonal implies that the crosscovariance function between $e_x(t)$ and $e_y(t)$ is zero at all lags. In the right-hand side of **), the order of $a_{y,t}[B,m]b_{x,t}[B,q]$ is not larger than q + m (smaller than q + m in case common roots have been removed from $a_{y,t}[B,m]$) and the order of $a_{x,t}[B,p]b_{y,t}[B,n]$ equals p + n. Hence the order s of the sum of polynomial products $a_{y,t}[B,m]b_{x,t}[B,q]$ and $a_{x,t}[B,p]b_{y,t}[B,n]$ cannot be larger than the maximum order of the summands: $s \leq \max[p + n, q + m]$.

The proof of the conjecture is the same as the schematic proof of the TGM, only the time-invariant polynomials occurring in the proof of the TGM have been replaced by time-varying polynomials in the proof of the conjecture. In fact, the proof of the conjecture is no longer schematic like the proof of the TGM, because the technical aspects associated with weak stationarity are not relevant for NARMA processes. In this sense the proof of the conjecture is even simpler than the proof of the TGM.

2.2.3 Addition of NARMA processes in structural equation models

In this section we touch upon a theme that will be discussed more fully in the next chapter, namely the relationship between time series analysis and standard structural equation model fitting. In a time series analysis in its barest form, one has available a single stretch of repeated measurements obtained with a single subject (system, case). The analysis then proceeds by considering the within-subject variation, taking the subject fixed and generalizing results across all time points at which the process under scrutiny is defined. In contrast, in the fit of longitudinal structural equation models one has available multiple stretches of repeated measurements replicated over many subjects (systems, cases). The analysis then proceeds by a consideration of between-subject variation, taking the time points fixed and generalizing results over the population from which the subjects have been sampled randomly.

Apart from all kinds of qualifications, statistical estimation in time series analysis consists of taking averages over time points, whereas estimation in longitudinal analysis involves taking averages over subjects. I refer to taking averages over time points as an analysis of within-subject variation, while taking averages over subjects is referred to as an analysis of between-subject variation. As far as statistical estimation is concerned, there are no fundamental differences between analyses of within-subject variation and between-subject variation. Of course there are various differences having to do with the sequential dependence of terms entering averages over time points in an analysis of within-subject variation, in comparison with the measurement independence of terms entering the averages over cases in an analysis of between-subject variation. But the principles underlying statistical estimation in time series analysis and signal analysis are the same as the principles underlying multivariate statistical analysis in general (despite sometimes large differences in implementation and detail). For further elaborations the reader is referred to the excellent overview by Wooldridge (1994).

Analyses of within- and between-subject variation not only share the same principles of statistical estimation theory, but also the same kinds of statistical models. Brillinger (1975) presents time series analogues of all standard multivariate statistical models, for instance regression and factor models. Honerkamp (1994) is another noteworthy source for additional information about this continuity of models across the two domains concerned. Hence it appears that with respect to statistical modeling and estimation there are no fundamental differences between analyses of within- and between-subject variation. In the next chapter I will present various forms of evidence that there do exist important differences between analyses of between- and within-subject variation, but these differences have nothing to do with statistical modeling and estimation proper.

Our observations about the continuity of statistical modeling and estimation across the domains of within- and between-subject variation indicate that the prospects of generalizing the TGM and the conjecture about addition of NARMA models (henceforth referred to as the TGM-C) appear to be good. That is, it would appear that the applicability of both the TGM and the TGM-C can be generalized straightforwardly to structural equation models for the analysis of between-subject variation. This is indeed the case, as will be argued below. But as an additional preliminary move it may be helpful to make some additional observations about the functional role of weak stationarity in time series analysis.

Basic time series analysis is based on the concept of weak stationarity (cf. Hannan, 1970). When only a single finite realization of a time series is available, the assumption of weak stationarity provides a rationale for taking averages over time points in statistical estimation. In this way weak stationarity fulfills the same role in time series analysis as the assumption of homogeneity in an analysis of between-subject variation. In the latter kind of analysis, the assumption that subjects are homogeneous in the relevant respects (i.e., the subjects constitute a homogeneous population) provides a rationale for taking averages over subjects in statistical estimation. The concept of weak stationarity has more aspects than just providing a warrant for taking averages over time points in statistical estimation; it is for instance related to the notions of stochastic stability (cf. Tong, 1990, chapter 4) and ergodicity (see next chapter). But here only its licensing role in statistical estimation is considered. It then follows that nonstationarity will invalidate any simple approach to statistical estimation by taking averages over time points. Indeed, analysis of nonstationary time series is a much more delicate affair (cf. Priestley, 1988). When a single finite realization of a nonstationary time series is available, statistical estimation only is possible if the time-dependency of parameters is sufficiently smooth (see the excellent exposition by Dahlhaus, 1997, for further details).

While the latter restriction on the time-dependency of parameters in a NARMA is mandatory in statistical estimation in time series analysis, no such restriction is necessary in the context of structural equation modeling. In longitudinal analysis of between-subject variation, the NARMA(1,0), for instance, is the simplex model encountered before, defined as

*) $a_t[B,1]y_i(t) = y_i(t) +, a_{1,t}\, y_i(t-1) = e_i(t)$, t=2,...,T; i=1,2,...

For convenience the initial condition $y_i(1) = e_i(1)$ has been omitted in *), while the time-varying coefficient $a_{1,t}$ corresponds to the notation of (2.11). In *) no restrictions are imposed on the way in which the coefficient $a_{1,t}$ depends upon time. Still, statistical estimation in this simplex model can proceed without problems because of the assumption of homogeneity underlying the analysis of between-subject variation concerned. It is assumed that *) applies to each subject i, in particular it is assumed that the time-varying coefficient is fixed across subjects i, thus allowing for the possibility to recover any arbitrary sequence of values for $a_{1,t}$, t = 2, ..., T.

In view of the arguments given above, I conjecture that the TGM-C can be generalized straightforwardly to NARMA processes occurring in structural equation models of between-subject variation. This generalized conjecture will be referred to as GC:

Generalized Conjecture GC

Let $x_i(t)$ and $y_i(t)$, for each i=1,2,..., be two weakly orthogonal zero mean NARMA(p,q) and NARMA(m,n) processes, respectively:

$$a_{x,t}[B,p]x_i(t) = b_{x,t}[B,q]e_{xi}(t)$$

$$a_{y,t}[B,m]y_i(t) = b_{y,t}[B,n]e_{yi}(t)$$

where $e_{xi}(t)$ and $e_{yi}(t)$ are white noise series with variance $\sigma_x(t)^2$ and $\sigma_y(t)^2$, respectively. Let $z_i(t)$ be the sum of $x_i(t)$ and $y_i(t)$: $z_i(t) = x_i(t) + y_i(t)$. Then $z_i(t)$ is a zero mean NARMA(r,s) process, where $r \leq p + m$ and $s \leq \max[p + n, q + m]$.

Proof.

As for TGM-C for each i=1,2, ...

2.2.4 Transforming the latent simplex into a manifest simplex

In this section a first application of the GC will be given. It will be applied to the quasi-simplex model, yielding an equivalent representation of this model without latent variables. In a sense to be specified shortly, application of the GC allows for the removal of latent variables in the quasi-simplex model. Because it was shown before that factor models and latent growth curve models are nested under the latent simplex model, it will be evident that application of the GC to the latter models also allows for the removal of latent variables. But in this section we focus attention solely on the quasi-simplex model and defer discussion of factor and latent growth curve models to the next section.

Let us first recall the definition (2.1) of the quasi-simplex model:

$$\begin{aligned} &y_i(t) = \eta_i(t) + \varepsilon_i(t),\ t=1,...,T \\ (2.1)\quad & \\ &\eta_i(t) = \beta_{t,t-1}\eta_i(t\text{-}1) + \varsigma_i(t),\ t=2,...,T;\ \eta_i(1) = \varsigma_i(1) \end{aligned}$$

where $y_i(t)$ is a zero mean univariate manifest variable observed at the fixed time points t=1,2,...,T, $\varepsilon_i(t)$ is zero mean Gaussian white noise measurement error, and $\varsigma_i(t)$ is a zero mean white noise innovation process. It is evident that the latent process $\eta_i(t)$ is a NARMA(1,0):

$$\begin{aligned} &a_{\eta,t}[B,1]\eta_i(t) = \varsigma_i(t),\ t=2,...,T \\ (2.12)\quad & \\ &a_{\eta,t}[B,1] = 1 + a_{\eta,t,1}B = 1 - \beta_{t,t-1}B \end{aligned}$$

where the coefficient $a_{\eta,t,1}$ corresponds to the representation given in (2.11) which is standard in time series analysis, and the coefficient $\beta_{t,t-1}$ corresponds to the representation (2.1) which is standard in structural equation modeling. Such differences in notation have their own logic and should be endured as contingent facts without further implications. Of course, $a_{\eta,t,1} = -\beta_{t,t-1}$.

The first line of (2.1) expresses $y_i(t)$ as the sum of the NARMA(1,0) process $\eta_i(t)$ and the white noise measurement error process $\varepsilon_i(t)$: $y_i(t) = \eta_i(t) + \varepsilon_i(t)$. It is clear that the measurement error $\varepsilon_i(t)$ is a NARMA(0,0) process that can be vacuously represented as a special instance of (2.11):

$$a_{\varepsilon,t}[B,0]\varepsilon_i(t) = b_{\varepsilon,t}[B,0]e_{\varepsilon}(t),\ t=1,2,...,T, \tag{2.13}$$
$$a_{\varepsilon,t}[B,0] = b_{\varepsilon,t}[B,0] = 1$$

Also, under the usual assumptions about structural equation models, $\eta_i(t)$ and $\varepsilon_i(t)$ are weakly orthogonal. Hence the quasi-simplex model consists of the addition of two NARMA processes obeying the requirements of the GC, from which it follows that $y_i(t)$ also is a NARMA process.

$$y_i(t) = \text{NARMA}(1,0) + \text{NARMA}(0,0) = \text{NARMA}(r,s)$$

where $r \leq 1 + 0$ and $s \leq \max[1 + 0, 0 + 0]$. Because there are no common roots, it follows that $y_i(t)$ is a NARMA(1,1) process:

$$a_{y,t}[B,1]y_i(t) = b_{y,t}[B,1]e_{yi}(t) \tag{2.14}$$

Let's pause here and try to evaluate what has been accomplished thus far. We started with the standard quasi-simplex model given by (2.1). In this quasi-simplex model there are at least three different kinds of random variable at each time point t: the manifest variable $y_i(t)$, the latent factor $\eta_i(t)$, and the measurement error $\varepsilon_i(t)$. Perhaps the latent innovation $\varsigma_i(t)$ should be added as a fourth kind of random variable. This standard quasi-simplex model involving these four different kinds of random variable, each with their own interpretation, has been rewritten as a NARMA(1,1) given by (2.14). In this NARMA there are at each time point t only two kinds of random variable: the manifest variable $y_i(t)$ and a new innovation variable $e_{yi}(t)$. It follows from the GC that (2.1) and (2.14) are equivalent. This will be illustrated later on in this section by means of a numerical example. Consequently we have two representations of the same structural equation model: one involving four types of random variable and one involving only two types of random variable. In the transformation from (2.1) to (2.14) which, as I will show shortly, is a one-to-one transformation (and hence invertible), two types of random variable have been removed. The latent factor $\eta_i(t)$ and the latent innovation $\varsigma_i(t)$ no longer occur in (2.14). Loosely speaking, the latent factor and innovation have been removed from the quasi-simplex without affecting its explanatory power. In the next section I will indicate some important implications of this result.

Proceeding with the main line of argument, we now face the task of elaborating the details of the transformation from (2.1) to (2.14). First, a simple numerical example will be given using a weakly stationary quasi-simplex model. The simplicity of this example will convey the details of the transformation concerned as transparently as possible. Then we move on to the elucidation of arbitrary nonstationary quasi-simplex models, again using a numerical example as stepping stone.

Consider the following weakly stationary quasi-simplex model:

$$y_i(t) = \eta_i(t) + \varepsilon_i(t),\ t=1,\ldots,5$$

$$\eta_i(t) = .8\eta_i(t-1) + \varsigma_i(t),\ t=2,\ldots,5;\ \eta_i(1) = \varsigma_i(1)$$

(2.15) $$\mathrm{var}[\varepsilon_i(t)] = 1,\ t=1,\ldots,5$$

$$\mathrm{var}[\varsigma_i(t)] = 1,\ t=2,\ldots,5$$

$$\mathrm{var}[\varsigma_i(1)] = 2.778$$

Both the measurement error $\varepsilon_i(t)$ and the latent innovations $\varsigma_i(t)$ are zero mean white noise processes and hence lack any sequential dependency. The autoregressive coefficient is invariant across time points: $\beta_{t,t-1} = .8$, t=2,...,5. Also the variance of $\varepsilon_i(t)$ and the variance of the latent innovations are invariant in time. The only exception is the variance of the latent innovation at the initial time point. The vacuous equation $\eta_i(1) = \varsigma_i(1)$ implies that $\varsigma_i(1)$ is not a genuine innovation, but is equal to the latent factor at the initial time point. Hence it should have the variance of the latent factor process. As explained in chapter 1, the variance of this weakly stationary first-order autoregressive factor process equals: $\mathrm{var}[\eta_i(t)] = \mathrm{var}[\varsigma_i(t)] / (1 - .8^2) = 2.778$. The true covariance matrix associated with this model is:

$$\begin{array}{c|ccccc|}
 & y(1) & y(2) & y(3) & y(4) & y(5) \\
y(1) & 3.78 & & & & \\
y(2) & 2.22 & 3.78 & & & \\
y(3) & 1.78 & 2.22 & 3.78 & & \\
y(4) & 1.42 & 1.78 & 2.22 & 3.78 & \\
y(5) & 1.14 & 1.42 & 1.78 & 2.22 & 3.78
\end{array}$$

According to the GC, this covariance matrix can also be explained by a weakly stationary ARMA(1,1). There are various ways in which this ARMA model can be represented as a structural equation model. In the final part of this chapter this issue will be discussed in more detail. Presently I will specify one particular implementation as a structural equation model of the weakly stationary ARMA(1,1) for T=5 time points. In the notation of (2.14) we have the following set of equations:

(2.16^a) $$y_i(t) + a_{y,1}y_i(t-1) = e_{y,1}(t) + b_{y,1}e_{y,1}(t-1),\ t=2,\ldots,5$$

$$y_i(1) + a_{y,1}y_i(0) = e_{y,1}(1) + b_{y,1}e_{y,i}(0)$$

As usual in this type of modeling, difficulties arise concerning the handling of the initial condition(s). The equation for t=1 contains $y_i(0)$, which is not available. This can be accommodated in a formal sense by moving the term $a_{y,1}y_i(0)$ to the right-hand side of the equation:

$$y_i(1) = e_{y,1}(1) + b_{y,1}e_{y,i}(0) - a_{y,1}y_i(0) = e_{y,1}(1) + \overset{\$}{b}\overset{\$}{e}_i(0) \tag{2.16^b}$$

where $\overset{\$}{b}\overset{\$}{e}_i(0) = b_{y,1}e_{y,i}(0) - a_{y,1}y_i(0)$. Note that $e_{y,i}(0)$ and $y_i(0)$ are mutually independent under the usual assumptions for ARMA (and structural equation) models.

The equations (2.16^a) and (2.16^b) can be implemented as a regular structural equation model in the following way. Let $\mathbf{y}_i = [y_i(1), ..., y_i(5)]'$ be the 5-dimensional vector of manifest variables. Let $\mathbf{I}_5$ be the (5,5)-dimensional unit matrix and $\mathbf{0}_{5,6}$ the (5,6)-dimensional zero matrix. Then

$$\mathbf{y}_i = \boldsymbol{\Lambda\eta}_i$$

where the (5,11)-dimensional matrix $\boldsymbol{\Lambda}$ is defined as $\boldsymbol{\Lambda} = [\mathbf{I}_5, \mathbf{0}_{5,6}]$. Note that the first five elements of the 11-dimensional vector $\boldsymbol{\eta}_i$ equal $\mathbf{y}_i$ (there is no measurement error ε_i). The remaining six elements of $\boldsymbol{\eta}_i$ are, respectively, $\overset{\$}{e}_i(0)$ and $e_{y,i}(t)$, t=1,...,5:

$$\boldsymbol{\eta}_i = [y_i(1),..., y_i(5), \overset{\$}{e}_i(0), e_{y,i}(1), ..., e_{y,i}(5)]'$$

Hence (2.16^a) and (2.16^b) can be written as the following structural equation model:

$$\begin{aligned}
\eta_{1,i} &= \eta_{7,i} + \beta_{1,6}\eta_{6,i} \\
\eta_{2,i} &= \beta_{2,1}\eta_{1,i} + \eta_{8,i} + \beta_{2,7}\eta_{7,i} \\
\eta_{3,i} &= \beta_{3,2}\eta_{2,i} + \eta_{9,i} + \beta_{3,8}\eta_{8,i} \\
\eta_{4,i} &= \beta_{4,3}\eta_{3,i} + \eta_{10,i} + \beta_{4,9}\eta_{9,i} \\
\eta_{5,i} &= \beta_{5,4}\eta_{4,i} + \eta_{11,i} + \beta_{5,10}\eta_{10,i} \\
\eta_{k,i} &= \varsigma_{k,i},\ k = 6,...,11
\end{aligned} \tag{2.17}$$

where

$$\beta_{2,1} = \beta_{3,2} = \beta_{4,3} = \beta_{5,4}$$

.

$$\beta_{2,7} = \beta_{3,8} = \beta_{4,9} = \beta_{5,10}$$

$\mathrm{var}[\varsigma_{k,i}]$ is invariant over k=6,...,11

Not surprisingly, this structural equation model representing a weakly stationary ARMA(1,1) yields an exact fit to the covariance matrix associated with the weakly stationary quasi-simplex presented above. Using the notation of (2.16^a) and (2.16^b), the parameter values thus recovered are: $a_{y,1}$ = -.8, $b_{y,1}$ = -.34, $b^\$$ = .77, $\mathrm{var}[e_{y,i}(t)]$ = 2.37, t=0,...,5.

How are the parameter values in the weakly stationary ARMA(1,1) related to those in the weakly stationary quasi-simplex? Following Granger & Morris (1976) this can be explained as follows. For the moment neglect initial conditions and consider the first equation in the (weakly stationary instance of the) quasi-simplex model (2.1): $y_i(t) = \eta_i(t) + \varepsilon_i(t)$. Premultiply this equation by $a_\eta[B,1]$ obtained from the weakly stationary case of (2.12), yielding:

$$a_\eta[B,1]y_i(t) = a_\eta[B,1]\eta_i(t) + a_\eta[B,1]\varepsilon_i(t).$$

But (2.12), interpreted in the weakly stationary sense, specifies that $a_\eta[B,1]\eta_i(t) = \varsigma_i(t)$. Consequently:

$$(2.18^a) \qquad a_\eta[B,1]y_i(t) = \varsigma_i(t) + a_\eta[B,1]\varepsilon_i(t).$$

On the other hand, (2.14) specifies that (again neglecting the initial conditions and taking the special case of weak stationarity):

$$(2.18^b) \qquad a_y[B,1]y_i(t) = b_y[B,1]e_{y,i}(t)$$

We now have two representations of what is supposed to be the same process $y_i(t)$, one given by (2.18^a) and one given by (2.18^b) (note that (2.18^b) is another way of writing (2.16^a), used to make the comparison between the two representations of $y_i(t)$ more transparent qua notation). Hence the left-hand sides of (2.18^a) and (2.18^b) should be equal, and also the right-hand sides of these two representations should be equal.

The equality of the left-hand sides of (2.18^a) and (2.18^b) is immediately obtained by taking $a_{\eta,1} = a_{y,1}$. Hence the autoregressive coefficient $a_{\eta,1} = -\beta_{t,t-1}$, t=2,...,5, in the weakly stationary quasi-simplex (2.15) should equal the autoregressive coefficient $a_{y,1} = -\beta_{k,k-1}$, k=2,...,5, in the ARMA(1,1) given by (2.16^a)-(2.17). In the quasi-simplex (2.15) underlying the covariance matrix given above, we took $\beta_{t,t-1}$ = .8, t=2,...,5, while in fitting the ARMA(1,1) given by (2.17) to this covariance matrix we obtained $a_{y,1} = -.8 = -\beta_{k,k-1}$, k=2,...,5. Hence in our numerical example the left-hand sides of (2.18^a) and (2.18^b) indeed agree.

The right-hand sides of (2.18^a) and (2.18^b) cannot be equated directly. Following Granger & Morris (1976), we instead equate the autocovariances up to lag 1 associated with the right-hand sides concerned (hence these right-hand sides are considered in themselves). First we write out the expressions more explicitly:

RHS-(2.18^a): $\varsigma_i(t) + \varepsilon_i(t) + a_{\eta,1}\varepsilon_i(t-1)$

RHS-(2.18^b): $e_{yi}(t) + b_{y,1}e_{y,i}(t-1)$

It is noted that $\varsigma_i(t)$, $\varepsilon_i(t)$ in RHS-(2.18^a) are zero mean white noise processes that are weakly orthogonal. It also is noted that $e_{yi}(t)$ in RHS-(2.18^b) is a zero mean white noise process. Hence the variances, i.e., autocovariances at lag zero, denoted by ACV(0), are:

ACV(0) RHS-(2.18^a):var$[\varsigma_i(t)] + \{1 + (a_{\eta,1})^2\}var[\varepsilon_i(t)]$

ACV(0) RHS-(2.18^b): $\{1 + (b_{y,1})2\}$var$[e_{yi}(t)]$

The autocovariances at lag one, denoted by ACV(1) are:

ACV(1) RHS-(2.18^a):$a_{\eta,1}$var$[\varepsilon_i(t)]$

ACV(1) RHS-(2.18^b): $b_{y,1}$var$[e_{yi}(t)]$

This completes the specification of the mapping between a weakly stationary quasi-simplex and a weakly stationary ARMA(1,1). Let’s apply it to our numerical example. The covariance matrix given above has been generated according to the weakly stationary quasi-simplex (2.15) in which var$[\varepsilon_i(t)] = 1$, $a_{\eta,1} = -.8$, and var$[\varsigma_i(t)] = 1$. We now want to derive from this information the parameter values in the ARMA(1,1) corresponding to (2.15). We therefore consider the parameter values in (2.15) as given. Remember that we already determined from the equality of the left-hand sides of (2.18^a) and (2.18^b) that $a_{y,1} = a_{\eta,1} = -.8$. Substitution of the parameter values of (2.15) in ACV(u) RHS-(2.18^a), u=0,1, and equating the value thus obtained to ACV(u) RHS-(2.18^b), u=0,1, yields:

$2.64 = \{1 + (b_{y,1})2\}var[e_{yi}(t)]$

$$-.80 = b_{y,1}\text{var}[e_{yi}(t)]$$

The two unknowns, $b_{y,1}$ and var[$e_{yi}(t)$], can be obtained from these two equations: $b_{y,1} = -.34$ and var[$e_{yi}(t)$] = 2.37. These are exactly the values for $a_{y,1}$, $b_{y,1}$, and var[$e_{yi}(t)$] obtained in fitting (2.17) to the covariance matrix associated with (2.15).

Of course, one can also proceed in the reverse direction and consider the parameter values in the ARMA(1,1) (2.17) as given. Then the parameter values in the quasi-simplex (2.15) can be determined in a similar vein by using the same mapping rules the other way around. In fact, this is the way in which Granger & Morris (1976) discuss this mapping in order to answer their question: “Can ARMA(1,1) = AR(1) + white noise?”. This is certainly not a trivial question, because the class of weakly stationary ARMA(1,1) models is strictly larger than the class of weakly stationary quasi-simplex models. I will address this important issue and some of its implications later on. For the moment it is noted that our main interest is in transforming quasi-simplex models to (N)ARMA models. Only for this particular subset of (N)ARMA models the inverse transformation is presently of interest.

Having the relationship between the weakly stationary quasi-simplex and the weakly stationary ARMA(1,1) in place, it is a relatively simple exercise to extend this to the specification of the mapping between general (nonstationary) quasi-simplex models and NARMA(1,1) models. In doing so, we will see the workings of the GC in full force. Unfortunately, the intricacies associated with initial conditions will present themselves in the same manner. The basic structure of the set of rules linking the weakly stationary quasi-simplex and ARMA(1,1), however, is not affected and carries over directly to the nonstationary case. Again I will start with a numerical example for the sake of concreteness.

Consider the following (nonstationary) quasi-simplex model:

$$\begin{aligned}
&y_i(t) = \eta_i(t) + \varepsilon_i(t),\ t=1,...,5 \\
&\eta_i(t) =.\beta_{t,t-1}\eta_i(t-1) + \varsigma_i(t),\ t=2,...,5;\ \eta_i(1) = \varsigma_i(1) \\
&\beta_{2,1} = .8,\ \beta_{3,2} = .9,\ \beta_{4,3} = 1.0,\ \beta_{5,4} = 1.1 \\
&\text{var}[\varepsilon_i(1)] = \text{var}[\varepsilon_i(2)] = 1 \\
&\text{var}[\varepsilon_i(3)] = 2 \\
&\text{var}[\varepsilon_i(4)] = \text{var}[\varepsilon_i(5)] = 3 \\
&\text{var}[\varsigma_i(1)] = 5,\ \text{var}[\varsigma_i(2)] = 1,\ \text{var}[\varsigma_i(3)] = 2, \\
&\text{var}[\varsigma_i(4)] = 3,\ \text{var}[\varsigma_i(5)] = 4
\end{aligned} \tag{2.19}$$

It is noted that (2.19) is nonstationary in almost all the respects which were considered in section 2.2.2. The autoregression coefficients $\beta_{t,t-1}$ are time-varying.

Also the variances of $\varepsilon_i(t)$ and $\varsigma_i(t)$ are time-varying. Only the order of the autoregression describing the $\eta_i(t)$ process is constant. The variances of $\varepsilon_i(t)$ at time points t=1,2 are equal: var[$\varepsilon_i(1)$] = var[$\varepsilon_i(2)$] = 1. The same equality restriction has been imposed on the variances of $\varepsilon_i(t)$ at time points t=4,5: var[$\varepsilon_i(4)$] = var[$\varepsilon_i(5)$] = 3. These two equality restrictions do not restrict the generality of (2.19) and could have been omitted. But it is well-known that in the general quasi-simplex model the variances of $\varepsilon_i(t)$ at the initial and final time points are not identifiable. Hence the equality restrictions concerned (or alternative restrictions guaranteeing identifiability) have to be imposed anyway a posteriori, as soon as it comes to model fitting, and then will affect the scaling of the remaining parameter values thus recovered. By using the equality restrictions a priori in generating the true covariance matrix, the parameter values given in (2.19) and those obtained in subsequent model fits will be directly comparable, without the need to rescale.

The true covariance matrix associated with (2.19) is:

$$\begin{array}{c|ccccc|}
 & y(1) & y(2) & y(3) & y(4) & y(5) \\
y(1) & 6.00 & & & & \\
y(2) & 4.00 & 5.20 & & & \\
y(3) & 3.60 & 3.78 & 7.40 & & \\
y(4) & 3.60 & 3.78 & 5.40 & 11.40 & \\
y(5) & 3.96 & 4.16 & 5.94 & 9.24 & 17.17
\end{array}$$

According to the GC, this covariance matrix can also be explained by a NARMA(1,1). I will specify one particular implementation as a structural equation model of this NARMA(1,1) for T=5 time points. In the notation of (2.14) we have the following set of equations for time points t=4 and t=5:

$$(2.20^{a}) \qquad y_i(t) + a_{y.t,1}y_i(t-1) = e_{y,i}(t) + b_{y,,t,1}e_{y,i}(t-1),\ t=4,5$$

The difficulties arising from the handling of initial conditions are more involved for the NARMA(1,1) than for the weakly stationary ARMA(1,1) given by (2.16). For the latter weakly stationary ARMA(1,1) it was possible to utilize the time-invariance of model parameters in order to keep the effects of the initial conditions limited to the first time point t=1. In particular the assumed constancy of var[$e_{y,i}(t)$] in (2.16) made it possible to keep the model equation at time points t=2 and t=3 out of the reach of the initial conditions. For the NARMA(1,1) we have to accept that at time points t=1, t=2 and t=3 the initial conditions complicate the possibility to recover all the parameter values that follow from the GC.

At time point t=1 it is no longer possible in the equation for $y_i(1)$ to distinguish between $e_{y,i}(1)$ and $e_{y,i}(0)$ because the variances of these variables are arbitrary. Consequently, the equation for $y_i(1)$, which according to the NARMA(1,1) would be $y_i(1) + a_{y,1,1}y_i(0) = e_{y,i}(1) + b_{y,1,1}e_{y,i}(0)$, reduces to:

(2.20^{b}) $\qquad y_i(1) = \overset{\$}{e}_i(1)$

At time point t=2 it has to be recognized that in the NARMA(1,1) equation $y_i(2) + a_{y,2,1}y_i(1) = e_{y,i}(2) + b_{y,2,1}e_{y,i}(1)$ the term $b_{y,2,1}e_{y,i}(1)$ is not identified because $e_{y,i}(1)$ is not available (only $\overset{\$}{e}_i(1)$ is available). Because $b_{y,2,1}e_{y,i}(1)$ constitutes part of the description of the relationship between $y_i(1)$ and $y_i(2)$, the fact that this term is not identified also affects the coefficient $a_{y,2,1}$. Hence at t=2 the following reduced equation for $y_i(2)$ is obtained:

(2.20^{c}) $\qquad y_i(2) + \overset{\$}{a}\,y_i(1) = \overset{\$}{e}_i(2)$

Finally at time point t=3, only the coefficient $b_{y,3,1}$ in the NARMA(1,1) equation $y_i(3) + a_{y,3,1}y_i(2) = e_{y,1}(3) + b_{y,3,1}e_{y,i}(2)$ is affected by the initial conditions because $e_{y,i}(2)$ is not available (only $\overset{\$}{e}_i(2)$ is available). This leads to the reduced equation:

(2.20^{d}) $\qquad y_i(3) + a_{y,3,1}y_i(2) = e_{y,i}(3) + \overset{\$}{b}\,\overset{\$}{e}_i(2)$

The equations (2.20^{a})-(2.20^{d}) can be implemented as a regular structural equation model in the following way. Let $\mathbf{y}_i = [y_i(1), ..., y_i(5)]'$ be the 5-dimensional vector of manifest variables. Let $\mathbf{I}_5$ be the (5,5)-dimensional unit matrix and $\mathbf{0}_5$ the (5,5)-dimensional zero matrix. Then

$$\mathbf{y}_i = \boldsymbol{\Lambda\eta}_i$$

where the (5,10)-dimensional matrix $\boldsymbol{\Lambda}$ is defined as $\boldsymbol{\Lambda} = [\mathbf{I}_5, \mathbf{0}_5]$. The 10-dimensional vector $\boldsymbol{\eta}_i$ is defined as:

$$\boldsymbol{\eta}_i = [y_i(1),..., y_i(5), \overset{\$}{e}_i(1), \overset{\$}{e}_i(2), e_{y,i}(3), e_{y,i}(4), e_{y,i}(5)]'$$

The equations (2.20^{a})-(2.20^{d}) can now be expanded as:

$$\eta_{1,i} = \eta_{6,i}$$

$$\eta_{2,i} = \beta_{2,1}\eta_{1,i} + \eta_{7,i}$$

$$\eta_{3,i} = \beta_{3,2}\eta_{2,i} + \eta_{8,i} + \beta_{3,7}\eta_{7,i}$$

(2.21)
$$\eta_{4,i} = \beta_{4,3}\eta_{3,i} + \eta_{9,i} + \beta_{4,8}\eta_{8,i}$$

$$\eta_{5,i} = \beta_{5,4}\eta_{4,i} + \eta_{10,i} + \beta_{5,9}\eta_{9,i}$$

$\eta_{k,i} = \varsigma_{k,i}$, k = 6,...,10

The fit of (2.21) to the covariance matrix associated with the quasi-simplex model (2.19) is exact. It also has the same number of free parameters as the quasi-simplex model. The obtained values of the autoregressive parameter values in both the notation of (2.20^a)-(2.20^d) and (2.21) are: $a^{\$} = -\beta_{2,1} = -.67$, $a_{y,3,1} = -\beta_{3,2} = -.90$, $a_{y,4,1} = -\beta_{4,3} = -1.0$, $a_{y,5,1} = -\beta_{5,4} = -1.1$. The obtained values of the moving-average parameters, given in both notations, are: $b^{\$} = \beta_{3,7} = -.36$, $b_{y,4,1} = \beta_{4,8} = -.45$, $b_{y,5,1} = \beta_{5,9} = -.46$. Finally, the variances of the innovations are, again expressed in both notations: $\mathrm{var}[e^{\$}_{i}(1)] = \mathrm{var}[\varsigma_{6,i}] = 6.0$, $\mathrm{var}[e^{\$}_{i}(2)] = \mathrm{var}[\varsigma_{7,i}] = 2.53$, $\mathrm{var}[e_{y,i}(3)] = \mathrm{var}[\varsigma_{8,i}] = 4.49$, $\mathrm{var}[e_{y,i}(4)] = \mathrm{var}[\varsigma_{9,i}] = 7.11$, $\mathrm{var}[e_{y,i}(5)] = \mathrm{var}[\varsigma_{10,i}] = 9.10$.

We now have obtained two models and their parameter values for the same process $y_i(t)$, t=1,...,5, namely the nonstationary quasi-simplex (2.19) and the NARMA(1,1) given by (2.20^a)-(2.20^d). For those parameters in the NARMA(1,1) that are not affected by the initial conditions, the mapping rules associated with the GC apply. These rules are the same as given before for the weakly stationary ARMA(1,1) model, but should be applied at each time point separately for the NARMA(1,1). I will not go through all the detailed steps again, because these have already been amply discussed, but instead concentrate on the computational aspects.

Neglecting initial conditions and their effects, the two models concerned are the nonstationary quasi-simplex (obtained by premultiplying (2.19), $y_i(t) = \eta_i(t) + \varepsilon_i(t)$, by $a_{\eta,t}[B,1]$ and making use of $a_{\eta,t}[B,1]\eta_i(t) = \varsigma_i(t)$):

(2.22^a) $$a_{\eta,t}[B,1]y_i(t) = \varsigma_i(t) + a_{\eta,t}[B,1]\varepsilon_i(t)$$

and the NARMA(1,1)

(2.22^b) $$a_{y,t}[B,1]y_i(t) = b_{y,t}[B,1]e_{y,i}(t).$$

The autoregressive polynomials $a_{\eta,t}[B,1]$ and $a_{y,t}[B,1]$ at the left-hand sides of (2.22^a) and (2.22^b) should be equal to each other at time points t=3,4,5. This follows from the GC and (2.20^a)-(2.20^d). The autoregressive parameter values at these time points in the nonstationary quasi-simplex (2.19) are: $-\beta_{3,2} = a_{\eta,3,1} = -.9$, $-\beta_{4,3} = a_{\eta,4,1} = -1.0$, $-\beta_{5,4} = a_{\eta,5,1} = -1.1$. The corresponding autoregressive parameter values in the NARMA(1,1), obtained from the fit of (2.21), are indeed the same at the time points concerned: $a_{y,3,1} = -.9$, $a_{y,4,1} = -1.0$, $a_{y,5,1} = -1.1$.

To establish the equality of the moving-average polynomials at right-hand sides of (2.22^a) and (2.22^b), it is helpful to write out the expressions at the time points t=4,5:

RHS-(2.22^a): $\varsigma_i(t) + \varepsilon_i(t) + a_{\eta,t,1}\varepsilon_i(t-1)$, t=4,5

RHS-(2.22^b): $e_{yi}(t) + b_{y,t,1}e_{y,i}(t-1)$, t=4,5

It is noted that $\varsigma_i(t)$, $\varepsilon_i(t)$ in RHS-(2.22^a) are zero mean white noise processes that are weakly orthogonal. Also $e_{yi}(t)$ in RHS-(2.22^b) is a zero mean white noise process. The autovariances at time points t=4,5, denoted by ACV(t,t) are:

ACV(t,t) RHS-(2.22^a): $\text{var}[\varsigma_i(t)] + \text{var}[\varepsilon_i(t)] + (a_{\eta,t,1})^2\text{var}[\varepsilon_i(t-1)]$

ACV(t,t) RHS-(2.22^b): $\text{var}[e_{yi}(t)] + (b_{y,t,1})^2\text{var}[e_{yi}(t-1)]$

The autocovariances between time points t and t-1, t=4,5, are denoted by ACV(t,t-1):

ACV(t,t-1) RHS-(2.22^a): $a_{\eta,t,1}\text{var}[\varepsilon_i(t-1)]$

ACV(t,t-1) RHS-(2.22^b): $b_{y,t,1}\text{var}[e_{yi}(t-1)]$

It is immediately clear that the number of equations for the auto(co-)variances of the right-hand sides of (2.22^a) and (2.22^b) associated with the time points t=4,5 is one less than the number of free parameters in these equations. With respect to the right-hand side of the nonstationary quasi-simplex (2.22^a) at these two time points there are a total of four equations for the auto(co-)variances: ACV(4,3), ACV(4,4), ACV(5,4) and ACV(5,5). These four equations have five free parameters: $\text{var}[\varepsilon_i(3)]$, $\text{var}[\varepsilon_i(4)]$, $\text{var}[\varepsilon_i(5)]$, $\text{var}[\varsigma_i(4)]$ and $\text{var}[\varsigma_i(5)]$. A similar count can be made for the right-hand side of the NARMA(1,1) at time points t=4,5, showing again that there are available four equations for the auto(co-variances) involving five free parameters. The only way to uniquely specify the equalities between the right-hand sides of (2.22^a) and (2.22^b) is to start at the initial time point t=1 and work our way up to the time points of interest, t=4,5 where the GC applies. This is always possible because in the present context of structural equation models for longitudinal analysis, the number of (fixed) time points under consideration is finite. In what follows I will not describe this march from t=1 to t=5 in full generality, but instead sketch the numerical computations associated with mapping the quasi-simplex given by (2.19) and (2.22^a) to the NARMA(1,1) given by (2.20) and (2.22^b). The inverse mapping (from

NARMA to quasi-simplex) should then be evident and is not considered in order to avoid too much repetition and its consequent boredom.

At time points t=1,2 the GC does not apply and therefore no mapping rules between the quasi-simplex and NARMA(1,1) are specified. We can within the context of each model simply march from t=1 onwards without considering the relationships with the other model. Because I want to illustrate the application of the GC with respect to the mapping *from* quasi-simplex *to* NARMA, only the start-up for the NARMA model is presented. For time point t=1 the relevant NARMA(1,1) equation is given by (2.20^b): $y_i(1) = \overset{\$}{e}_i(1)$.

Hence var[$\overset{\$}{e}_i(1)$] = var[$y_i(1)$] = 6. For time point t=2 we have according to (2.20^c): $y_i(2) + \overset{\$}{a} y_i(1) = \overset{\$}{e}_i(2)$. Because cov[$y_i(2)$, $y_i(1)$] = 4 = -$\overset{\$}{a}$var[$y_i(1)$], it follows that $\overset{\$}{a}$ = -4/6 = -.67. We also have that var[$y_i(2)$] = 5.2 = $(-\overset{\$}{a})^2$var[$y_i(1)$] + var[$\overset{\$}{e}_i(2)$], whence var[$\overset{\$}{e}_i(2)$] = 2.53. Until now we have recovered the parameter values as obtained from the fit of (2.21).

At time point t=3 there is the first encounter with the GC as it is applied to the left-hand sides of (2.22^a) and (2.22^b):

*) $$y_i(3) + a_{\eta,3,1}y_i(2) = \varsigma_i(3) + \varepsilon_i(3) + a_{\eta,3,1}\varepsilon_i(2)$$

**) $$y_i(3) + a_{y,3,1}y_i(2) = e_{y,i}(3) + \overset{\$}{b}\overset{\$}{e}_i(2)$$

The GC stipulates that $a_{\eta,\,3,1}$ = -.9 = $a_{y,3,1}$. Because we are still marching within the context of the NARMA, we proceed as before under the condition that $a_{y,3,1}$ = -.9. From **) it follows that cov[$y_i(3)$, $y_i(2)$] = 3.78 = .9var[$y_i(2)$] + $\overset{\$}{b}$var[$\overset{\$}{e}_i(2)$], yielding $\overset{\$}{b}$ = -.36. Finally, from var[$y_i(3)$] = 7.4 = .9^{2}var[$y_i(2)$] + 1.8 $\overset{\$}{b}$var[$\overset{\$}{e}_i(2)$] + var[$e_{y,i}(3)$] + $(\overset{\$}{b})^2$var[$\overset{\$}{e}_i(2)$], yielding var[$\overset{\$}{e}_i(3)$] = 4.49. This latter parameter value, var[$\overset{\$}{e}_i(3)$] = 4.49, provides for the missing bit of information in the execution of the mapping rules associated with the GC. Again, the parameter values recovered up to the present time point t=3 in out march agree exactly with the ones obtained from the fit of (2.21).

We next move to the time point t=4 and apply the mapping rules associated with the GC in exactly the same way as has been described earlier for the relationship between the weakly stationary quasi-simplex and the ARMA(1,1). At t=4 the left-hand sides of (2.22^a) and (2.22^b), describing the autoregressive parts of each model, are equated in the usual way. Equating the right-hand sides at t=4 involves comparison of ACV(4,4) RHS-(2.22^a) with ACV(4,4) RHS-(2.22^b) and comparison of ACV(4,3) RHS-(2.22^a) with ACV(4,3) RHS-(2.22^b). The parametric expressions for these auto(co-)variances have been specified above. As far as the NARMA(1.1) is concerned (remember that we consider the mapping *from* quasi-simplex *to* NARMA), this comparison involves two equations containing two free parameters: var[e_{yi}(4)] and $b_{y,4,1}$. The remaining parameter in the expressions concerned, var[e_{yi}(t-1)], is now known from the previous step in our march. Hence all parameters of the NARMA(1,1) at t=4 can be determined from the parameter values in the quasi-simplex. The latter conclusion immediately generalizes to time point t=5 (and to any additional subsequent time points in general).

At the close of this section I would like to stress that we have accomplished something that to the best of my knowledge has not been achieved before in the broad field of structural equation modeling. It has been shown in detail that the quasi-simplex, i.e., the latent univariate simplex, is equivalent to a NARMA(1,1). The quasi-simplex is composed of a latent factor process and measurement error, while the NARMA(1,1) lacks such a latent factor process. Hence we have rewritten a structural equation model involving common latent variables as a model involving only manifest variables. In this sense, we have "removed the common latent variables from a structural equation model". This accomplishment has interesting implications for ongoing deliberations about "the status of latent random variables in general, and common factors in particular". Some of these implications will be addressed in the next section. Presently it should be recognized that there are also ambiguities associated with the proper denotation of the construct "latent random variable". The factor process η_i(t) occurring in the quasi-simplex model is a latent random process. But strictly speaking the innovations process e_{yi}(t) occurring in the NARMA model also is a latent process. The latter e_{yi}(t) process is akin to the residuals in a standard regression model, and in the context of regular regression analysis it does not appear to be customary to emphasize that the residuals are latent variables. There would seem to be some kind of difference between a common factor and a regression residual, despite that both have to be considered as latent random variables. But as far as I know, a complete and satisfactory specification of this difference (if any) is not yet available.

2.2.5 Removing the factor from a factor model

During the period of my assignment at the Pennsylvania State University, I received one day an honorable invitation to present a lecture at the Statistics Department headed by the eminent C.R. Rao. The lecture was very well attended by members of the department and afterwards most of us, including CR himself, engaged in lively discussion and good food. There was one particular topic of discussion I remember most vividly. It concerned what can be loosely described as "the status of common factors and other latent variables". It was explained to me in various ways that many mathematical statisticians have doubts about the appropriateness of models involving common latent factors. After having listened to the arguments presented to me, I asked who would return home afterwards by car. Many gave an affirmative answer. I then asked whether anybody would not trust the workings of their cars while driving home. There was nobody present who expressed ontological doubts about the robustness of cars. I then indicated that important aspects of the workings of modern cars are controlled by so-called adaptive state controllers, where the states concerned are latent factors. Hence my final, rhetorical, question was, how intelligent scientists could have doubts about the appropriateness of latent variables on the one hand, and yet would trust vital process control based on the same kind of latent variables on the other hand. Maybe a new state of mind was born in the minds in the population of mathematical statisticians, that day in Pennsylvania. A state of mind for which they helped create themselves many formal tools, while being unconscious of some of its implications (barring active repression in some Freudian sense). I felt like a real psychologist amidst mathematicians.

The contents of the present section also have a bearing on issues related with the status of common factors, but the approach is not the dialectical one recounted above. It will be shown that models with latent common factors can be rewritten as models in which these latent factors no longer occur. The tool to accomplish the removal of latent factors from a factor model is again the GC. First a factor model is reformulated as a restricted latent simplex model in the way explained earlier in this chapter. Then the GC is applied to this restricted latent simplex, yielding a NARMA(1,1) that is equivalent to the restricted latent simplex, and hence to the original factor model, but that does no longer contain latent factors. This preview may indicate that the discussion in the present section does not introduce any new formalities, but instead involves a further application of the results obtained in earlier sections. The factor model deserves special treatment because, for better for worse, it has an intimate and long-standing relationship with psychology. It will be shown, to the best of my knowledge for the first time, how this factor model can be stripped of its essential ingredients and turned into a model involving only manifest variables and residuals.

We start again with a numerical example and for this the 1-factor model considered in section 2.1.2 is taken. Let $\mathbf{y_i} = [y_{1i}, y_{2i}, y_{3i}, y_{4i}]'$ be the 4-dimensional vector of manifest variables and consider the 1-factor model

$$\mathbf{y}_i = \boldsymbol{\lambda}\eta_i + \boldsymbol{\varepsilon}_i$$

$$\boldsymbol{\lambda}\text{ ‘} = [1, 1, 2, 1]$$

a)

$$\mathrm{cov}[\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_i'] = \mathrm{diag}[1, 2, 3, 4]$$

$$\mathrm{var}[\eta_i] = 1$$

The true covariance matrix associated with model a) has been given in section 2.1.2 and is repeated below:

$$\begin{array}{c c} & \begin{array}{cccc} y_1 & y_2 & y_3 & y_4 \end{array} \\ \begin{array}{c} y_1 \\ y_2 \\ y_3 \\ y_4 \end{array} & \left[\begin{array}{cccc} 2 & & & \\ 1 & 3 & & \\ 2 & 2 & 7 & \\ 1 & 1 & 2 & 5 \end{array}\right] \end{array}$$

It also has been explained in section 2.1.2 that this 1-factor model is equivalent to the following restricted quasi-simplex model:

$y_i(t) = \eta_i(t) + \varepsilon_i(t)$, t=1,...,4

var[$\eta\_$] = 1

b)

$\eta_i(t) = \beta_{t,t-1}\eta_i(t-1)$, t=2,3,4

$\beta_{21} = 1$, $\beta_{32} = 2$, $\beta_{43} = .5$

Neglecting initial conditions and their effects (hence for t larger than τ, where the minimum value of τ will be determined shortly), the latter restricted quasi-simplex b) can (after premultiplication by $a_{\eta,t}[B,1]$, etc.) be written as:

*) $a_{\eta,t}[B,1]y_i(t) = a_{\eta,t}[B,1]\varepsilon_i(t)$, $t > \tau$,

where $a_{\eta,t}[B,1] = (1 - \beta_{t,t-1}B)$. According to the GC, the nonstationary quasi-simplex *) is equivalent to the NARMA(1,1)

**) $a_{y,t}[B,1]y_i(t) = b_{y,t}[B,1]e_{y,i}(t)$, $t > \tau$

Note that *) differs from (2.22^a) in the previous section in one important respect: there is no longer a latent innovations process $\varsigma_i(t)$ occurring at the right-hand side of *). This implies that *) and **) can be equated simply by taking $a_{\eta,t}[B,1] = a_{y,t}[B,1]$ and $a_{\eta,t}[B,1] = b_{y,t}[B,1]$. Hence in **) the autoregressive polynomial $a_{y,t}[B,1]$ is equal to the moving-average polynomial $b_{y,t}[B,1]$, and both polynomials are equal to $a_{\eta,t}[B,1]$ in *). It may look as if *) could be reduced to $y_i(t) = \varepsilon_i(t)$ by dividing out the common polynomial, and **) could be similarly reduced to $y_i(t) = e_{y,i}(t)$. This is not allowed, however, due to the effects of initial conditions on the restricted quasi-simplex and the NARMA(1,1). As will be explained below, these differences in initial conditions are inherited at later time points, making the reductions under consideration invalid.

But let us first pause for a moment and consider *) and **). It then is noted that *both* have exactly the same NARMA(1,1) structure: for $t > \tau$ the expression *) for the restricted quasi-simplex model is, apart from irrelevant notational differences, exactly the same as the expression **) for the NARMA(1,1). Both *) and **) simply *are* the same NARMA(1,1) expressions for $t > \tau$. Consequently it is expected that the $e_{y,i}(t)$ process in the NARMA(1,1) **) will be the same (in some appropriate stochastic sense) as the measurement error process $\varepsilon_i(t)$ in the restricted quasi-simplex *). It also is noted that **) has the same form as (2.22^b) considered in the previous section in the context of rewriting the general quasi-simplex as NARMA(1,1). It was shown in that section that the effect of initial conditions on the moving-average polynomial at the right-hand side of (2.22^b) were still noticeable at time point t=3. Hence $\tau = 3$ with respect to the right-hand side of (2.22^b).

With these preliminary observations in hand, we proceed by writing out the NARMA(1,1) associated with the restricted quasi-simplex for each time point t=1,...,4. To start with, the steps in the discussion following (2.22^{b}) (where $\tau = 3$ for the right-hand side of the NARMA associated with the general quasi-simplex) are repeated, yielding

$$y_i(1) = e^{\$}{}_i(1)$$

$$y_i(2) + a_{y,2,1}y_i(1) = e^{\$}{}_i(2)$$

$$y_i(3) + a_{y,3,1}y_i(2) = e_{y,i}(3) + b^{\$}e^{\$}{}_i(2)$$

$$y_i(4) + a_{y,4,1}y_i(3) = e_{y,i}(4) + a_{y,4,1}e_{y,i}(3)$$

For time point t=4 the restriction that the autoregressive polynomial in *B* equals the moving-average polynomial in *B* is obeyed. For t=3 this restriction has not yet been expressed in the right-hand side. The problem is that only $e^{\$}{}_i(2)$ is available from the antecedent time point t=2, not $e_{y,i}(2)$ which is needed to express the restriction. Turning our attention to t=2, suppose that we substitute $e^{\$}{}_i(2) = e_{y,i}(2) + e^{\#}{}_i(2)$, yielding $y_i(2) + a^{\$}y_i(1) = e_{y,i}(2) + e^{\#}{}_i(2)$, where $a^{\$}$ has to be substituted for $a_{y,2,1}$ because of the addition of $e_{y,i}(2)$. It then no longer holds that $a_{y,2,1}$ in **) equals $a_{\eta,2,1}$ in *). However, it now is possible to express the restriction about the equality of autoregressive and moving-average polynomial at t=3: $y_i(3) + a_{y,3,1}y_i(2) = e_{y,i}(3) + a_{y,3,1}e_{y,i}(2)$. The complete set of equations for the NARMA(1,1) equivalent of the restricted quasi-simplex then becomes:

$$y_i(1) = e^{\$}{}_i(1)$$

$$y_i(2) + a^{\$}y_i(1) = e_{y,i}(2) + e^{\#}{}_i(2)$$

c)

$$y_i(3) + a_{y,3,1}y_i(2) = e_{y,i}(3) + a_{y,3,1}e_{y,i}(2)$$

$$y_i(4) + a_{y,4,1}y_i(3) = e_{y,i}(4) + a_{y,4,1}e_{y,i}(3)$$

Of course, the artificial residual variable $e^{\#}{}_i(2)$ is uncorrelated with the remaining residual variables. In particular, $\text{cov}[e^{\#}{}_i(2), e_{y,i}(2)] = 0$. Note that we have thus reduced the value of τ for the right-hand side of **) from $\tau = 3$ to $\tau = 2$. This procedure to reduce τ to $\tau = 2$ always works, irrespective of the total number T of time points at which the analysis is carried out.

The fit of c) to the true covariance matrix associated with the 1-factor model a) is exact. Model c), which is the NARMA(1,1)-equivalent of the 1-factor model, has the same number of free parameters as the 1-factor model a) and the restricted quasi-simplex b). The parameter values obtained in the fit of c) are: $a^{\$} = -.50$, $a_{y,3,1} = -2.0$, $a_{y,2,1} = -.50$, var[$e^{\$}_{i}(1)$] = 2.0, var[$e^{\#}_{i}(2)$] = .50, var[$e_{y,i}(2)$] = 2.0, var[$e_{y,i}(3)$] = 3.0, and var[$e_{y,i}(4)$] = 4.0.

It is noted that in model c), for t > 2, the autoregressive polynomials in the left-hand side of the NARMA(1,1) are equal to moving-average polynomials in the right-hand side. The coefficients in these polynomials are equal to the autoregressive coefficients in the restricted quasi-simplex equivalent of the 1-factor model described in section 2.1.2. Furthermore, it is noted that for t > 1 the residual process $e_{y,i}(t)$ in this NARMA is in a distributional sense the same as the measurement error process $\varepsilon_i(t)$, i.e., both processes have the same 3-variate Gaussian distribution. Consequently, **) can be expressed more specifically as:

$$y_i(1) = e^{\$}_{i}(1)$$

(2.23)
$$y_i(2) + a^{\$} y_i(1) = \varepsilon_i(2) + e^{\#}_{i}(2)$$

$$a_{y,t}[B,1]y_i(t) = a_{y,t}[B,1]\varepsilon_i(t),\ t > 2, ..., T$$

It therefore can be concluded that for arbitrary T the 1-factor model can be rewritten as a NARMA(1,1) expressed as (2.23). The common factor η_i no longer occurs in (2.23), only the manifest variables $y_i(t)$ and (for t > 2) the measurement errors $\varepsilon_i(t)$. In addition two residuals occur at t=1 and t=2, respectively. The common factor η_i has been removed from the 1-factor model by expressing it in the form (2.23).

There remains one final point that has to be worked out yet. Namely the handling of initial conditions within the present context. For the restricted NARMA we have:

$$y_i(1) = e^{\$}_{i}(1)$$

$$y_i(2) + a^{\$} y_i(1) = \varepsilon_i(2) + e^{\#}_{i}(2)$$

whereas for the restricted quasi-simplex we have:

$$y_i(1) = \eta_1 + \varepsilon_i(1)$$

$$y_i(2) + a_{\eta,1,1}\eta_1 = \varepsilon_i(2)$$

These two sets of equations should yield expressions for variances and covariance of

$y_i(t)$, t=1,2, which have equal values. At t=1 this implies var[$y_i(1)$] = var[$e^{\$}_i(1)$], which equals var[$y_i(1)$] = var[$\eta_1$] + var[$\varepsilon_i(1)$]. Hence

i) $\quad \mathrm{var}[e^{\$}_i(1)] = \mathrm{var}[\eta_1] + \mathrm{var}[\varepsilon_i(1)]$

From cov[$y_i(1)$, $y_i(2)$] = $(-a^{\$})^2$var[$y_i(1)$] and cov[$y_i(1)$, $y_i(2)$] = $(-a_{\eta,1,1})^2$var[η_1] it follows that

j) $\quad a^{\$} = a_{\eta,1,1}\mathrm{var}[\eta_1] / \mathrm{var}[y_i(1)]$

Note that the ratio ρ = var[η_1] / var[$y_i(1)$] defines the reliability in classical test theory and abroad. Finally, from var[$y_i(1)$] = $(-a^{\$})^2$var[$y_i(1)$] + var[$\varepsilon_i(2)$] + var[$e^{\#}_i(2)$] and var[$y_i(1)$] = $(-a_{\eta,1,1})^2$var[η_1] + var[$\varepsilon_i(2)$] it follows that

k) $\quad \mathrm{var}[e^{\#}_i(2)] = (-a_{\eta,1,1})^2\mathrm{var}[\eta_1]\{1 - \rho\}$

Using i), j) and k) one can compute the parameters in the restricted NARMA from those in the restricted quasi-simplex. Hence these equations define the mapping *from* the restricted quasi-simplex *to* the restricted NARMA. To obtain the inverse mapping from restricted NARMA to restricted quasi-simplex, one needs to start at the final time T and work one’s way back to the initial time point. I will not describe the rather boring details.

The conclusion that for arbitrary T the 1-factor model can be rewritten as a NARMA(1,1) has been formulated for a T-technique longitudinal factor model involving repeated measurements of a univariate manifest variable $y_i(t)$, t=1,...,T. As has been remarked at the close of section 2.1.2, nothing special hinges on the interpretation of the manifest variables in a 1-factor model in terms of repeated measurements. Hence it follows immediately that any 1-factor model for a p-variate vector-valued manifest variable $\mathbf{y}_i$ can be rewritten as a NARMA(1,1) expressed as (2.23), where the index t in the latter expression now runs over the elements of $\mathbf{y}_i$: t=1,2,...,p. This means that the 1-factor model is equivalent to a model involving linear combinations of the manifest variables and the measurement errors. This result has a variety of implications, of which I will mention a few, but first we collect our fruits in a fancy basket.

I propose to refer to the transformation removing the common factor from a 1-factor model as the Houdini transformation. The disappearance of the common factor itself is not a very interesting accomplishment (in fact, each of us will disappear from the surface of the earth at some future time). But it is the fact that the disappearance can be undone which is remarkable. The transformation back from (2.23) to model b) and then back to model a) always is possible. This turns the transformation into a genuine analogue of acts of Houdini. In the end, the Houdini transformation turns out to be exceedingly simple. As the late David Fulker used to say: "It can be computed on the back of an envelope". Specifically:

The Houdini transformation for 1-factor models

- Given a 1-factor model, denote the vector of factor loadings by $\boldsymbol{\lambda}' = [\lambda_1, \lambda_2, \dots, \lambda_p]$. Compute $\beta_{k,k-1} = \lambda_k / \lambda_{k-1}$, k=2,…,p. These are the beta-coefficients in the restricted quasi-simplex associated with this 1-factor model.
- Determine the initial conditions in (2.23) according to i), j) and k).
- Given the beta-coefficients obtained from the first step, construct $y_{k,i} - \beta_{k,k-1}y_{k-1,i} = \varepsilon_{k,i} - \beta_{k,k-1}\varepsilon_{k-1,i}$, k=3,…,p.

This definition of the Houdini transformation has to be amended with special rules for k=1,2. I leave this to the reader to elaborate on the basis of what has been said before about the handling of initial conditions. It appears that for k > 2 the Houdini transformation involves the taking of a simple linear combination of the k-th and (k-1)-th manifest variables, and equating this to the same linear combination of the k-th and (k-1)-th measurement errors.

To mention, at the close of this section, some possible implications of all this, I first would like to refer back to the introduction of this section. The fact that a common latent factor can be removed from a model, without affecting the goodness of fit of this model, says something about the status of such a latent factor. In particular, it implies something about the relationship of this latent factor with the manifest variables. Presently, I will not elaborate what this something might be, but this issue is addressed somewhat further in the final part of this chapter where identifiability of state-space models will be considered from a general perspective. Another, perhaps related, point is what the implications can be of the Houdini transformation with respect to the indeterminacy of factor scores (Krijnen, 1999). Is this indeterminacy still present in some ways in the NARMA(1,1) equivalent of the 1-factor model? And, as a final example of possible implications of the Houdini transformation, consider a 1-factor model in which the factor loadings are equal: $\lambda_1 = \dots = \lambda_p$. Suppose also that $\text{var}[\varepsilon_{1,i}] = \dots = \text{var}[\varepsilon_{p,i}]$. In that case the elements of $\mathbf{y}_i$ constitute parallel measurements of the same construct (Lord & Novick, 1968). It then follows that the NARMA(1,1) which according to the Houdini transformation is equivalent to this parallel measurement factor model is: $y_{k,i} - y_{k-1,i} = \varepsilon_{k,i} - \varepsilon_{k-1,i}$, k=3,…,p, where the right-hand side now has stationary “auto”-covariance function of the index k. For the initial conditions, letting $\rho = \text{var}[\eta_1] / \text{var}[y_i(1)]$ and remembering that the expression for t=1 is vacuous, it holds that $y_i(2) - \rho y_i(1) = \varepsilon_i(2) + e^{\#}{}_i(2)$, where $\text{var}[e^{\#}{}_i(2)] = \text{var}[\eta_1]\{1 - \rho\}$. Similar expressions can be derived for tau-equivalent and congeneric measurements. These may cast a new perspective on the concept of reliability in classical test theory. I consider these implications, and several other ones not mentioned, as possibly interesting topics for future research.

2.2.6 Addition of multiple latent factors yields a univariate latent simplex

There are some important properties of the GC which until now have been left implicit. One of these concerns the applicability of the GC to a denumerable set of NARMA processes. Application of the GC to a pair of NARMA processes taken from this set yields again a NARMA process. If to the NARMA process thus obtained a third NARMA process taken from the set is added, this yields again a NARMA process. And so forth and so on. This shows that the GC stipulates that the operation of addition of NARMA processes has a closure property in the algebraic sense. One could also say that the GC provides for a connection with asymptotic theory, in particular the (central) limit theorems. The latter connection with asymptotic theory would seem to raise new questions. For instance, let NARMA(r_n,s_n) denote the result of summing n NARMA processes, each of arbitrary order r_i, s_i, i=1,...,n. Then the GC implies that r_n and s_n are nondecreasing functions of n. What kind of asymptotic theory (if any) is obtained if n increases indefinitely? Despite the intrinsic interest of this question, we do not have to address it because only finite sums of NARMA processes will be considered in what follows.

A second property of the GC concerns the assumption that the pair of NARMA processes which are added together are weakly orthogonal, i.e., their cross-covariance function is zero. This assumption can be dropped, as is already suggested by Granger & Morris (1976) with respect to the addition of ARMA processes. Of course, allowing for n NARMA processes to be correlated will complicate the asymptotic theory about their sum even further. The details of dropping the assumption concerned will not be worked out here, however, because we will only consider finite sums of weakly orthogonal NARMA processes. I hope to elaborate generalization of the GC to the addition of finite sums of correlated NARMA processes in the near future.

The reason why these topics are brought up at the beginning of the present section is the following. Consider a model having multiple common factors, for instance an orthogonal 2-factor model. Then, as shown in section 2.1.3, this model can be rewritten as a restricted latent bivariate simplex model. At the latent level the latter model consists of two weakly orthogonal restricted NARMA(1,0) processes to which the GC can be applied. This yields a univariate NARMA(2,1) at the latent level. Finally, the GC can be applied again with respect to the addition of this NARMA(2,1) and the NARMA(0,0) measurement error process, yielding a NARMA(2,2). Here we have two applications of the GC, one to add the two reatricted NARMA(1,0) processes representing the two common factors, and another one to add the univariate NARMA(2,1) sum process thus obtained and the measurement error process.

In what follows I will only detail the addition of latent variables associated with the communal part of models, in particular addition of a pair of orthogonal common factors. Hence subsequent application of the GC to measurement errors, as was explained in the previous section, will no longer be considered (but can be carried out, of course). Addition of a pair of common factors, addition of a pair consisting of a common factor and a latent simplex, and addition of a pair of latent simplexes, each yield a latent univariate NARMA(2,1) and in that respect behave the same under application of the GC. Hence proceeding with the second and final step in the complete Houdini transformation of these models, i.e., recursive application of the GC to the addition of NARMA(2,1) and NARMA(0,0) measurement error, would in each case yield manifest NARMA(2,2) models. But to reiterate, this final step will not be considered in the present section in an attempt to keep the text within manageable proportions.

We consider the addition of two orthogonal common factors, using the following numerical illustration:

$$\mathbf{y}_i = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i$$

$$\mathbf{y}_i = [y_{1,i}, ..., y_{5,i}]', \boldsymbol{\eta}_i = [\eta_{1,i}, \eta_{2,i}]', \boldsymbol{\varepsilon}_i = [\varepsilon_{1,i}, ..., \varepsilon_{5,i}]'$$

$$\boldsymbol{\Lambda}' = \begin{bmatrix} 1 & 2 & 3 & 2 & 1 \\ 0 & 3 & 1 & 1 & 3 \end{bmatrix}$$

$$\text{cov}[\boldsymbol{\varepsilon}_i. \boldsymbol{\varepsilon}_i'] = \text{diag}[1, 2, 3, 4, 5]$$

$$\text{cov}[\boldsymbol{\eta}_i, \boldsymbol{\eta}_i'] = \mathbf{I}_2$$

This is an exploratory orthogonal 2-factor model for a 5-variate manifest variable $\mathbf{y}_i$. Note that $\lambda_{12} = 0$, hence the model already includes the minimum identifiability constraint. This constraint could have been omitted. But it is inconsequential, its only convenient effect is that the parameter values given above are not affected by scaling in subsequent model fits (see the analogous comment below (2.19) in section 2.2.4). The true covariance matrix associated with this orthogonal 2-factor model is:

$$\begin{array}{c|ccccc} & y_1 & y_2 & y_3 & y_4 & y_5 \\ \hline y_1 & 2 & & & & \\ y_2 & 2 & 15 & & & \\ y_3 & 3 & 9 & 13 & & \\ y_4 & 2 & 7 & 7 & 9 & \\ y_5 & 1 & 11 & 6 & 5 & 15 \end{array}$$

Following the approach described in section 2.1.3, the following instance of the restricted latent bivariate simplex (2.5) also yields an exact fit to this covariance matrix:

$$\eta_{1,i}(t) = \beta_{t,t-1}\eta_{1,i}(t-1), t=2,...,5; \eta_{1,i}(1) = \varsigma_{1,i}(1)$$

$$\eta_{2,i}(t) = \delta_{t,t-1}\eta_{2,i}(t-1), t=3,...,5; \eta_{2,i}(2) = \varsigma_{2,i}(2)$$

$$y_i(1) = \eta_{1,i}(1) + \varepsilon_i(1)$$

$$y_i(t) = \eta_{1,i}(t) + \eta_{2,i}(t) + \varepsilon_i(t), t=2,...,5$$

$$\beta_{2,1} = 2.00, \beta_{3,2} = 1.50, \beta_{4,3} = .67, \beta_{5,4} = .50$$

$$\delta_{3,2} = .33, \delta_{4,3} = 1.00, \delta_{5,4} = 3.00$$

$$\text{var}[\varsigma_{1,i}(1)] = 1.00, \text{var}[\varsigma_{2,i}(2)] = 9.00, \text{cov}[\varsigma_{1,i}(1), \varsigma_{2,i}(2)] = 0$$

$$\text{cov}[\boldsymbol{\varepsilon}_i. \boldsymbol{\varepsilon}_i'] = \text{diag}[1, 2, 3, 4, 5]$$

The restricted latent bivariate simplex consists of the univariate components $\eta_{1,i}(t)$, t=1,...,5, and $\eta_{2,i}(t)$, t=2,...,5, where $\eta_{1,i}(t)$ and $\eta_{2,i}(t)$ are weakly orthogonal NARMA(1,0) processes. Hence the GC can be applied to these processes, yielding a latent univariate NARMA(2,1). To see how this works in the present example, it is convenient to shift to the notation in terms of polynomials in the backshift operator *B* used in presenting the GC.

Define $\beta_t[B,1] = 1 - \beta_{t,t-1}B$ and $\delta_t[B,1] = 1 - \delta_{t,t-1}B$. Then the first and second univariate component simplex can be represented as, respectively:

$$\eta_{1,i}(1) = \varsigma_{1,i}(1)$$

$$\beta_t[B,1]\eta_{1,i}(t) = 0,\ t=2,...,5$$

$$\eta_{2,i}(1) = 0,\ \eta_{2,i}(2) = \varsigma_{2,i}(2)$$

$$\delta_t[B,1]\eta_{2,i}(t) = 0,\ t=3,4,5$$

We now consider the sum $\xi_i(t) = \eta_{1,i}(t) + \eta_{2,i}(t)$, t=1,...,5, which according to the GC is a NARMA(2,1). If initial conditions could be neglected, it would follow that

$$\beta_t[B,1]\delta_t[B,1]\xi_i(t) = 0,\ t > 2$$

implying that for t > 2 the latent univariate $\xi_i(t)$ process obeys an autonomous second-order difference equation. A difference equation is autonomous in case it lacks external perturbations such as the random innovations $\varsigma_i(t)$. Unfortunately, as will be seen shortly, the initial conditions destroy the simple pattern of the coefficients in this second-order difference equation for $\xi_i(t)$, t > 2.

The restricted latent NARMA(2,1) can be fitted to the covariance matrix associated with the orthogonal 2-factor model by means of the following structural equation model:

$$\xi_i(1) = \nu_i(1)$$

$$\xi_i(2) = \gamma_{2,1}\xi_i(1) + \nu_i(2)$$

$$\xi_i(t) = \gamma_{t,t-1}\xi_i(t-1) + \gamma_{t,t-2}\xi_i(t-2),\ t=3,4,5$$

$$y_i(t) = \xi_i(t) + \varepsilon_i(t),\ t=1,...,5$$

where $\text{cov}[\nu_i(1), \nu_i(2)] = 0$ and the measurement errors $\varepsilon_i(t)$ are defined in the same way as in the orthogonal 2-factor model. This restricted NARMA(2,1) model yields an exact fit to the covariance matrix given above and has the same number of free parameters as the orthogonal 2-factor model used to generate this covariance matrix. The parameter values thus obtained are: $\gamma_{2,1} = 2.0$, $\gamma_{3,1} = 2.33$, $\gamma_{3,2} = .33$, $\gamma_{4,2} = .14$, $\gamma_{4,3} =$

.57, $\gamma_{5,3} = -5.0$, $\gamma_{5,4} = 8.0$, var[$\nu_i(1)$] = 1.0, var[$\nu_i(2)$] = 9.0, cov[$\boldsymbol{\varepsilon}_i$. $\boldsymbol{\varepsilon}_i$'] = diag[1, 2, 3, 4, 5].

I now will explain in some detail how the parameter values in the restricted latent univariate NARMA(2,1) can be derived from the parameter values in the restricted latent bivariate simplex. It has already been demonstrated in previous sections how the parameter values in a restricted latent bivariate simplex are derived from the parameter values in an orthogonal 2-factor model, hence this step will not be considered again here. Suffice it to say that by deriving the parameter values in the restricted latent NARMA from those in the restricted latent bivariate simplex, the three sets of parameter values in the three equivalent models under consideration (orthogonal 2-factor, restricted latent bivariate simplex, restricted latent univariate NARMA) have one-to-one pairwise relationships. Given the values of one set of parameters, the values in the remaining two sets of parameters can be derived exactly. The derivation I will give of the parameter values in the restricted latent NARMA from those in the restricted latent bivariate simplex is tedious and lacks elegance. Because much of the material presented in this book is rather new and has been elaborated in the process of writing, it may turn out that the derivation below can, with some additional effort, be improved considerably in terms of elegance. The reader who is willing to accept the one-to-one correspondence between the parameter values of the restricted latent bivariate simplex and the restricted latent NARMA can skip the rest of this section without harm.

We are going to specify the mapping *from* the restricted latent bivariate simplex *to* the restricted latent univariate NARMA. Only the mapping from the β- and δ-coefficients to the γ-coefficients have to be specified, because the mapping of the variances of the latent innovations and the measurement error variances is trivial. At time point t = 1 it simply holds that

$$\xi_i(1) \quad = \eta_{1,i}(1) = \nu_i(1)$$

At time point t = 2 we have:

$$\begin{aligned}\xi_i(2) \quad &= \eta_{1,i}(2) + \eta_{2,i}(2)\\ &= \beta_{2,1}\eta_{1,i}(1) + \eta_{2,i}(2)\\ &= \beta_{2,1}\xi_i(1) + \nu_i(2)\end{aligned}$$

Hence $\gamma_{2,1} = \beta_{2,1} = 2.0$. At time point t = 3 the following derivation can be given:

$$\begin{aligned}\xi_i(3) \quad &= \eta_{1,i}(3) + \eta_{2,i}(3)\\ &= \beta_{3,2}\eta_{1,i}(2) + \delta_{3,2}\eta_{2,i}(2)\\ &= \beta_{3,2}\beta_{2,1}\xi_i(1) + \delta_{3,2}[\xi_i(2) - \beta_{2,1}\xi_i(1)]\\ &= \beta_{2,1}(\beta_{3,2} - \delta_{3,2})\xi_i(1) + \delta_{3,2}\xi_i(2)\end{aligned}$$

Hence $\gamma_{3,1} = \beta_{2,1}(\beta_{3,2} - \delta_{3,2}) = 2.33$ and $\gamma_{3,2} = \delta_{3,2} = .33$

Proceeding to time point t = 4, the computations become increasingly messy. We have

$$\xi_i(4) \quad = \eta_{1,i}(4) + \eta_{2,i}(4)$$

The first term at the right-hand side can be rewritten as

$$\eta_{1,i}(4) = \beta_{4,3}\beta_{3,2}\beta_{2,1}\xi_i(1)$$

$$= \beta_{4,3}\beta_{3,2}\beta_{2,1}[\xi_i(3) - \delta_{3,2}\xi_i(2)][(\beta_{2,1}(\beta_{3,2} - \delta_{3,2})]^{-1}$$

$$= \beta_{4,3}\beta_{3,2}\beta_{2,1}[\xi_i(3) - \delta_{3,2}\xi_i(2)]Q$$

where $Q = [(\beta_{2,1}(\beta_{3,2} - \delta_{3,2})]^{-1}$ is used to ease the notation. The second term on the right-hand side is rewritten as

$$\eta_{2,i}(4) = \delta_{4,3}\delta_{3,2}\eta_{2,i}(2)$$

$$= \delta_{4,3}\delta_{3,2}[\xi_i(2) - \beta_{2,1}\xi_i(1)]$$

$$= \delta_{4,3}\delta_{3,2}[\xi_i(2) - \beta_{2,1}[\xi_i(3) - \delta_{3,2}\xi_i(2)]Q]$$

Collecting terms yields

$$\xi_i(4) = \gamma_{4,3}\xi_i(3) + \gamma_{4,2}\xi_i(2)$$

$$\gamma_{4,3} \quad = [\beta_{4,3}\beta_{3,2}\beta_{2,1} - \delta_{4,3}\delta_{3,2}\beta_{2,1}][(\beta_{2,1}(\beta_{3,2} - \delta_{3,2})]^{-1}$$

$$\gamma_{4,2} \quad = [\beta_{2,1}\delta_{4,3}(\delta_{3,2})^2 - \beta_{4,3}\beta_{3,2}\beta_{2,1}\delta_{3,2}][(\beta_{2,1}(\beta_{3,2} - \delta_{3,2})]^{-1} + \delta_{4,3}\delta_{3,2}$$

Using the values of the β- and δ-coefficients we then get $\gamma_{4,2} = .14$ and $\gamma_{4,3} = .57$.

I think this should be sufficient to convey the pattern of derivation of the parameter values in the restricted latent univariate NARMA from the parameter values in the restricted latent bivariate simplex. The derivation for the final time point t = 5 does not involve anything new and therefore is left as an exercise. We have obtained the remarkable result that an orthogonal 2-factor model, i.e., a model having a 2-dimensional common factor space, has been rewritten as a model involving a 1-dimensional latent process described by a restricted NARMA(2,1). This result can be generalized straightforwardly: an orthogonal q-factor model, q > 0, can be rewritten as a model involving a univariate latent restricted NARMA(q,q-1). Of course, by adding the measurement error in the final step, each orthogonal q-factor model can be Houdini transformed into a NARMA(q,q).

Before leaving the topic of detailed applications of the GC in the context of structural equation models, I would like to address the addition of a latent simplex and a common factor. One could interpret the latent simplex as describing a state-like process, the predictability of which across consecutive time points is less than perfect.

In contrast, the common factor then would describe a trait-like process, the predictability of which at the latent level is perfect (see discussion in section 2.1.6). Such interpretations have been given, for instance, to the latent simplex and common factor components of the longitudinal quantitative genetical model of Hewitt, Eaves, Neale, & Meyer (1988). In 1987 Dorret Boomsma and I had introduced the quantitative genetical quasi-simplex as a convenient model for longitudinal phenotypic data (Boomsma & Molenaar, 1987). Shortly after the appearance of the Hewitt et al. paper I met the authors at a conference in Boulder, where I showed them that the latent simplex and common factor components in their model can be added, yielding a univariate latent NARMA(2,1) process. In my opinion this transformation proved that there was no qualitative difference between their model and ours, only a difference in degree (or more specifically, a difference in the order of the underlying NARMA process). I am not sure whether the authors shared my point of view. Anyway, this little history shows that the addition of a latent simplex and a common factor would appear to be the earliest instance of the use of the GC. The reader is referred to Rovine & Molenaar (2001) for further details.

2.2.7 Discussion of the GC and some of its implications

In this closing section on the presentation of techniques to manipulate latent variables, it may be worthwhile to qualify the obtained results in various ways. First and foremost, it has to be reiterated that the techniques concerned are new, at least as far as I know. I never encountered an equivalent of the GC in the published literature, nor am I aware of publications on the systematic addition of latent variables. New results are like young persons: they are beautiful and promising, but possible flaws only become manifest with sufficient maturity. The techniques to manipulate latent variables are immature in almost every respect and may inhere several latent flaws. It is up to the working psychometricians and structural equation modelers to decide what will remain of the whole exercise after closer scrutiny. This remark brings me to another preliminary point. One should not interpret applications of the GC as a stimulus to forget about models involving latent variables and only consider their analogues in terms of manifest variables. Latent variable models can have transparent structures which may be difficult to discern in manifest variable analogues. I will return to this point in the next section.

The GC has been formulated in terms of dynamic NARMA processes, i.e., models for time-dependent processes. A dynamic NARMA model belongs to the class of so-called causal models of time-dependent processes. A causal time series model for a univariate process x(t), t=0,±1,... provides an explanation of the dynamics of x(t) in terms of {x(t-k-1), **z**(t-k)}, where k ≥ 0. Accordingly, x(t) at each time point t is explained by previously occurring values x(t-k-1), k ≥ 0, and instantaneous or previously occurring realizations of extraneous influences **z**(t-k), k ≥ 0, where the r-variate process **z**(t), r ≥ 1, usually includes a latent innovation process and possibly additional time-dependent influences. A causal time series model does *not* explain the dynamics of x(t) at each time point t in terms of future values x(t+m) and **z**(t+m), m ≥ 1. It therefore obeys the simple characterization of physical causality according to which an effect never can precede its cause. Causality thus understood, namely that effects should not lead their causes in time, is called Granger causality in econometric time series analysis (cf. Lütkepohl, 1993). In space-time models of physical wave processes the restriction that effects should follow their causes in time gives rise to so-called dispersion relationships. Such a dispersion relationship has been used to identify causal wave models of the electro-cortical potential field in Molenaar (1987).

Despite the centrality of causal time series models of time-dependent processes, their use in other areas of signal analysis is less prominent. In particular the spatial modeling of patterns requires consideration of noncausal variants of (N)ARMAs. Take for instance a given picture made up of a grid of pixels. The assumption that the picture is given, rules out any consideration of its build-up and eventual subsequent changes in time (the latter dynamic aspect is called pattern formation in mathematical biology; e.g., Murray, 1993, Chapters 14-17). A spatial (N)ARMA describing the covariance between neighboring pixels will in general be noncausal, because there exists no lead-lag direction in space that has an intrinsic relationship with physical causality. Rosenblatt (2000, Chapter 8) presents an in-depth discussion of the consequences of noncausality for spatial ARMA model estimation.

As I remarked in the foregoing sections, any q-factor model can be rewritten as a restricted q-variate latent simplex. This formal equivalence holds irrespective of the contingent fact whether or not the manifest variables constitute repeated measurements obtained in some longitudinal design. Consider for instance a q-factor model for a p-variate manifest variable $\mathbf{y}_i$ obtained in a cross-sectional design. The p component variables of $\mathbf{y}_i$ lack any time-dependence in the sense which is relevant for physical causality. Paraphrasing the jargon of relativity theory: the component variables in $\mathbf{y}_i$ are not time-like, at the most they are space-like. Hence it might be considered to be more natural to rewrite the q-factor model for $\mathbf{y}_i$ as a restricted latent *noncausal* q-variate simplex. Moreover, according to the same reasoning the GC should be reformulated (and proven, although this will involve only a slight adaptation of the proof given earlier) in terms of noncausal NARMA models. This point of view, namely that the manipulation of latent variables in cross-sectional factor models requires an analogue of the GC which is formulated in terms of noncausal NARMA models, should be taken seriously. I suspect that a similar point of view is shared by Browne in his masterful discussion of circumplex models, when he defines these models in a noncausal way (Browne, 1992). Notwithstanding this need to elaborate a noncausal extension of the GC and the Houdini transformation, it should be recognized that causal (N)ARMA models have their place in the manipulation of space-like latent variable models. Our successful applications of the GC and the Houdini transform, yielding causal NARMA equivalents of standard factor models, bear witness to this. For a useful compilation of early benchmark papers on spatial causal ARMA models, the reader is referred to Mitra & Ekstrom (1978). See Elliott, Aggoun & Moore (1995, Chapter 9) for an up-to-date exposition of spatial modeling.

While I expect that causal and noncausal NARMA models will turn out to play mutually supporting, *not* mutually exclusive, roles in the manipulation of latent variables, there is another perspective from which noncausal models might be assigned independent importance. This perspective concerns the possible interpretation of the kind of structure that is obtained after application of the Houdini transformation to a cross-sectional factor model. Stated more specifically, how could one interpret the manifest restricted NARMA(1,1) that is obtained after application of the Houdini transform to a 1-factor model for $\mathbf{y}_i$? Obviously, this NARMA(1,1) describes a kind of interaction between component variables in $\mathbf{y}_i$. One could say that the latent factor has been dissolved into a pairwise interaction defined on a lattice composed of the manifest univariate components of $\mathbf{y}_i$. In a similar vein, the restricted NARMA(q,q) obtained after application of the Houdini transformation to a q-factor model for $\mathbf{y}_i$ can be interpreted as describing a (q+1)-term interaction defined on a lattice carrying the manifest univariate component variables in $\mathbf{y}_i$. The geometrical (topological, metrical) properties of such a lattice are inherited from the experimental design according to which $\mathbf{y}_i$ has been obtained. A longitudinal design will assign stronger (metrical) properties to the lattice than a cross-sectional design. In this way the interplay between algebraic group theory and experimental designs could be given a natural place in the analysis of structural equation models. But what I consider to be an even more interesting aspect of the manifest NARMA equivalents obtained by the Houdini transformation is the prospect to apply concepts and techniques of field theory to the lattice structures thus obtained. Statistical field theory (e.g., Le Bellac, 1991) has recently made its way into psychometrical realms in the guise of computational techniques for probabilistic graphical models (Opper & Saad, 2001). Probabilistic graphical models not only are intimately linked up with the issue of causality (e.g., Pearl, 2000), but also can be assigned lattice properties in the context of structural equation model (cf. Koster, 1999).

In my view the recent applications of mean field computational techniques drawn from statistical field theory will prove to have wider implications. Field theoretical concepts also may turn out to be powerful tools for the interpretation of structural equation models and the ongoing discussion about causality. The GC then may be helpful in making explicit the lattice structures implied by common factor models and their likes. In fact, statistical field theory may have even more to offer. It always struck me that there appears to be a close connection between the basic expressions underlying item-response theory and the solutions of elementary lattice fields in statistical physics. For instance, there is almost a one-to-one formal correspondence of the solution of the Ising model (a lattice with nearest neighbor interaction between binary-valued sites; e.g., Kindermann & Snell, 1980, Chapter 1) and the Rasch model (Fischer, 1974).

Another important qualification of the results presently obtained concerns extension of the GC to the addition of dependent NARMA processes. As indicated before, this possibility was mentioned, but not further considered, by Granger & Morris (1976). The proof of the GC for dependent NARMA processes largely can proceed along the lines of the schematic proof given in section 2.2.2. But the computational details of the mappings thus defined will become more complex and for the moment still constitute *terra incognito*. Yet elaboration of the GC for dependent NARMA processes as well as the analogous Houdini transformation is important in applications of the techniques for manipulation of latent variables in longitudinal factor models with multivariate manifest variables at each time point. Consider the standard longitudinal factor model given in section 2.1.6 in which $\mathbf{y}_i(t)$ denotes a p-variate vector of observations for subject i at time t; t=1,2,...,T. Then the longitudinal factor model is defined by:

$$\begin{aligned} &\mathbf{y}_i(t) = \mathbf{\Lambda}_t \boldsymbol{\eta}_i(t) + \boldsymbol{\varepsilon}_i(t), \text{ t=1,...,T; i=1,2,...} \\ (2.24) \quad & \\ &\boldsymbol{\eta}_i(t) = \mathbf{B}_{t,t-1}\boldsymbol{\eta}_i(t-1) + \boldsymbol{\zeta}_i(t), \text{ t=2,...,T} \end{aligned}$$

where $\mathbf{\Lambda}_t$ is a (p,q)-dimensional matrix of factor loadings at time t, $\boldsymbol{\eta}_i(t)$ is a q-variate latent factor at time t, $\boldsymbol{\varepsilon}_i(t)$ is p-variate measurement error at time t, $\mathbf{B}_{t,t-1}$ is the (q,q)-dimensional matrix of regression weights linking $\boldsymbol{\eta}_i(t)$ to $\boldsymbol{\eta}_i(t-1)$, and $\boldsymbol{\zeta}_i(t)$ denotes q-variate innovation at time t. If p = q = 1 then the quasi-simplex model is obtained, for which the Houdini transformation has been specified in section 2.2.4. If q > 1 then the standard longitudinal factor model involves a q-variate simplex at the latent level. The pattern of dependence between the q univariate component processes making up $\boldsymbol{\eta}_i(t)$ is described by $\mathbf{B}_{t,t-1}$, t=2,...,T and $\boldsymbol{\Psi}_t$ = cov[$\boldsymbol{\zeta}_i(t)$, $\boldsymbol{\zeta}_i(t)'$], t=1,...,T ($\boldsymbol{\eta}_i(1) = \boldsymbol{\zeta}_i(1)$).

For arbitrary identifiable $\mathbf{B}_{t,t-1}$ and $\boldsymbol{\Psi}_t$ this pattern of dependence between the q univariate component processes in $\boldsymbol{\eta}_i(t)$ can be quite intricate. A special case is obtained by restricting $\mathbf{B}_{t,t-1}$, t=2,...,T to be a sequence of T-1 (q,q)-dimensional diagonal matrices, while $\boldsymbol{\Psi}_t$, t=1,...,T still is arbitrary. In this case a straightforward generalization of the GC for dependent NARMA processes can be formulated according to which the q univariate components in $\boldsymbol{\eta}_i(t)$ can be added. I think that this is the kind of dependence which Granger & Morris (1976) had in mind when they mentioned the possibility to extend their theorem to the addition of dependent ARMA processes. For arbitrary $\mathbf{B}_{t,t-1}$ and $\boldsymbol{\Psi}_t$, however, I expect that more powerful transformation techniques will be necessary in order to arrive at a representation involving weakly independent univariate component processes. Such a transformation technique is available for weakly stationary multivariate processes (Brillinger, 1975, Chapter 9; Molenaar, 1987).

The discussion in this section shows that the proposed techniques to manipulate latent variables only constitute the first few steps into a large unexplored area. Even with respect to the steps actually made there remain several additional aspects that require further scrutiny and elaboration. For instance concerning the best way to handle the case p > 1 in the standard longitudinal factor model. As to that, in the next section some references are made to relevant results from algebraic systems theory.

2.3 General state space methods to manipulate latent variables

Part of the theory of state-space models is devoted to the question: “Given a specific process, what systems can generate it as an output process when the input process is restricted to be simple in some technical sense?” (Caines, 1988, p. 199). For instance, suppose that a given process is weakly stationary and that input processes are required to be white noise processes. Then the set of possible systems that might have generated the given process under the influence of white noise input includes those of the ARMA type. Another way to put this is that weakly stationary processes can be generated or realized by means of ARMA systems driven by white noise input. Accordingly, the part of the theory of state-space models concerned with this question is called realization theory. Realization theory aims to provide for the delineation of important equivalence classes of systems, the complete characterization of each equivalence class, and the specification of different representations of each element of an equivalent class. It is especially this latter aim, specification of different model representations of a system type, that is of present interest, because it covers (among a great many of other things) the ways in which different representations (for instance ARMA and state-space) are related to each other. My aim in this section is very modest, namely to present and discuss a fundamental theorem in the realization theory of linear stochastic systems on the relationship between multivariate NARMA models and state-space models. It will be shown that the Houdini transformation is a special case of this theorem.

In what follows I will mainly draw on two sources: Caines (1988) and Hannan & Deistler (1988). To the best of my knowledge, Caines (1988) is the only source in which the relationship between multivariate NARMA models and state-space models is addressed head-on. Many publications only address the transformation in one direction, namely rewriting a multivariate NARMA model as state-space model. But it is especially the reverse direction of this relationship that is of interest for our present concerns: rewriting a state-space model as a multivariate NARMA model. Hannan & Deistler (1988, Theorem 1.2.1, p. 16) also address this relationship, but in an indirect way, without proofs, and restricted to the weakly stationary case.

Consider the following linear nonstationary state-space system:

$$\begin{aligned} \mathbf{y}(t) &= \boldsymbol{\Lambda}(t)\boldsymbol{\eta}(t) + \boldsymbol{\varepsilon}(t) \\ \boldsymbol{\eta}(t+1) &= \mathbf{B}(t)\boldsymbol{\eta}(t) + \boldsymbol{\zeta}(t) \end{aligned} \qquad (2.25)$$

where $\mathbf{y}(t)$ is a p-variate manifest (output) process, $\boldsymbol{\varepsilon}(t)$ is p-variate white noise measurement error, $\boldsymbol{\varepsilon}(t) \sim \aleph(\mathbf{0}, \boldsymbol{\Theta}_t)$, $\boldsymbol{\eta}(t)$ is a q-variate state process, $\boldsymbol{\zeta}(t)$ is q-variate white noise innovation, $\boldsymbol{\zeta}(t) \sim \aleph(\mathbf{0}, \boldsymbol{\Psi}_t)$, and $\boldsymbol{\Lambda}(t)$ and $\mathbf{B}(t)$ are matrices of appropriate dimensions at each time t. The similarity of this state-space model to the standard longitudinal factor model (2.24) given in the previous section should be obvious. The stationary analogue of (2.25) is given by (1.2) and (1.3) in Chapter 1. It is our aim to show that (2.25) can be rewritten as a p-variate NARMA model (to be defined below). This will be accomplished in a number of steps, each of which is described in a somewhat heuristic fashion. No proofs will be given, only detailed references to Caines (1988) and Hannan & Deistler (1988) where the proofs concerned can be found. The reason for doing so is that the proofs, although rather simple, are not very illuminating. Moreover, proving the equivalence between (2.25) and the class of p-variate NARMA models should be conceived of as a proof of the existence of mapping rules between the two types of representation. The actual construction of these mapping rules is an entirely different, and much more difficult, matter that is made even more difficult due to the intricate effects of initial conditions. It is in this constructive phase, as yet unrealized in the present context, where the hard work will have to be done.

The first step in rewriting (2.25) as a p-variate NARMA model involves the introduction of a restrictive assumption. It is required that the matrix $\mathcal{O}(t; q)$ has full rank q for each time t, where $\mathcal{O}(t; q)' = [\boldsymbol{\Lambda}(t)', \{\boldsymbol{\Lambda}(t+1)\mathbf{B}(t+1)\}', \{\boldsymbol{\Lambda}(t+2)\mathbf{B}(t+1)\mathbf{B}(t+2)\}', ..., \{\boldsymbol{\Lambda}(t+q)\mathbf{B}(t+1)\mathbf{B}(t+2)...\mathbf{B}(t+q-1)\}']$. If (2.25) obeys this assumption then it is called observable. To see what this implies, suppose that measurement error is lacking in (2.25), i.e., $\mathbf{y}(t) = \boldsymbol{\Lambda}(t)\boldsymbol{\eta}(t)$, and that (2.25) is observable. Then $\boldsymbol{\eta}(t)$ can exactly be determined (observed) from $\mathbf{y}(t; q) = [\mathbf{y}(t)', \mathbf{y}(t+1)', ..., \mathbf{y}(t+q-1)']'$ as

$$\boldsymbol{\eta}(t) = [\mathcal{O}(t; q)'\mathcal{O}(t; q)]^{-1}\mathcal{O}(t; q)'\mathbf{y}(t; q)$$

(see Shumway & Stoffer, 2000, p. 328, for the stationary case).

The second step consists in rewriting (2.25) in a different form, the so-called prediction error form. Define the one-step ahead prediction error $\boldsymbol{\nu}(t) = \mathbf{y}(t) - \boldsymbol{\Lambda}(t)\boldsymbol{\eta}(t \mid t-1)$, where $\boldsymbol{\Lambda}(t)\boldsymbol{\eta}(t \mid t-1)$ is the prediction of $\mathbf{y}(t)$ based on information up to time t-1, i.e., based on {$\mathbf{y}(k)$, k < t}. Note that $\boldsymbol{\nu}(t)$ is a p-variate white noise process. Then (2.25) can be rewritten as:

$$\begin{aligned} &\mathbf{y}(t) = \boldsymbol{\Lambda}(t)\boldsymbol{\eta}(t \mid t\text{-}1) + \boldsymbol{\nu}(t) \\ (2.26)\qquad & \\ &\boldsymbol{\eta}(t+1 \mid t) = \mathbf{B}(t)\boldsymbol{\eta}(t \mid t\text{-}1) + \mathbf{K}(t)\boldsymbol{\nu}(t) \end{aligned}$$

The (q,p)-dimensional matrix $\mathbf{K}(t)$ is defined recursively, in the same way as the matrix of the same name in the definition (1.5) of the Kalman filter in Chapter 1. Although the Kalman filter given in (1.5) is based on a stationary state-space model, the same set of expressions hold for the nonstationary state-space model (2.25) after substituting time-varying model parameters for their time-homogeneous analogues (cf. Hannan & Deistler, 1988, p. 91). The proof of the equivalence of (2.25) and (2.26) is given in Hannan & Deistler (1988, p. 16-18). Their proof, consisting of simple substitution steps, is given for the stationary state-space model, but carries over straightforwardly to the nonstationary case.

The third and final step consists in the proof of the following theorem adapted from Caines (1988, Theorem 4.3, part 2, p. 111-114): Let (2.26) be observable for all times t. Then it can be rewritten as a p-variate NARMA(q,q) model given by

$$\begin{aligned} &\mathbf{A}_t[B, q]\mathbf{y}(t) = \mathbf{C}_t[B, q]\boldsymbol{\nu}(t) \\ (2.27)\qquad &\mathbf{A}_t[B, q] = \mathbf{I}_p + \mathbf{A}_{t,1}B + \ldots + \mathbf{A}_{t,q}B^q \\ &\mathbf{C}_t[B, q] = \mathbf{I}_p + \mathbf{C}_{t,1}B + \ldots + \mathbf{C}_{t,q}B^q \end{aligned}$$

where $\boldsymbol{\nu}(t)$ is the p-variate white noise prediction error in (2.26) and where for each time t $\mathbf{A}_{t,i}$ and $\mathbf{C}_{t,i}$, i=1,...,q, are (p,p)-dimensional autoregressive and moving-average coefficient matrices, respectively. The proof is obtained by writing out the second equation of (2.26) for t+q+1 consecutive time points in the following way (letting $\boldsymbol{\xi}(t)$ denote $\boldsymbol{\eta}(t \mid t\text{-}1)$ in order to avoid cumbersome notation):

$$\boldsymbol{\xi}(t+1) = \mathbf{B}(t)\boldsymbol{\xi}(t) + \mathbf{K}(t)\boldsymbol{\nu}(t)$$

$$\boldsymbol{\xi}(t+2) = \mathbf{B}(t)\mathbf{B}(t+1)\boldsymbol{\xi}(t) + \mathbf{B}(t+1)\mathbf{K}(t)\boldsymbol{\nu}(t) + \mathbf{K}(t+1)\boldsymbol{\nu}(t+1)$$

.....

$$\boldsymbol{\xi}(t+q+1) = \mathbf{B}(t)\mathbf{B}(t+1)\ldots\mathbf{B}(t+q)\boldsymbol{\xi}(t) + \mathbf{B}(t+1)\ldots\mathbf{B}(t+q)\mathbf{K}(t)\boldsymbol{\nu}(t) + \ldots + \mathbf{B}(t+q)\mathbf{K}(t+q)\boldsymbol{\nu}(t+q)$$

Because this system of equations obeys the observability criterion, and given that $\mathbf{y}(t) = \boldsymbol{\Lambda}(t)\boldsymbol{\xi}(t) + \boldsymbol{\nu}(t)$, it follows that there exist appropriate (p,p)-dimensional matrices $\mathbf{A}_{t,i}$ and $\mathbf{C}_{t,i}$, i=1,...,q, such that (2.27) is obtained. In case the state-space system is stationary, the part in the proof played by the observability criterion can be replaced by an appeal to the Cayley-Hamilton theorem (cf. Padulo & Arbib, 1974; p. 283; Caines, 1988, p. 814-815).

It appears that the proof of Caines’ theorem involves two simple steps: a) transform the state-space model into prediction error form and b) use the observability criterion to conclude that at each time t the state-space in this model is spanned by the rows of $\mathcal{O}(t; q)$, implying that the system of q+1 equations for $\xi(k)$, k=t+1,...,t+q+1, given above is linearly dependent. One thus obtains a proof of the existence of one-to-one mapping rules between (2.25) and (2.27), i.e., between linear nonstationary state-space models and p-variate NARMA(q,q) models. With Caines' theorem in hand, the specification of mapping rules can proceed without further worries about their existence. In this sense, the theorem corroborates a Platonic interpretation of mathematical objects, including the difficulties encountered in actually specifying their shadowy projections in our sensuous world. In the special case of stationary processes use can be made of polynomial matrix algebra (Blomberg & Ylinen, 1983), but even here the handling of initial conditions will complicate the mapping rules thus obtained. We already encountered the top of this iceberg of complications due to initial conditions in sections 2.2.5 and 2.2.6. This will become much worse in the more general case.

This last remark brings us to the question about the relationship between the Houdini transformation and Caines' theorem. This question has an easy and unambiguous answer: the Houdini transformation is a special case of Caines' theorem in that the latter implies the existence of the former. Consider the Houdini transformation for the 1-factor model specified in section 2.2.5 and suppose, to simplify matters, that the observations consist of longitudinal univariate data obtained at T measurement occasions. The 1-factor model then is given by: $y_i(t) = \lambda_t\eta_i + \varepsilon_i(t)$, t=1,...,T, which is of the same form as the first equation in (2.25) with p = 1 and q = 1. As discussed in Chapter 1 and at the start of section 2.2.3, at this level of model specification the shift from a within-subjects perspective (state-space model) to a between-subjects perspective (longitudinal factor model) is immaterial. According to Caines' theorem, this 1-factor model can be rewritten as a univariate NARMA(1,1), which is the same conclusion as obtained by invoking the Houdini transformation. Of course, our Houdini transformation for the 1-factor model also specifies the details of the mapping rules involved, making use of the GC. These constructive details are lacking in Caines' theorem.

3. The (lack of) relationship between longitudinal and time series analysis

In the previous two chapters we saw that the standard longitudinal factor model and the linear state-space model have the same formal structure. The matrix-algebraic equations making up the longitudinal factor model (2.24) are the same as those defining the state-space model (2.25). This formal equivalence has been exploited in Chapter 1 to show the relationship between the regression factor score estimator and the Kalman filter (this relationship will be extended to longitudinal factor score estimators in the next chapter). In the same vein, in Chapter 2 the first steps have been made towards a realization theory for structural equation models. In the present chapter I will further pursue the common homology of longitudinal and state-space models, but this time in a more discriminative fashion. Despite the fact that state-space and longitudinal factor models bear close family relationships to each other regarding structural form and estimation theory, they differ in one important aspect. In the wordings of Chapter 1, longitudinal factor models describe the variation between homogeneous systems (between-system variation, BSV), whereas state-space models describe the variation of a single system (within-system variation, WSV). This is not so much a difference at the level of algebraic or mathematical-statistical theory, but concerns the possibly different characteristics of these two types of variation themselves.

In what follows I will argue that the actual structure of BSV has to be considered to be different from the analogous structure of WSV, unless certain strict criteria have been met. Stated more succinctly, it will be shown that longitudinal factor analysis of BSV yields results that should be considered to be unrelated to those obtained in state-space analysis of WSV, unless explicitly established otherwise. I draw my arguments from the mathematical-statistical literature, from psychometric sources, from theoretical psychology, and from results obtained with simulation studies. But first and foremost I intend to present a coherent set of reasons why the structure of variation between and within systems will differ in general, and why this state of affairs has important consequences for applied psychometrics and psychology.

3.1 *Ensembles again*

In Chapter 1 the notion of ensemble was introduced to provide a common framework for the discussion of factor models and state-space models. An ensemble was defined as the set $\{\mathbf{y}_i(t);\ t=0,\pm1,...;\ i=1,2,...\}$ composed of a countable infinite number of realizations of a random p-variate process $\mathbf{y}(t)$. Accordingly, an ensemble can be conceived of as an extensive entity (in the thermodynamical sense of extensive as being directly proportional to the size of something; cf. Thompson, 1988, p.31-32) corresponding to the well-known concept of population in sampling-theoretical statistics. The concept of ensemble arose in statistical mechanics and thermodynamics, where its precise status has been the subject of some controversy. In contrast, ensembles appear to be relatively unknown in psychometrics. What, then, are the reasons to consider it in the present context? Why not stick to the simple and clear twin concepts of population and domain of generalization? The answers I will give to these questions are tuned to the main theme of this chapter, namely the relationship between results obtained with analyses of BSV and of WSV. This seemingly redundant announcement is made because some aspects of the following discussion not only have not been addressed elsewhere in the psychometrical literature on structural equation modeling, but also may initially seem to be somewhat unrelated to our main theme.

I have always felt a bit uneasy about the concept of infinite population in statistics. It denotes the set of entities which are homogeneous in all relevant aspects, a random sample of which is drawn for explicit study. It also constitutes the domain of generalization of the results obtained with the random sample. The characterization "homogeneous in all relevant aspects" will be considered later on; for the moment let us take it at face value (ambiguities included). But what *is* a population? Is it a set of real entities, or is it instead a set of virtual entities? If it is a set of real entities, then are these entities supposed to interact with each other? And can different populations interact? If it is a set of virtual entities, then is there a separate population for each experimental set-up (like the way in which state preparation figures in quantum physics; cf. Bellantine, 1990, Chapter 8)? Questions such as these are hardly raised in the psychometrical literature, so I had to look elsewhere for possible answers. It turned out that similar questions concerning the status of ensembles have given rise to important theoretical developments in statistical mechanics (cf. Farquhar, 1964). It is not so much the case that one can give unique answers to questions such as whether or not an ensemble should be considered to be virtual, because different, but consistent, answers have been given to this particular question. Yet it becomes evident that the way in which one conceptualizes ensembles (e.g., as real or virtual sets) has major implications for the further theoretical built-up of statistical mechanics. Perhaps what holds for ensembles in statistical mechanics also pertains to the similar concept of population in statistics. Perhaps the concept of population is not so unproblematic after all.

The main reason I have for considering the concept of ensemble is that it provides one with a mental picture that is helpful in understanding the relationship between analyses of BSV and WSV. In contrast to the rather static concept of population, an ensemble is inherently tied up with dynamics. A population is a set of indices or a domain, whereas an ensemble is a manifold of trajectories of some dynamical system. A well-known example of the dynamical nature of ensembles is given by the action of Hamiltonian systems in so-called phase space. Such action induces a smooth manifold of trajectories with invariant measure (Liouville theorem, cf. Cornfield, Fomin, & Sinai, 1982, p. 48). In the present context we will only need a much more simple kind of ensemble. The ensemble $\{\mathbf{y}_i(t); t=0,\pm1,...; i=1,2,...\}$ introduced above consists of a countably infinite number of time-dependent trajectories, where each trajectory is the realization of a p-variate process $\mathbf{y}(t)$. The dynamical nature of this ensemble is evident. In what follows it will be called our standard ensemble.

The standard ensemble provides a natural setting for discussion of the relationship between analyses of BSV and WSV. It already served that purpose in Chapter 1 where the relationship between the regression factor score estimator and the Kalman filter was derived. An analysis of BSV typically starts with fixing T time points, where T usually is small (including the possibility that T=1). A sample of N realizations is drawn, where N usually is large, yielding the observations $\{\mathbf{y}_i(t); t=1,2,...,T; i=1,2,...,N\}$. Next all statistics (means, covariances, etc.) are obtained by averaging over i=1,...,N. For instance, the covariance matrix in a longitudinal factor analysis at T time points is estimated by constructing for each i the Tp-dimensional supervector $\mathbf{y}_i$ = [$\mathbf{y}_i$(1)', ..., $\mathbf{y}_i$(T)']', after which the (Tp,Tp)-dimensional longitudinal covariance matrix is estimated by averaging products $\mathbf{y}_i\mathbf{y}_i$' over i=1,...,N (neglecting inessential mean corrections for the moment). In contrast, an analysis of WSV typically starts with fixing one system i=1 that is observed at T time points, where T usually is large. This yields the data {$\mathbf{y}_1$(t); t=1,2,...,T}. Next all statistics (mean function, covariance function, etc.) are obtained by averaging over t=1,...,T. For instance, the (p.p)-dimensional covariance function at lag u in a state-space analysis is estimated by averaging products $\mathbf{y}_1$(t)$\mathbf{y}_1$(t+u) over t=1,...,T-|u|.

Despite the differences between an analysis of BSV (statistics are averages over systems) and of WSV (statistics are averages over time points), the description just given makes clear that both are based on data sampled from the same standard ensemble. Sampling data from an ensemble can be likened to placing a window of finite extent over the ensemble and keeping the part within this window as data. The way in which such a window is placed in an analysis of BSV differs from that in an analysis of WSV. But in both types of analysis the window is placed over the same ensemble, and therefore this ensemble provides a natural setting for comparison of these types of analysis.

Having presented my reasons for considering ensembles, I should hasten to add a qualification and also mention a caveat. The qualification concerns the mere heuristic use which will be made of the concept of ensemble. In what follows, no appeal will be made to the profound theoretical elaborations of the various kinds and roles of ensembles in statistical mechanics. This does not imply that I consider these theoretical issues unimportant for psychometrics and structural equation modeling (as indicated by some of my critical remarks in this section about the concept of population), but any serious consideration would take us too far from the main topic of this book. The caveat concerns the ontological status of the trajectories in our standard ensemble: are they best conceived of as being real or virtual? Until now different values of the subscript i in the ensemble {$\mathbf{y}_i$(t); t=0,±1,...; i=1,2,...} have been treated as referring to different systems. This is allowed, but then it should be acknowledged that for each given value of i = S (for each individual system S), $\mathbf{y}_S$(t), t=0,±1,... denotes a random process that itself is composed of an infinite number of possible trajectories. The latter subensemble of possible trajectories of i = S constitutes a set of virtual entities, while the different systems i=1,2,... themselves can be taken to be real. It follows that the standard ensemble has more structure than is made explicit in our notation.

3.2 (Non)-ergodicity

In this section I will start with the presentation of a number of arguments which all converge to the same conclusion, namely that there are in general no lawful relationships between results of analyses of BSV and WSV. We first take a look at a mathematical-statistical theory that explicitly deals with this relationship, namely ergodic theory. The concept of ergodicity was introduced by Boltzmann in his work on the foundations of statistical mechanics. Sklar (1993) and Guttman (1994) give excellent nontechnical discussions of the intricacies associated with the ergodic hypothesis in statistical mechanics. The ergodic hypothesis roughly states that in a pure gas, kept at a fixed temperature in a container shielded from the environment, averages taken along the trajectory of a single gas molecule in phase space asymptotically approach averages taken with respect to the distribution of molecules in phase space. The trajectory of a single molecule m describes its “life history” and is an intrinsically dynamical entity. Taking averages along such a trajectory over increasingly large time intervals can be schematically represented as

$$(3.1) \qquad f[\mathbf{y}_m(*)] = \lim_{T\rightarrow\infty} T^{-1}\sum_t f[\mathbf{y}_m(t)]$$

where f[.] denotes a sufficiently smooth function. Taking averages with respect to the distribution of molecules in phase space can be schematically represented as

$$(3.2) \qquad f[\mathbf{y}_*] = N^{-1}\sum_i f[\mathbf{y}_i]$$

where N is understood to be very large (Avogadro’s number). The time argument has been omitted in (3.2) to accentuate the difference with (3.1); one can understand (3.2) as holding at some fixed time t.

Although this simplistic sketch of the ergodic hypothesis leaves out all interesting issues, it makes clear that the hypothesis captures the essence of our main question in this chapter concerning the relationship between analyses of BSV and WSV. The ergodic hypothesis proclaims the equality of (3.1) and (3.2), where in our terminology (3.1) pertains to an analysis of WSV and (3.2) pertains to an analysis of BSV. Hence any answer to the ergodic hypothesis would seem to imply an answer to our present question. It will turn out that the ergodic hypothesis (in some suitably modified form) can only be accepted if the “life histories” of molecules (and of systems in our standard ensemble) obey very strict criteria. This implies that in all those cases where these criteria are not met, equality of (3.1) and (3.2) does not obtain. And by implication, it then follows that in these cases analysis of BSV yields results that are different from those obtained in analysis of WSV.

What *are* the conditions under which the ergodic hypothesis holds? These are laid down in a number of ergodic theorems proved by eminent mathematicians such as Birkhoff, von Neumann, Hopf, and Kingman. Petersen (1983) and Cornfield, Fomin, & Sinai (1982) give excellent overviews of this material. These deep results, however, will not be needed in the present context in which we focus on the simplest possible case of Gaussian processes $\{\mathbf{y}_i(t); t=0,\pm1,...; i=1,2,...\}$. A Gaussian process is ergodic if it obeys the following restrictions: a) it is weakly stationary (and hence strictly stationary), and b) its spectrum has no jumps (cf. Hannan, 1970, p. 201). The latter restriction b) rules out the presence of sinusoidal trends. If a Gaussian process obeys a) and b), then (3.1) equals (3.2) for this process. More specifically, the mean function and, more importantly, the covariance function of this process as obtained from (3.1) converge to their analogues defined by (3.2). Hence for an ensemble of ergodic Gaussian systems, the covariance function estimated along the trajectory of one individual system (i.e., derived from WSV) converges to the covariance function averaged across all systems in the ensemble (i.e., derived from BSV). This is the content of a theorem in Hannan (1970, p. 203, Theorem 2).

Evidently, the restriction guaranteeing that Gaussian processes are ergodic is rather severe: they have to be weakly stationary. More specifically, weak stationarity is a sufficient criterion for ergodicity of Gaussian processes. We will not dwell on necessary criteria, because that would require explicit consideration of the ergodic theorems. Instead, I will use the following principle: a Gaussian process is called almost certainly non-ergodic (ACNE; pun intended) if it is not weakly stationary. This principle accommodates certain special cases such as discussed in Gray (1988).

It will be clear that many empirical processes suffer from ACNE. Developmental processes, learning processes, evolutionary processes, these are all examples of processes affected by ACNE because they are, almost by definition, nonstationary in various respects. But there are many other examples of a more special nature, such as the requirement that neural networks have to be nonergodic in order to function properly (Amit, 1989). The diagnosis for the existence of lawful relationships between results obtained in analyses of BSV and WSV does not look good.

To specify where this leads us with respect to the relationship between longitudinal factor analysis and state-space analysis,
consider again the longitudinal factor model given by (2.24):

$$\mathbf{y}_i(t) = \Lambda_t\boldsymbol{\eta}_i(t) + \boldsymbol{\varepsilon}_i(t),\ t=1,...,T;\ i=1,2,...$$

$$\boldsymbol{\eta}_i(t) = \mathbf{B}_{t,t-1}\boldsymbol{\eta}_i(t-1) + \boldsymbol{\zeta}_i(t),\ t=2,...,T$$

where Λ_t is a (p,q)-dimensional matrix of factor loadings at time t, $\boldsymbol{\eta}_i(t)$ is a q-variate latent factor at time t, $\boldsymbol{\varepsilon}_i(t)$ is p-variate Gaussian measurement error at time t: $\boldsymbol{\varepsilon}(t) \sim\aleph(\mathbf{0}, \Theta_t)$, $\mathbf{B}_{t,t-1}$ is the (q,q)-dimensional matrix of regression weights linking $\boldsymbol{\eta}_i(t)$ to $\boldsymbol{\eta}_i(t-1)$, and $\boldsymbol{\zeta}_i(t)$ denotes q-variate Gaussian innovation at time t: $\boldsymbol{\zeta}(t) \sim\aleph(\mathbf{0}, \Psi_t)$. This model is ergodic if it has the restricted form:

$$\mathbf{y}_i(t) = \Lambda\boldsymbol{\eta}_i(t) + \boldsymbol{\varepsilon}_i(t),\ t=1,...,T;\ i=1,2,...$$

(3.3)

$$\boldsymbol{\eta}_i(t) = \mathbf{B}\boldsymbol{\eta}_i(t-1) + \boldsymbol{\zeta}_i(t),\ t=2,...,T$$

where Λ is invariant over time, $\boldsymbol{\varepsilon}_i(t)$ has constant covariance: $\boldsymbol{\varepsilon}(t) \sim\aleph(\mathbf{0}, \Theta)$, $\mathbf{B}$ is invariant over time, and $\boldsymbol{\zeta}_i(t)$ has constant covariance: $\boldsymbol{\zeta}(t) \sim\aleph(\mathbf{0}, \Psi)$. In addition, the absolute value (modulus) of each eigenvalue of $\mathbf{B}$ has to be strictly less than 1. In a similar vein, consider again the linear nonstationary state-space system given by (2.25):

$$\mathbf{y}(t) = \Lambda(t)\boldsymbol{\eta}(t) + \boldsymbol{\varepsilon}(t)$$

$$\boldsymbol{\eta}(t+1) = \mathbf{B}(t)\boldsymbol{\eta}(t) + \boldsymbol{\zeta}(t)$$

where $\mathbf{y}(t)$ is a p-variate manifest (output) process, $\boldsymbol{\varepsilon}(t)$ is p-variate Gaussian white noise measurement error: $\boldsymbol{\varepsilon}(t) \sim\aleph(\mathbf{0}, \Theta_t)$, $\boldsymbol{\eta}(t)$ is a q-variate state process, $\boldsymbol{\zeta}(t)$ is q-variate Gaussian white noise innovation: $\boldsymbol{\zeta}(t) \sim\aleph(\mathbf{0}, \Psi_t)$, and $\Lambda(t)$ and $\mathbf{B}(t)$ are matrices of appropriate dimensions at each time t. This state-space system is ergodic if it has the restricted form:

$$\mathbf{y}(t) = \Lambda\boldsymbol{\eta}(t) + \boldsymbol{\varepsilon}(t)$$

(3.4)

$$\boldsymbol{\eta}(t+1) = \mathbf{B}\boldsymbol{\eta}(t) + \boldsymbol{\zeta}(t)$$

where $\boldsymbol{\varepsilon}(t)$ has constant covariance:: $\boldsymbol{\varepsilon}(t) \sim\aleph(\mathbf{0}, \Theta)$, $\boldsymbol{\zeta}(t)$ has constant covariance: $\boldsymbol{\zeta}(t) \sim\aleph(\mathbf{0}, \Psi)$, while Λ and $\mathbf{B}$ are invariant over time. In addition, the absolute value (modulus) of each eigenvalue of $\mathbf{B}$ has to be strictly less than 1. Model (3.3) describes the structure of BSV, whereas (3.4) describes the structure of WSV. Asymptotically, application of (3.3) and (3.4) yields converging results (in some appropriate probabilistic sense). As soon as one or more parameter matrices are not time-invariant, or eigenvalues of $\mathbf{B}$ have modulus larger than 1, (3.3) and (3.4) reduce to (2.24) and (2.25), respectively. The latter nonstationary models suffer from ACNE and their application can no longer be expected to yield converging results.

3.3 The restrictive nature of classical test theory

In this section an aspect of classical test theory is discussed that is not often emphasized in the psychometric literature. It will be shown that this aspect has a direct bearing on our question concerning the relationship between analyses of BSV and WSV. The focus will be on the definition of true score which found its most eloquent expression in Lord & Novick (1968). They define the true score of a fixed person as the expected value of the observed score of this person with respect to the propensity distribution of this person's observed scores. The latter propensity distribution is characterized as a "... distribution function defined over repeated statistically independent measurements on the same person" (Lord & Novick, 1968, p. 30). It is assumed that the repeated measurements do not affect the person in that in each replication the person responds without any aftereffects of previous assessments (e.g., due to memory, habituation, etc.). Stated more formally, it is assumed that in each replication the observed score is an independent realization of the same random variable.

It is clear from this definition of the true score of a fixed person that it is based on a stochastic process describing the WSV of this person. The stochastic process is supposed to lack any sequential dependencies and in this sense it is akin to a white noise process. But contrary to a regular white noise process, the process underlying a propensity distribution has in general nonzero mean function. In addition, this process is supposed to be strictly stationary in that it obeys the same propensity distribution at each time point (measurement occasion). This implies that the stochastic process has constant mean function (equal to the true score of the person), while its covariance function is stationary: constant variance at each time point and zero covariance at all nonzero lags. We will refer to this particular stochastic process underlying a propensity distribution as the Lord & Novick process.

With respect to this WSV definition of true score (and of error score as the difference between observed score and true score), Lord & Novick (1968, p. 32) make the following remark: "The true and error scores defined above are not those primarily considered in test theory ... . They are, however, those that would be of interest to a theory that deals with individuals rather than with groups (counseling rather than selection) ... ." I consider this a noteworthy remark because it might indicate that Lord & Novick appreciate the possible lack of relationship between test theories based on WSV and BSV. The true and error scores that *are* considered in test theory are defined in terms of BSV: "Primarily, test theory treats individual differences or, equivalently, measurements over people" (Lord & Novick, 1968, p. 32). The first quoted remark of Lord & Novick would seem to imply that such a test theory based on BSV may not be relevant to individual assessment and counseling.

Classical test theory as developed in Lord & Novick (1968) is based on analysis of BSV. A population of persons is considered in which each person has its own propensity distribution of scores, characterized by a person-specific mean and variance. Taking into consideration that a Lord & Novick process $y_i(t)$ underlies each person-specific propensity distribution in this population, we obtain an ensemble $\{y_i(t); t=0,\pm1,...; i=1,2,...\}$. For each person i, $y_i(t)$ denotes a stochastic Lord & Novick process with stationary mean function (equal to the true score of this person) and stationary covariance function (equal to a white noise covariance function). Hence within each person this Lord & Novick process is at least weakly stationary. But the propensity distributions characterizing different persons in the population have person-specific means and variances, hence their underlying Lord & Novick processes are heterogeneous across persons. Consequently, the total ensemble $\{y_i(t); t=0,\pm1,...; i=1,2,...\}$ cannot be ergodic, even if the individual Lord & Novick processes are considered to be Gaussian processes. This is immediately obvious if one considers taking averages along the trajectory of one individual Lord & Novick process. The statistics thus obtained (mean, variance, etc.) pertain to one particular person in the ensemble, but certainly not to any other person in the ensemble or to the ensemble as a whole. In reverse, taking averages over persons, i.e., taking averages over the values of trajectories in the ensemble at a fixed time point, yields statistics that do not pertain to the individual propensity distributions and associated Lord & Novick processes characterizing each person in the population. Here we have a case in which nonergodicity is not due to nonstationarity, but to heterogeneity. Some rather surprising effects of heterogeneity on the relationship between analyses of BSV and WSV will be discussed in the next section.

I conclude that classical test theory is based on a heterogeneous ensemble. This implies that this ensemble is nonergodic: there are no direct relationships between classical test theory based on analysis of BSV and the structure of WSV characterizing each individual person in the population. Obviously this state of affairs has important consequences for the application of tests in individual counseling (as suggested by the quoted remark of Lord & Novick). Insofar as these tests have been constructed according to the guidelines of classical test theory, they cannot be expected to yield measures in individual assessment of a single person that obey the characteristics proclaimed by the theory.

To give an example, suppose that the propensity distribution of person P has a high mean τ_P and vanishing variance. Hence an observed score y_P of P equals the true score τ_P of P and has reliability equal to one. Suppose also that the reliability across persons in the population to which P belongs is not perfect, .80 say. Then P’s true score estimate is given by the Kelley estimator (Lord & Novick, 1968, p. 65): est-τ_P = $.80y_P + (1-.80)\mu$, where μ is the mean score in the population. Under the stated assumption that the true score τ_P of P is much higher than μ, it follows that est-τ_P will differ from τ_P, despite our assumption that the observed scores of P have reliability equal to one (P’s propensity distribution having vanishing variance). Reasoning along similar lines, consider the subset of persons having the same extremely high observed score y_{max}. Part of this subset will consist of persons having propensity distributions with large variances, whose observed score is substantially higher than the means of their propensity distributions. This part will consist of more persons than the dual part consisting of persons having propensity distributions with large variance and observed scores substantially lower than the means of their propensity distributions. Another part of this subset will be persons having propensity distributions with small variances, whose observed score will be close to the means of their propensity distributions. Yet the Kelley estimate of the true scores of all persons in this subset is the same. On average this would do injustice to the persons whose propensity distributions have small variance (assuming that the test measures a desirable trait). This kind of reasoning resembles the explanation of the regression paradox (Lord, 1963). It shows that the ordering of observed scores may not correspond to the ordering of true scores.

In closing this section I would like to address the question why Lord & Novick do not pursue their original concept of true score. Their answer is the following (Lord & Novick, 1968, p. 13): “In mental testing we can perhaps repeat a measurement once or twice, but if we attempt further repetitions, the examinee’s responses change substantially because of fatigue or practice effects”. This answer is noteworthy for a number of reasons. First, in Chapter 5 of their book Lord & Novick consider tests composed of a varying number of items, where each item constitutes a measurement. In fact, even an interpretation of true score of a person in terms of the average score on a test of infinite length is considered (Lord & Novick, 1968, p, 108). This suggest that it is possible in mental testing to repeat a measurement much more than once or twice without changing the psychological processes involved. Second, measures of mental information processing such as response latencies require extended initial practicing before the actual assessment procedures can begin. The rationale is that as long as a fixed experimental condition does not yield a stationary sequence of outcomes for a person, assessment of the target information process still is confounded by extraneous factors. Hence it is at least conceivable that the effects of practice in mental testing can signal the initial presence of confounding factors, which should be allowed to decay in extended repeated measurements.

I think that Lord & Novick are too pessimistic about the prospects of a test theory based on WSV. Moreover, their insistence that repeated measurements of the same person should yield statistically independent scores is unwarranted. Weakly stationary time series of scores are quite appropriate to determine individual propensity distributions. This has been exploited in an early paper by Drösler (1978) and is common practice in psychophysiological signal analysis. A general methodology for individual diagnosis, based on a concept of person as a bundle of behavioral processes, has been sketched by de Groot (1954). The reader is referred to Molenaar, Huizenga, & Nesselroade (2002) for further elaboration. I do not at all share the reservations of Lord & Novick concerning the possibility of a test theory for individual diagnosis and prediction. However, their position is derived from the premise that there exists a fundamental difference between, on the one hand, a test theory based on BSV and tailored to selection and, on the other hand, a test theory based on WSV and tailored to individual diagnosis, counseling and prediction. I fully endorse this latter premise.

3.4 Heterogeneity in analysis of BSV

In the previous section we encountered heterogeneity (of individual propensity distributions and associated Lord & Novick processes) as a source of nonergodicity. Presently, the issue of heterogeneity will be considered in the more general context of factor analysis of BSV and WSV. It will be shown that factor analysis is surprisingly insensitive to the presence of substantial heterogeneity between persons. A proof of this insensitivity for the standard factor model will be sketched. I also will give some reasons why one could expect heterogeneity to be ubiquitous in natural populations.

Consider the following heterogeneous standard 1-factor model (of BSV):

(3.5) $$\mathbf{y}_i = \lambda_i\eta_i + \varepsilon_i,\ i=1,2,...$$

where λ_i is a p-dimensional vector of factor loadings for person i, η_i is a univariate latent factor, $\eta_i \sim\aleph(\mathbf{0}, \phi_i)$, ε_i is p-variate Gaussian measurement error, $\varepsilon(t) \sim\aleph(\mathbf{0}, \Theta_i)$. It is noted that all parameter matrices in the expression for the covariance matrix associated with (3.5) are person-specific:

(3.6) $$\Sigma_i = \lambda_i\phi_i\lambda_i' + \Theta_i,\ i=1,2,\ldots$$

To ease the presentation, suppose that only the vector of factor loadings λ_i is person-specific, while $\phi_i = \phi$ and $\Theta_i = \Theta$, i=1,2,… Note that $\mathbf{y}_i$ is taken to be centered, for convenience only: $E[\mathbf{y}_i] = \mathbf{0}$. Then it follows from a theorem in Kelderman & Molenaar (2001) that in an analysis of BSV, (3.6) is indistinguishable from the standard homogeneous factor model:

(3.7) $$\Sigma = \lambda\varphi\lambda' + \Theta,\ i=1,2,\ \ldots$$

if the elements of λ_i in (3.6) are independently normally distributed. That is, if $\lambda_i \sim\aleph(\lambda, \text{diag-}\Delta)$, where diag-$\Delta$ denotes the diagonal (p,p)-dimensional covariance matrix of λ_i.

The theorem in Kelderman & Molenaar (2001) is slightly more general. It pertains to a 1-factor model including a person-specific mean vector and person-specific measurement error variances. These additional person-specific parameters can accommodate person-specific propensity distributions in classical test theory (cf. Lord & Novick, 1968, p. 535, eq. 24.3.2). Still the theorem does not cover multi-factor models (in which $q > 1$) and longitudinal factor models. I expect, however, that the theorem can be generalized to cover these cases as well (the proof given in Kelderman & Molenaar, 2001, is a straightforward exercise in the derivation of raw fourth-order moments).

It may come as a surprise that the standard 1-factor model (3.7) will fit data generated by the heterogeneous factor model (3.5)-(3.6). Indeed, it is a basic assumption of standard factor analysis that factor loadings are fixed in the population (invariant over persons). Wholesale violation of this assumption might be expected to lead to misfit of the postulated model. Yet this is not what happens under the conditions of the theorem in Kelderman & Molenaar (2001). And, to reiterate, I expect that under similar conditions it can be proved that standard multifactor models and longitudinal factor models also will fit data generated by their heterogeneous counterparts. The latter expectation has been corroborated in simulation studies reported in Molenaar (1997, 1999). In reality the situation may be even more extreme: data simulated according to person-specific q_i-factor models, i.e., heterogeneous factor models in which the number q_i of latent factors also is person-specific, still can result in acceptably fitting standard q-factor models in which q is small, even if the random variation of, e.g., factor loadings is correlated (nondiagonal Δ). The general picture emerging from these results (and similar ones obtained in related studies; cf. Hamaker, Dolan, & Molenaar, 2002) is that is appears to be very hard to detect the presence of person-specific heterogeneity in standard factor analysis of BSV.

Returning to the heterogeneous 1-factor model defined by (3.6), suppose that the randomly varying factor loadings in λ_i are correlated: $\lambda_i \sim \aleph(\lambda, \Delta)$, where Δ is a full (p,p)-dimensional covariance matrix. It is proven by Kelderman & Molenaar (2001) that in this case the models (3.6) and (3.7) are no longer indistinguishable in an analysis of BSV. This result, however, does not imply that some standard 1-factor model will not yield an acceptable fit to data generated by a heterogeneous factor model with correlated factor loadings. It only implies that the estimated factor loadings thus obtained will not converge to the mean of the random loadings in (3.6). In applications to real data, where the true mechanism of data generation is unknown, this implication does not bring much comfort.

It is noted that the law $\lambda_i \sim \aleph(\lambda, \text{diag-}\Delta)$ allows for arbitrary large variances of the random factor loadings. Hence it may occur that the estimated factor loadings in the standard model (3.7) are all positive, while many of these loadings are vanishing or negative for individual persons. Obviously this will affect the estimation of individual factor scores, as well as the quality of individual assessments based on model (3.7). I fear that there is only one general remedy against this case of ACNE, namely the application of single-subject factor analysis of WSV (Molenaar, 1985). Only in this way sufficiently homogeneous subsets of subjects can be detected in which application of (3.7) is warranted (Nesselroade & Molenaar, 1999). Even then, weak stationarity is required to use the results of application of (3.7) in such homogeneous subsets of persons for valid generalization to individual diagnosis and prediction.

Although it does not belong to the main theme of this book, I would like to take this opportunity to consider some possible sources of the kinds of heterogeneity as discussed in this section (readers not interested in this issue can skip to the next chapter without harm). Are there reasons to expect the presence of substantial heterogeneity with respect to, for instance, the loadings of a factor model of mental test scores? I think there are. Consider the following scenario. Associated with the production of mental test scores is neural activity (the premise of brain imaging and cognitive neuroscience). This neural activity itself takes place in neural networks in the brain, the architecture of which has emerged during embryogenesis, followed by subsequent changes during adaptation to environmental influences. There exist strong converging evidence, both of empirical and theoretical nature, showing that the genesis and adaptation of neural networks is controlled by self-organizing growth processes (so-called nonlinear epigenesis; cf. Molenaar, Boomsma, & Dolan, 1993; Molenaar & Raijmakers, 1999). It is an inherent characteristic of self-organizing epigenetical processes that they generate endogenous variation in their outcomes; variation that is irreducible to (independent of) the effects of genetical and environmental factors (cf. Molenaar, 1986). Hence neural networks emerging during embryogenesis and changing later during adaptation are characterized by irreducible structural variation. For instance, Edelman (1987) shows that the difference between neural connections in the left side and right side of one's brainstem are as large as the difference between the right sides of the brainstem of different subjects.

Given that developmental and learning processes generate endogenous structural variation at the neural level, and given that the production of mental test scores is associated with (supervenes on) the activity of neural networks, one can construct a plausible scenario according to which heterogeneity of the extent and kind as considered above can be expected to exist in a population of persons. The assumption that this heterogeneity at the level of neural architecture manifests itself in, for instance, heterogeneous factor loadings can be defended by an appeal to standard (engineering) interpretations of the components of state-space models (e.g., Padulo & Arbib, 1974). Of course, the factor model is a special instance of the state-space model.

4. Closing remarks

Before trying to take stock of what has been accomplished in the foregoing chapters, I will first discuss some points that require a kind of finishing touch. The first one of these points concerns the relationship between the regression estimator for factor scores and the Kalman filter. It was shown in chapter 1 that for cross-sectional factor models these are the same, but this equality does not simply carry over to longitudinal factor models. In section 4.1 it is explained that the regression estimator for longitudinal factor scores is the same as the Kalman smoother, not the Kalman filter. Another point is more of a definitional nature: in section 4.2 I introduce a pragmatic distinction between state-space models and state-space representations. This will enable us to address some persistent misunderstandings about the relationship between state-space models and dynamic factor models. Then, in the final section 4.3, I will summarize the main results of this book and take the opportunity to speculate a little bit about its possible consequences.

4.1 The Kalman smoother for longitudinal factor models

The discussion in chapter 1 concerning the relationship between the regression estimator of factor scores and the Kalman filter only pertains to cross-sectional factor models. This relationship does not carry over straightforwardly to longitudinal factor models. The reason for this is quite simple. The Kalman filter is an estimator of the schematic form: estimate(t) = f[observation(t) | estimate(t-1)]. Now consider a longitudinal factor model defined at T time points. The Kalman filter estimates of the factor scores at the first time point t=1 are only based on the observations obtained at t=1: $\boldsymbol{\eta}_i(1 \mid 1) = f[\mathbf{y}_i(1) \mid \boldsymbol{\eta}_i(0 \mid 0)]$, where $\boldsymbol{\eta}_i(0 \mid 0)$ denotes (lack of) *a priori* information about initial values. The Kalman filter estimates of the factor scores at the second time point t=2 are only based on the observations obtained at t=1 and t=2: $\boldsymbol{\eta}_i(2 \mid 2) = f[\mathbf{y}_i(2) \mid \boldsymbol{\eta}_i(1 \mid 1)]$. And so forth for t=3,...,T. Only the Kalman filter estimate of the factor scores at t=T is based on the observations obtained at all time points. In contrast, the regression estimator for longitudinal factor scores is based on the supervector $\mathbf{y}_i = [\mathbf{y}_i(1)', ..., \mathbf{y}_i(T)']'$ and therefore makes use of the observations obtained at all time points in estimating the factor scores at each time point. Clearly the statistical performance of this estimator will be better than the performance of its Kalman filter alternative.

The Kalman filter is a recursive estimator that can be applied in real-time because each new observation is processed as soon as it arrives. In applications to longitudinal data, however, all observations are (considered to be) given before the analysis starts and therefore constitute, what Koopmans (1974) calls, historic time series. For such historic time series it makes sense to apply the Kalman filter twice: the usual forward recursion from the initial to the final time point followed by a backward recursion from the final to the initial time point. This is accomplished by means of the fixed interval Kalman smoother. The fixed interval Kalman smoother is used “... when the time interval T of the measurements (i.e., the data span) is fixed, and we seek optimal estimates at some, or perhaps all, interior points. This is the typical problem encountered when processing noisy measurement data off-line” (Brown, 1983, p. 275).

Application of the fixed interval Kalman smoother yields recursive estimates of longitudinal factor scores at each time point t that are based on the observations at all time points: $\boldsymbol{\eta}_i(t \mid T)$, t=1,2,...,T. Dolan & Molenaar (1991) provide a formal proof and evidence from a simulation study, showing that the fixed interval Kalman smoother is identical to the regression estimator for longitudinal factor scores.

It is noted that the need to replace the Kalman filter by the fixed interval Kalman smoother only arises if one wants to obtain regression estimates of longitudinal factor scores in a recursive way, where the (forward and backward) recursions are defined over adjacent time points. In contrast, if all observations are stacked in the supervector $\mathbf{y}_i = [\mathbf{y}_i(1)', ..., \mathbf{y}_i(T)']'$ then the resulting factor model for $\mathbf{y}_i$ constitutes a special instance of a confirmatory oblique factor model. Factor score estimation according to the regression method in the latter confirmatory oblique factor model can proceed by means of the Kalman filter in the way as described in chapter 1. Hence we obtain another equivalence: the recursive fixed interval Kalman smoother for longitudinal factor score estimation, $\boldsymbol{\eta}_i(t \mid T)$, t=1,2,...,T, equals the Kalman filter estimate $\boldsymbol{\eta}_i$ given by (1.7) for factor score estimation in the corresponding confirmatory oblique factor model for the supervector $\mathbf{y}_i$: $\boldsymbol{\eta}_i = [\boldsymbol{\eta}_i(1 \mid T)', ..., \boldsymbol{\eta}_i(T \mid T)']'$. I do not know whether this equivalence has been noticed before in the published literature.

4.2 State-space models and representations

In this section a distinction is made between state-space models and state-space representations. This distinction is not at all fundamental, but it will serve as a heuristic device to discuss some intricacies that have been left implicit in the previous chapters. Possible intricacies that have to do with the use of the concept of state-space model in those chapters.

Until now we have mainly focused on the commonalties between longitudinal factor models (including limiting cases of cross-sectional factor models) and state-space models. The standard longitudinal factor model has a particular form which can schematically represented as:

$$\text{*)} \quad \begin{aligned} &\mathbf{y}_{i,t} = \boldsymbol{\Lambda}_t\boldsymbol{\eta}_{i,t} + \boldsymbol{\varepsilon}_{i,t} \quad i=1,2....; t=1,2,...,T \\ &\boldsymbol{\eta}_{i,t+1} = \mathbf{B}_{t+1,t}\boldsymbol{\eta}_{i,t} + \boldsymbol{\zeta}_{i,t} \quad \boldsymbol{\eta}_{i,1} = \boldsymbol{\zeta}_{i,1} \end{aligned}$$

It is noted that, according to *), at each time point t the p-variate observation $\mathbf{y}_{i,t}$ is only directly influenced by the q-variate longitudinal factor score $\boldsymbol{\eta}_{i,t}$. Hence longitudinal factor scores directly affect only contemporaneous observations; there are no direct relationships between $\mathbf{y}_{i,t}$ and antecedent longitudinal factor scores $\boldsymbol{\eta}_{i,s}$, s < t. There are only indirect relationships with antecedent longitudinal factor scores represented by the latent q-variate AR(1): $\boldsymbol{\eta}_{i,s+1} = \mathbf{B}_{s+1,s}\boldsymbol{\eta}_{i,s} + \boldsymbol{\zeta}_{i,s}$ $\boldsymbol{\eta}_{i,1} = \boldsymbol{\zeta}_{i,1}$, s < t.

The state-space models in this book have been represented by a form similar to *):

$$\text{**)} \quad \begin{aligned} &\mathbf{y}(t) = \boldsymbol{\Lambda}(t)\boldsymbol{\eta}(t) + \boldsymbol{\varepsilon}(t) \quad t=0,\pm1,... \\ &\boldsymbol{\eta}(t+1) = \mathbf{B}(t)\boldsymbol{\eta}(t) + \boldsymbol{\zeta}(t) \end{aligned}$$

Like in *), there are in **) only indirect relationships between $\mathbf{y}(t)$ and $\eta(s)$, $s < t$, which are implied by the latent q-variate AR(1) state process: $\eta(s+1) = \mathbf{B}(s)\eta(s) + \zeta(s)$, $s < t$. I will refer to **) as a state-space *model* and contrast it with a state-space *representation* to be described below.

Both the standard longitudinal factor model *) and the state-space model **) share the same kinds of restrictions. The restriction that the latent q-variate η-process in both models obeys a first-order autoregressive model is not essential: generalization to higher-order autoregressions is straightforward (cf., e.g., Shumway & Stoffer, 2000, Chapter 4; Durbin & Koopman, 2001, Chapter 3). The restriction that the p-variate measurement error process (ε-process) in both models is a white noise process also is not essential. This can be replaced by more general p-variate NARMA alternatives (cf., e.g., Jazwinsky, 1970, Chapter 7). The same remarks can be made with respect to the latent q-variate innovation process (ζ-process) in both models: this also can be replaced by more general q-variate NARMA alternatives (cf., e.g., Jazwinsky, 1970, Chapter 7). Also, the ε–process and the ζ–process can be allowed to be cross-correlated in both models (cf., e.g., Goodwin & Sin, 1984). It is noted in passing that, although all these variants have been elaborated and applied in state-space modeling, they have not all been considered in longitudinal factor analysis. Hence there is ample room for innovative work here.

The restriction that I want to concentrate upon in distinguishing between state-space models and state-space representations concerns the lack of delayed direct influences of the η–process on observations in *) and **). Incorporation of such delayed direct influences in **) can be expressed according to the following schematic representation:

$$\S)\quad \begin{aligned} \mathbf{y}(t) &= \textstyle\sum_u \Lambda_t(u)\eta(t-u) + \varepsilon(t) \quad t=0,\pm1,...;\ u \geq 0 \\ \eta(t+1) &= \mathbf{B}_t\eta(t) + \zeta(t) \end{aligned}$$

where at each time point t $\Lambda_t(u)$, u=0,1,..., denotes a sequence of (p,q)-dimensional matrices of loadings expressing the delayed direct influences of previous realizations $\eta(t-u)$, u=0,1,…, on $\mathbf{y}(t)$. A similar generalization can be given for the longitudinal factor model *), but I will not consider this further.

The model given by §) is no longer a state-space model. It is called a dynamic factor model and for the weakly stationary case in which $\Lambda_t(u) = \Lambda(u)$, u=0,1,..., it is discussed in Brillinger (1975, Chapter 9), Molenaar (1985; 1987) and Molenaar, de Gooijer & Schmitz (1992). The delayed direct effects of the η–process on observations in §) can accommodate differential lead-lag patterns (or phase relationships) between the elements of $\mathbf{y}(t)$. This will be illustrated by the wave model for electrocortical potential fields developed by Nunez (1981; 1995; see also Molenaar, 1993). Suppose that the elements $y_j(t)$, j=1,...,p, of $\mathbf{y}(t)$ consist of registrations of the time-varying electrocortical potential field at different locations l_j, j=1,...,p on the head. Suppose also that $\eta(t)$ denotes the activity of neural sources modulating this potential field through long-range cortico-cortical connections. Then the activity of a neural source, i.e., an element $\eta_k(t)$ of $\eta(t)$ at a particular location L_k in the brain, will manifest itself later in registrations (elements of $\mathbf{y}(t)$) at locations at larger distances from L_k. Stated otherwise, delays in the modulating effect of neural source $\eta_k(t)$, k=1,...,q, on potential field registration $y_j(t)$, j=1,...,p, are distance-dependent, i.e., a function of $|L_k - l_j|$. Moreover, the delayed modulating effect due to the activity of a neural source at time t may persist during a finite interval after onset, t' > t, and thus give rise to aftereffects that overlap with the newly arriving modulating effects induced by the activity of this source at subsequent time points t'. Hence at each time point t and location l_j, the observed potential field $y_j(t)$ consists of a weighted superposition of contemporaneous and delayed effects of latent sources, transmitted with finite velocity along a network of long-range cortico-cortical connections. The biological properties of this connective network give rise to aftereffects, i.e., even if the activity of a neural source would be reduced to a single spike at a fixed time t_s, it still would give rise to delayed effects during a finite time interval t_s+u, $u \geq 0$. This results in intricate patterns of phase relationships between the elements of $\mathbf{y}(t)$ which are captured by the dynamic factor model §).

The state-space model **) is a special instance of a dynamic factor model §) in which the lag u only takes the value u=0. This implies in the context of our example that activity of a neural source $\eta_k(t)$ has instantaneous effects at all registration points across the head (no distance-dependent delays are possible). Hence it is obvious that the dynamic factor model includes the state-space model as a special case. Yet it has sometimes been conjectured that the dynamic factor model is a special instance of the state-space model (e.g., Immink, 1986). Although the latter conjecture is incorrect, there is a sense in which it could be saved. If in §) the sum is over a finite set of values for the lag-index u, it can be shown that the resulting dynamic factor model can be rewritten into state-space form (Molenaar, 1985, Appendix). One then obtains a state-space representation of this dynamic factor model. I prefer to call this a state-space *representation*, because the state vector now is a supervector covering the latent $\boldsymbol{\eta}$-process at a range of time points. Accordingly, a dynamic q-factor model has a state-space representation in which the dimension of the state vector is a multiple of q, depending upon the (presumably finite) range of the lag-index u in §). In contrast, the state-space *model* **) in which the state vector is q-dimensional constitutes the limiting form of the dynamic q-factor model in which the range of the lag-index u in §) only consists of the value u=0. Hence the latter (dynamic factor and state-space) *models* both involve latent processes of the same dimension q.

To reiterate, the distinction between state-space models and representations is not at all fundamental. It only serves to bring some order in the manifold collection of state-space forms. Almost any linear time series model (including the NARMA models discussed in Chapter 2) can be rewritten into state-space form. For practical purposes it therefore is worthwhile to introduce some classification scheme for state-space forms. In fact, this also would be worthwhile to consider for the several possible variants of longitudinal factor models mentioned above.

4.3 General discussion

The major part of this book has been devoted to an introduction of new perspectives on structural equation modeling. These perspectives are inspired by dynamical systems theory, in particular realization theory (chapter 2) and ergodic theory (chapter 3). I have tried to show how each perspective yields new and interesting results, although it has to be admitted that the results obtained thus far are preliminary in most respects. Hence one should consider the discussion in chapters 2 and 3 mainly as invitations to apply realization theory and ergodic theory in the context of structural equation modeling.

Realization theory for linear stochastic systems involves the application of sophisticated linear algebraic ideas and techniques (cf. Caines, chapter 4), often leading to constructive algorithms to compute equivalent representations. Given the formal equivalence between state-space models and structural equation models, these algorithms would seem to have potentially useful applications in the latter realm too. For instance, the special issue on linear systems and control of Linear Algebra and its Applications (1989) contains a wealth of information with possibly interesting applications to structural equation modeling.

The possibility to remove latent variables from structural equation models by means of the invertible (one-to-one) transformations described in chapter 2 needs to be worked out in a much more general fashion. The transformations themselves need to be casted into a transparent linear algebraic format that covers both causal and noncausal NARMA representations. Such a general abstract format is essential in order to accommodate the Houdini transformation to complex structural equation models of the multi-indicator multi-wave varieties. Noncausal NARMA representations would seem to be more natural for cross-sectional latent variable models. I presented a possible interpretation of NARMA equivalents of latent variable models in terms of nearest-neighbor interaction between the elements of the manifest p-variate vector **y**. Because in a cross-sectional setting the order of the elements of **y** is arbitrary, one actually obtains p! distinct NARMA equivalents, and hence one could consider the question whether there exists an optimal ordering in some sense. Moreover, there arises the especially challenging question what happens to factor indeterminacy (cf. Krijnen, 19**) if a factor model is transformed into a NARMA equivalent. Does factor indeterminacy leave behind any trace in the NARMA representations concerned? If so, then the only possible locus for indeterminacy in a NARMA representations is in the residual terms. This then would cast doubts on our usual interpretation of residuals.

Although realization theory is basically an algebraic theory, it does have interesting consequences for statistical estimation. The commonalties and differences between statistical estimation in state-space models and ARMA models are discussed in great depth in Hannan & Deistler (1988). Also, within the class of state-space models, or within the class of multivariate ARMA models, different possible representations (so-called canonical forms) have different consequences for statistical estimation. The implications of these theoretical results for estimation in latent variable models, in comparison with estimation in their NARMA equivalents, still appear to be unexplored.

As I have tried to show in chapter 3, ergodic theory has immediate consequences for the range of applicability of results obtained with structural equation modeling of BSV. The classical ergodicity theorems imply that only under strong restrictive stationarity assumptions can one expect that results obtained with structural equation modeling of BSV will generalize to WSV. Developmental and learning processes do not obey these stationarity assumptions and hence are prime candidates of nonergodic processes. It will be evident that this has major implications for individual assessment, counseling and prediction. Classical test theory is based on analysis of BSV and hence psychological tests constructed according to this theory may not be valid for individual assessment of learning and development. I expect that the same qualifications can be made with respect to the fruits of item-response theory.

In the discussion of classical test theory in chapter 3, the issue of heterogeneity in a population of subjects came up. Heterogeneity is not the same as nonergodicity, but appears to have similar effects in that it precludes the generalization of results obtained in analyses of BSV to WSV. Although parameters in structural equation models of BSV are assumed to be invariant across subjects, it turns out that extreme violation of this assumption is hard to detect. As I indicated in chapter 3, it can even be proven that under certain circumstances this heterogeneity cannot be detected at all in analyses of BSV. It is noted that multilevel models and mixture models do not offer a panacea for the kind of heterogeneity under scrutiny. If factor loadings in a factor model are randomly varying across subjects, then the implied covariance matrix between manifest variables becomes subject-specific. In contrast, the covariance between level and shape in a latent growth model is assumed to be invariant across subjects. Consequently, this latent growth model is not able to accommodate the subject-specific covariances in a factor model with random factor loadings. A mixture model is based on the assumption that there exist homogeneous subpopulations, where each subpopulation is large in some sense. In contrast, a factor model with random factor loadings is akin to a mixture model in which each subpopulation consists of a single subject.

The problems created by nonergodicity and heterogeneity can only be solved by dedicated study of WSV. In psychology such WSV-based approaches are called N=1 techniques and are considered to be somewhat inferior to BSV-based approaches. Psychology is about the only science where WSV is not considered to be a phenomenon worthy of serious study. I will not venture into discussion of the stated (and unstated) reasons for this particular attitude. Suffice it to say that the N=1 paradigm can yield nomothetic theory, i.e., general theory, though the required methodology will be different from the popular N=many approach. Fortunately, the required methodological and statistical tools are available to seriously start curing the chronic psychometric problems caused by ACNE.

Ergodic theory itself has become more and more detached from its origins in statistical mechanics. Modern ergodic theory is part of the general mathematical theory of dynamical systems (Arnold, 1998; Borovkov, 1998; Keller, 1998). Statistical mechanics is moving into entirely new directions in which the arbitrary large classical ensembles (thermodynamic limit) are replaced by small ensembles of interacting particles (e.g., Tsallis, 2002). Some of these developments have already entered the social sciences (Helbing, 1995).

Realization theory and ergodic theory are not the only viable perspectives on structural equation modeling inspired by system theory. Optimal control and nonlinear recursive estimation are also promising candidates. The possibility of optimal control arises if the manifest variables include a subset of so-called control variables that can be manipulated at will. It then is possible to compute the values of the control variables so that a performance criterion is optimized. Whittle (1990) presents an excellent overview of optimal control and its dual relationship with estimation. Molenaar (1987) presents an empirical application to therapeutic process control. To the best of my knowledge, Press (1972) is the only source where optimal control is considered in the context of analysis of BSV.

Nonlinear recursive estimation involves the extension of Kalman filtering techniques to nonlinear dynamical systems (e.g., Sage & Melsa, 1971, chapter 9). To illustrate the use of nonlinear recursive estimation in structural equation modeling, let’s go back to chapter 1 where the relationship was considered between the regression estimator of factor scores in a cross-sectional model and the Kalman filter. Both the regression estimator and the Kalman filter are derived under the assumption that the parameter matrices in a factor model are known. But usually only estimates of these parameters are available. To investigate how the use of uncertain parameter estimates in the regression estimator (Kalman filter) affects its performance, the following approach can be used. Add the uncertain parameter estimates to the state vector, which yields an extended state vector composed of the original factors and the parameter estimates. Note that the state-space model thus extended becomes nonlinear, even if the initial model is linear. Apply nonlinear recursive estimation to the state-space model thus extended to obtain factor score (state) estimates. Preliminary work of mine along these lines shows that the regression estimator (Kalman filter) appears to be quite robust against parameter uncertainty. Nonlinear recursive filtering techniques prove to be quite versatile for addressing this and similar questions.

Structural equation modeling is one of the most challenging and innovative fields in the methodology of social science. I hope that this book provides a stimulus to incorporate system theoretical perspectives to the ongoing theoretical study of structural equation models.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317.

Amit, D.J. (1989). *Modeling brain function: The world of attractor neural networks*. Cambridge: Cambridge University Press.

Anderson, B.D.O., & Moore, J.B. (1979). *Optimal filtering*. Englewood Cliffs, NJ: Prentice-Hall.

Anderson, T.W.A. (1971). *The statistical analysis of time series*. New York: Wiley.

Arnold, L. (1998). *Random dynamical systems*. Berlin: Springer-Verlag.

Bartholomev, D. J. (1987). *Latent variable models and factor analysis*. Oxford: Oxford University Press.

Bekker, P.A., Merckens, A., & Wansbeek, T.J. (1994). *Identification, equivalent models, and computer algebra*. San Diego: Academic Press.

Bellantine, L.E. (1990). *Quantum mechanics*. Englewood Cliffs, NJ.: Prentice Hall.

Bentler, P.M., Lee, S-Y. (1979). A statistical development of three-mode factor analysis. *British Journal of Mathematical and Statistical Psychology, 32,* 87-104.

Bijleveld, C.C.J.H., & van der Kamp, L.J.Th. (1998). *Longitudinal data analysis: Designs, models and methods*. London: Sage.

Billingsley, P. (1995). *Probability and measure* (3rd ed.). New York: Wiley.

Bookstein, F.L. (1990). An interaction effect is not a measurement. *Behavioral and Brain Sciences, 13*, 121-122.

Boomsma, D.I. and Molenaar, P.C.M. (1987). The genetic analysis of repeated measures I: Simplex models. *Behavior Genetics, 17,* 111-123.

Borovkov, A.A. (1998). *Ergodicity and stability of stochastic processes*. Chichester: Wiley.

Box, G.E.P., & Jenkins, G.M. (1970) *Time series analysis, forecasting and control*. San Francisco: Holden-Day.

Brillinger, D.R. (1975). *Time series: Data analysis and theory*. New York: Holt, Rinehart, & Winston.

Brown, R.G. (1983). *Introduction to random signal analysis and Kalman filtering*. New York: Wiley.

Browne, M.W. (1992). Circumplex models for correlation matrices. *Psychometrika, 57,* 469-497.

Caines, P.E. (1988). *Linear stochastic systems*. New York: Wiley.

Cattell, R.B. (1946). *The description and measurement of personality*. New York: World Book.

Cattell, R.B. (Ed.). (1966). *Handbook of multivariate experimental psychology*. Chicago: Rand McNally.

Cohen, L. (1995). *Time-frequency analysis*. Englewood Cliffs, N.J.: Prentice Hall.

Cornfield, I.P., Fomin, S.V., & Sinai, Y, G. (1982). *Ergodic theory*. New York: Springer-Verlag.

Cronbach (1968)

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Annals of Statistics, 25,* 1-37.

Dolan, C.V., & Molenaar, P.C.M. (1991). A note on the calculation of latent trajectories in the quasi-Markov simplex model by means of the regression method and the Kalman filter. *Kwantitatieve Methoden, 38,* 29-44.

Dorfman, J.R. (1999). *An introduction to chaos in nonequilibrium statistical mechanics*. Cambridge: Cambridge University Press.

Drösler. J. (1978). Extending the temporal range of psychometric prediction by optimal linear filtering of mental test scores. *Psychometrika, 43,* 533-549.

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

Edelman, G.M. (1987). *Neural Darwinism: The theory of neuronal group selection*. New York: Basic Books.

Elliott, R.J., Aggoun, L., & Moore, J.B. (1995). *Hidden Markov models: Estimation and control*. New York: Springer-Verlag.

Ellis, J.L., & Junker, B.W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika, 62,* 495-523.

Farquhar, I.E. (1964). *Ergodic theory in statistical mechanics*. London: Wiley.

Fischer, G. (1974). *Einführung in die Theorie psychologischer tests: Grundlagen und Anwendungen*. Bern: Verlag Hans Buber.

Fuhrmann, P.A., Kimura, H., & Willems, J.C. (Eds.). (1989). Linear systems and control [Special issue]. *Linear Algebra and its Applications, Vols. 122-124.*

Gianola, D., Im, S., Fernando, R.L., & Foulley, J.L. (1990). Mixed model methodology and the Box-Cox theory of transformation: A Bayesian approach. In D. Gianola & K. Hammond (Eds.), *Adavances in statistical methods for genetic improvement of livestock* (pp. 15-40). New York: Springer-Verlag.

Goodwin, G.C., & Sin, K.S. (1984). *Adaptive filtering, prediction and control*. Englewood Cliffs, NJ: Prentice-Hall.

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

Gray, R.M. (1988). *Probability, random processes, and ergodic properties*. New York: Springer-Verlag.

Groot, A.D. de (1954). Scientific personality diagnosis. *Acta Psychologica*, *10,* 220-241.

Guttman, Y.M. (1994). *The concept of probability in statistical physics*. Cambridge: Cambridge University Press.

Hannan, E.J. (1970). *Multiple time series*. New York: Wiley.

Hannan, E.J., & Deistler, M. (1988). *The statistical theory of linear systems*. New York: Wiley.

Heck, A. (1996). *Introduction to Maple*. New York: Springer-Verlag.

Helbing, D. (1995). *Quantitative sociodynamics: Stochastic methods and models of social interaction processes.* Dordrecht: Kluwer.

Hewitt, J.K., Eaves, L.J., Neale, M.C., & Meyer, J.M. (1988). Resolving causes of developmental continuity or "tracking" I. Longitudinal twins studies during growth. *Behavior Genetics 18*, 133-151.

Honerkamp, J. (1994). *Stochastic dynamical systems: Concepts, numerical methods, data analysis.* New York: Wiley-VCH Inc.

Immink, W. (1986). *Parameter estimation in Markov models and dynamic factor models*. Unpublished doctoral dissertation. Utrecht: University of Utrecht.

Jazwinsky, A.H. (1970). *Stochastic processes and filtering theory*. New York: Academic Press.

Jenkins & Watts (1968). *Spectral analysis and its applications*. San Francisco: Holden-Day.

Jöreskog, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 32,* 443-482.

Jöreskog, K.G. (1970). Estimation and testing of simplex models. *British Journal of Mathematical and Statistical Psychology, 59,* 121-145.

Jöreskog, K.G. (1979). Statistical estimation of structural models in longitudinal developmental investigations. In J.R. Nesselroade & P.B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 303-351). New York: Academic Press.

Jöreskog, K.G., & Sörbom, D. (1989). *Lisrel VII: A guide to the program and applications.* Chicago: SPSS.

Keilson, J. (1965). *Green's function methods in probability theory*. London: Griffin.

Keller, G. (1998). *Equilibrium states in ergodic theory*. Cambridge: Cambridge University Press.

Kindermann, R., & Snell, J.L. (1980). *Markov random fields and their applications. Contemporary Mathematics* (Vol. 1). Providence, Rhode Island: American Mathematical Society.

Koopmans, L.H. (1974). *The spectral analysis of time series*. New York: Academic Press.

Le Bellac, M. (1991). *Quantum and statistical field theory.* Oxford: Oxford University Press.

Lee, S.Y., & Poon, W.Y. (1993). Structural equation models with hierarchical data. In K. Haagen, D.J. Bartholomev, & M. Deistler (Eds.), *Statistical modeling and latent variables*. Amsterdam: Elsevier.

Lord, F.M. (1963). Elementary models for measuring change. In C.W. Harris (Ed.), *Problems in measuring change* (pp. 21-38). Madison: The University of Wisconsin Press.

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

Lütkepohl, H. (1993). *Introduction to multiple time series analysis.* Berlin: Springer-Verlag.

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

Mandys, F., Dolan, C.V., & Molenaar, P.C.M. (1994). Two aspects of the simplex model: Goodness of fit to linear growth curve structures and the analysis of mean trends. *Journal of Educational and Behavioral Statistics, 19,* 201-215.

Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change. In L.M. Collins & A.G. Sayer (Eds.), *New methods for the analysis* *of change* (pp. 203-240). Washington: American Psychological Association.

Mitra, S.J., & Ekstrom, M.R. (Eds.). (1978). *Two-dimensional digital signal processing*. Stroudsburg: Dowden, Hutchington & Ross.

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

Molenaar, P.C.M. (1986). On the impossibility of acquiring more powerful structures: A neglected alternative. *Human Development, 29,* 245-251.

Molenaar, P.C.M. (1987). Dynamic assessment and adaptive optimisation of the therapeutic process. *Behavioral Assessment, 9,* 389-416.

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

Molenaar, P.C.M. (1993). Dynamic factor analysis of psychophysiological signals. In J.R. Jennings, P. Ackles, & M.G.H. Coles (Eds.), *Advances in Psychophysiology* (Vol. 5, pp. 229-302). London: Jessica Kingsley Publishers.

Molenaar, P.C.M. (1997). Time series analysis and its relationship with longitudinal analysis. *International Journal of Sports Medicine, 18,* 232-237.

Molenaar, P.C.M. (1999). Comment on fitting MA time series by structural equation models. *Psychometrika, 64,* 91-94.

Molenaar, P.C.M. (1999). Longitudinal analysis. In H.J. Ader & G.J. Mellenbergh (Eds.), *Research methodology in the social, behavioural and life sciences* (pp. 143-167). London: Sage.

Molenaar, P.C.M., Boomsma, D.I., & Dolan, C.V. (1993). A third source of developmental differences. *Behavior Genetics, 23,* 519-524.

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

Molenaar, Huizenga, Nesselroade

Molenaar, P.C.M., & Raijmakers, M.E.J. (1999). Additional aspects of third source variation for the genetic analysis of human development and behavior. *Twin Research, 2,* 49-52.

Murray, J.D. (1993). *Mathematical biology* (2nd ed.). Berlin: Springer-Verlag.

Nesselroade, J. R., & Molenaar, P. C. M. (1999). Pooling lagged covariance structures based on short, multivariate time series for dynamic factor analysis. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 223-250). Thousand Oaks, CA: Sage Publications, Inc.

Nunez, P.L. (1981). *Electric fields of the brain: The neurophysics of EEG.* New York: Oxford University Press.

Nunez, P.L. (1995). *Neocortical dynamics and human EEG rhythms.* New York: Oxford University Press.

Opper, M., & Saad, D. (Eds.). (2001). *Advanced mean field methods: Theory and practice.* Cambridge, Mass.: MIT Press.

Padulo, L., & Arbib, M.A. (1974). *System theory: A unified state-space approach to continuous and discrete systems.* Philadelphia, PA: Saunders.

Papoulis, A. (1985). *Signal analysis* (2nd ed.). Auckland: McGraw-Hill.

Parzen, E. (1984). *Time series analysis of irregularly observed data.* New York: Springer.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge: Cambridge University Press.

Petersen, K. (1983). *Ergodic theory.* Cambridge: Cambridge University Press.

Press, S.J. (1972). *Applied multivariate analysis*. London: Holt, Rinehart, & Winston.

Priestley, M.B. (1988). *Non-linear and non-stationary time series analysis.* London: Academic Press.

Priestley, M.B., & Subba Rao, T. (1975). The estimation of factor scores and Kalman filtering for discrete parameter processes. *International Journal of Control, 21,* 971-975.

Rogosa, D., & Willett, J.B. (1985). Satisfying a simplex structure is simpler than it should be. *Journal of Educational Statistics, 10,* 99-107.

Rovine, M.J., & Molenaar, P.C.M. (1998). The covariance between level and shape in the latent growth curve model with estimated basis vector coefficients. *Methods of Psychological Research Online, 3,* 95-108.

Rovine, M.J., & Molenaar, P.C.M. (2000). A structural modeling approach to a multilevel random coefficients model. *Multivariate Behavioral Research, 35,* 51-88.

Rovine, M.J., & Molenaar, P.C.M. (2000). *Addition of latent variables paper*. Submitted.

Sage, A., & Melsa, J. (1971). *Estimation theory with applications to communications and control.* New York: McGraw-Hill.

Shumway, R.H., & Stoffer, D.S. (2000). *Time series analysis and its applications*.

New York: Springer-Verlag.

Sklar, L. (1993). Physics and chance: Philosophical issues in the foundations of statistical mechanics. Cambridge: Cambridge University Press.

Suppes, P., & Zinnes, J.L. (1963). Basic measurement theory. In R.D.Luce, R.R. Bush, & E. Galanter (Eds.), *Handbook of Mathematical Psychology* (Vol. 1, pp. 1-76). New York: Wiley.

Rosenblatt, M. (2000). *Gaussian and non-Gaussian linear time series and random fields.* New York: Springer-Verlag.

Thompson, C.J. (1988). *Classical equilibrium statistical mechanics.* Oxford: Oxford University Press.

Titterington, D.M., Smith, A.F.M., & Makov, U.E. (1985). *Statistical analysis of finite mixture distributions.* New York: Wiley.

Tong, H. (1990). *Non-linear time series: A dynamical system approach.* Oxford: Oxford University Press.

Tsallis, C. (2002). Entropic nonextensivity: a possible measure of complexity. *Chaos, Solitons and Fractals, 13,* 371-391.

Van der Maas, H.L.J., & Molenaar, P.C.M. (1992). Stagewise cognitive development: An application of catastrophe theory. *Psychological Review, 99,* 395-417.

Wohlwill, J.F. (1973). *The study of behavioral development.* New York: Academic Press.

Wooldridge, J.M. (1994). Estimation and inference for dependent processes. In R.F. Engle & D.L. McFadden (Eds.), *Handbook of econometrics* (Vol. IV, pp. 2639-2738). Amsterdam: Elsevier.