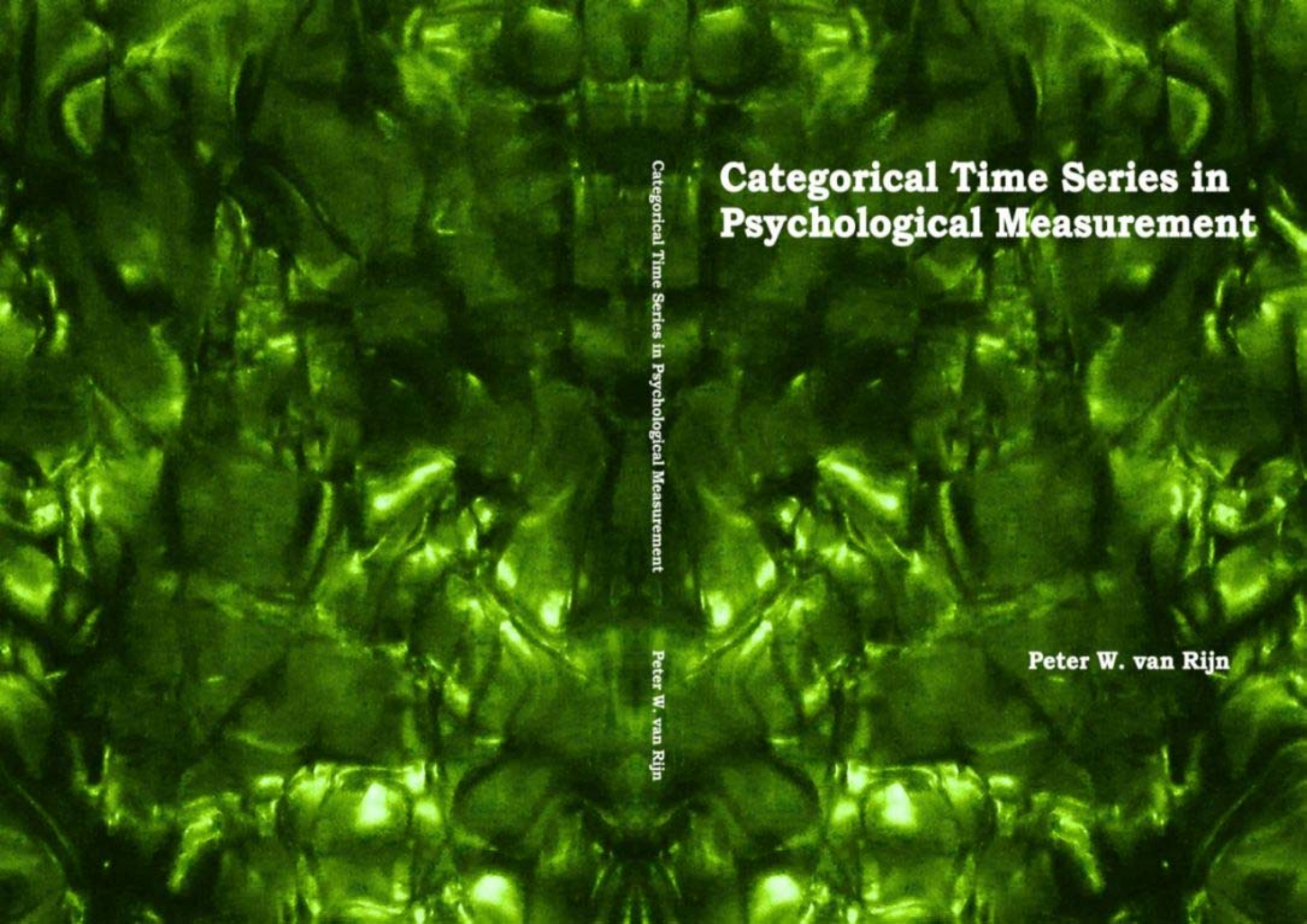
Categorical time series in psychological measurement

van Rijn, P.W.


Categorical Time Series in Psychological Measurement
Peter W. van Rijn
Categorical Time Series in Psychological Measurement
Peter W. van Rijn

Categorical Time Series in Psychological Measurement

Peter W. van Rijn

Categorical Time Series in Psychological Measurement

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties
ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op donderdag 24 april 2008, te 12:00 uur

door

Peter Wilhelmus van Rijn

geboren te Alkmaar

Promotiecommissie

Promotores:	Prof. dr. P.C.M. Molenaar
	Prof. dr. H.L.J. van der Maas
Copromotor:	Dr. C.V. Dolan
Overige leden:	Prof. dr. G.J. Mellenbergh
	Prof. dr. G.K.J. Maris
	Prof. dr. H. Kelderman
	Dr. D. Borsboom
	Dr. E.L. Hamaker

Faculteit der Maatschappij- en Gedragswetenschappen

Dankwoord

Dit proefschrift is niet compleet zonder een dankwoord. Ik wil een aantal mensen bedanken die een voor mij belangrijke rol hebben gespeeld bij de totstandkoming van dit proefschrift. Allereerst was dit proefschrift niet van de grond gekomen zonder mijn promotor Peter Molenaar. Ik ben hem dankbaar voor zijn ideeën en zeer brede wetenschappelijke expertise. Ik heb van hem geleerd om ook eens buiten je eigen vakgebied te kijken. Ik wil mijn promotor Han van der Maas bedanken, met name voor zijn volharding in te vragen wanneer het nu af was. Mijn copromotor Conor Dolan ben ik zeer dankbaar vanwege zijn hulp bij het programmeren, analyseren en schrijven. Zijn commentaar was altijd snel en scherp. Ook ben ik hem dankbaar voor een aantal mooie fietstochten en -routes.

Ik heb veel geleerd van mijn UvA-kamergenoten en met veel plezier op een kamer gezeten met Lourens Waldorp, Raoul Grasman, Denny Borsboom, Maarten Speekenbrink, Michiel Hol, Dave Hessen en Jan-Willem Romeijn. Ik heb gelukkig altijd last van hen gehad.

Ik wil ook de afdeling Psychologische Methodenleer bedanken en de mensen die eraan verbonden zijn of zijn geweest. In het bijzonder Don Mellenbergh, Pieter Koele, Wulfert van den Brink, Jaap van Heerden, Johan Hoogstraten, Marijke Engels, Harry Vorst, Jelte Wicherts, Ellen Hamaker, Niels Smits en ook Ingmar Visser. Het IOPS bedank ik voor het mogelijk maken van het bezoeken van conferenties in Japan en Sardinië.

Ik bedank mijn huidige werkgever Cito voor de mogelijkheid om mijn proefschrift af te ronden, in het bijzonder Piet Sanders en Anton Béguin. Mijn collega's van de afdeling Psychometrisch Onderzoek en Kenniscentrum wil ik ook bedanken. Ik reis met plezier naar Arnhem.

Ernst-Jan Jonkman bedank ik voor een levenslange vriendschap. Ook Maarten van Essen, Sander Borst, Jeroen Mesker en Remko Frijns wil ik bedanken voor hun vriendschap en de nodige afleiding van de wetenschap (Spanje!). Iedereen waarmee ik in bands heb gezeten gedurende het schrijven van dit proefschrift, bedankt: The Skidmarks (Ernesto, Pacobs!, Sander, Jeroen), Check 1-2, Parlandos (Freddie, Coco, Murph, Annie, Costello), the Sherpas (Rocko), en the Dam.

Bij het schrijven van een proefschrift sneuvelt soms een liefde, soms zelfs twee. Ik wil Linda de Boer bedanken voor haar vriendschap. Ndedi Sijsma bedank ik voor haar steun in de moeilijkere fases van de totstandkoming van dit proefschrift.

Dit proefschrift draag ik op aan mijn vader, die er helaas niet meer is, en mijn moeder.

Peter W. van Rijn
Amsterdam, 5 maart 2008

CONTENTS

1

Introduction

Statistical models dealing with latent variables are often used in contemporary psychometrics (Bartholomew & Knott, 1999; Marcoulides & Moustaki, 2002; Skondral & Rabe-Hesketh, 2004; Lee, 2007). Two major fields of psychometrics in which statistical models feature latent variables are factor analysis and item response theory. Factor analysis (FA) was initiated by Spearman's (1904) influential investigations of general mental ability. Currently, factor analytic methods are widely used in the social and behavioral sciences (see Cudeck & McCallum, 2007, for a historical account), and firmly grounded in statistical theory (e.g., Lawley & Maxwell, 1970; Basilevsky, 1994). Item response theory (IRT) developed out of classical test theory (CTT; Gulliksen, 1950), with important contributors as Lord (1952) and Birnbaum (1968) in the United States, and Rasch (1960) in Europe. Both CTT and IRT are discussed with mathematical rigour in the classic treatise of Lord and Novick (1968) with contributions by Birnbaum. Although IRT is applied throughout the social and behavioral sciences, large scale applications are found more frequently in educational measurement settings. Recent accounts and developments of IRT can be found in, e.g., Fischer and Molenaar (1995), Hambleton and van der Linden (1997), and De Boeck and Wilson (2004). Mellenbergh (1994) provides an interesting account of IRT that also refers to FA by making reference to the framework of generalized linear modelling (McCullagh & Nelder, 1989). On a more theoretical level, Borsboom, Mellenbergh, and van Heerden (2003) tackle the nature of latent variables.

In psychology, the latent variable models used in FA and IRT generally concern uncertainties about measurable aspects of variables of interest. Perhaps the most well known method of investigation in psychology is the questionnaire method in which individuals is asked to answer questions that are designed to be indicative of one or more latent psychological variables. In many cases, the uncertainties about measurable aspects originate from the variation that arises when measurements are taken from different, yet exchangeable individuals. In other words, it is likely that individuals respond differently to the posed questions. Hopefully, these observed differences are largely attributable to differences

in the psychological variable of interest and not to other sources. This can be investigated by the fit of the selected latent variable model. A thus obtained latent variable is then composed of variation between individuals. For example, we administer a personality questionnaire to a group of individuals and use a latent variable model in the analysis to compute extraversion scores. If the model fits satisfactorily, these scores can then be interpreted meaningfully, and, e.g., we can conclude that one individual scores higher on extraversion than another. However, it is important to state that such scores are only meaningful in reference to the population from which the group was sampled.

This thesis is concerned with statistical models from the fields of FA and IRT. Its main concern is however not with psychological measurements obtained from different individuals, but with measurements that are repeatedly taken from the same individual at different points in time. That is, this thesis is concerned with measurements that form a time series (e.g., Hamilton, 1994; Lütkepohl, 2005).[1] There are two important differences with the aforementioned situation in which measurements are taken from different individuals. The first important difference is that the measurements are no longer exchangeable, because the particular order in which they arise now plays an important role. More specifically, early measurements can influence later measurements, yet not the other way around. So, the order of the measurements should be accounted for by the selected latent variable model. The second important difference is that after the successful application of such a model, latent scores are to be interpreted only in reference to the studied individual's trajectory instead of to the population from which the individual was sampled. For example, we can only conclude that an individual's extraversion scores are higher now than they were before, and can be predicted to some extent.

In this thesis, special interest goes out to a situation that arises often in psychology, that is, the situation in which the measurements can be classified into only a limited number of categories. To wit, this thesis is mainly about categorical time series. This introductory chapter consists of a short motivation for the investigation of such time series based on developments and focuses in psychology and psychometrics. The present chapter ends with an overview of this thesis.

[1] That is, time series in which time is discrete, and the analyses are conducted in the time domain. In addition, the main focus is on psychological measurements, rather than physiological measurements.

1.1 Psychology

Psychologists have since long been interested in changes of psychological variables over time. An example of short term specific changes can be found in the forgetting curve of Ebbinghaus (1885) that describes the functional relationship between time and the retention of nonsense syllables. A second specific example can be found in the analysis of the schizofrenia symptoms of a single patient over time by Holtzmann (1963). More long term and general changes are described by, for instance, Piaget's theory of cognitive development, which deals with the development of intelligent behaviors in children (Inhelder & Piaget, 1958).

From the inception of psychology as a scientific discipline, psychological researchers have been struggling with the delineation of their research area. A major problem was the formulation of definitions and operationalizations of psychological variables in order to allow for replicable scientific research. Hand in hand with developments in the philosophy of science in the first half of the twentieth century, such as the movement of logical positivism, it was felt that if psychology was to become a true scientific discipline, the onus of psychology should be on empirical research. Psychological research was fitted into fashionable paradigms of scientific method, most notably, the hypothetico-deductive method advocated by Popper (1935). The development of the statistical method of null hypothesis testing , and mathematical statistics in general, by Fisher (1935) and others fitted in tightly with the developments in psychological research.

Taking account of these developments, the focus in psychology on differences between individuals in psychological variables such as general mental ability is not surprising. Van Rijn and Molenaar (Chapter 4 of this thesis; 2005) argue that this focus can also be ascribed to the scientific ideal of general nomothetic knowledge: The theories of scientific psychology should apply to all human individuals. In addition, they argue that this focus is one-sided, and that an ideographic approach, in which differences within a single individual are analysed with statistical methods, deserves to be studied in its own right. Admittedly, the application of such an ideographic approach to psychological research can be problematic for various reasons. For instance, due to repeated nature of the measurements, several kinds of confounding effects such as habituation are likely to occur. More particular, many research questions pertain specifically to differences between individuals, and not within individuals, so that an ideographic approach does not make sense. Although these problems make the reticence of psychologists to pursue ideographic investigations understandable, it is evident that many interesting

research questions in which latent variables are involved are open to ideographic methods of investigation. In addition, ideographic approaches can also provide an interesting contribution to the discussion on theoretical issues concerning latent variables (see Borsboom, Mellenbergh, & van Heerden, 2003). It is not the purpose of this thesis to provide an overview of interesting ideographic topics in psychology. Rather, its purpose is to present a selection of ideographic methods of investigation, an investigation into its possibilities and shortcomings, and eventually compare them with standard psychological methods.

1.2 PSYCHOMETRICS

In 2007, a special issue on psychometrics appeared of the Handbook of Statistics (Rao & Sinharay, 2007). It is illustrative of the lack of interest in time series among present day psychometricians that in the 34 chapter counting volume of well over a thousand pages, only two short sections are concerned with repeated measurements. This is all the more striking, because repeated measurements provided some vexing problems in psychometrics (see, e.g., Harris, 1963). It is not that there are no methods for analysing time series in the field of psychometrics, in fact, they are numerous. For example, Anderson (1963) provides an account of the use of factor analysis for multivariate time series. Molenaar (1985) discusses a method for the analysis of dynamic factors.

The analysis of single individuals is often associated with a less formal side of psychology, and not with the more scientific side in which mathematical models are fitted to psychological measurements. In various other branches of sciences, particularly econometrics, formal analysis of single systems with statistical methods is well developed. It seems that the field of psychometrics is somewhat hesitant when it comes to the analysis of time series in the form of psychological measurements. Yet, the matter of formally approaching the analysis of intra-individual variation has been raised by numerous authors (Hamaker, Dolan, & Molenaar, 2005; Molenaar, 2004; Wood & Brown, 1994; Holtzmann, 1963). Still, this type of analysis has not found a niche in mainstream psychometrics. The purpose of the present thesis is to contribute in its apprehension. Since the numerous examples of analysis of intra-individual variation are mainly concerned with continuous variables and factor analysis, and the fact that there are a lot of psychological measures that can have a discrete nature, the focus in this thesis is on categorical time series and item response theory.

1.3 Overview

Essentially, this thesis can be divided into two parts. The first part consists of chapter two, and the second part of chapters three, four, and five. In the first part, the proclaims of modelling multivariate normal and categorical time series within the framework of structural equation modelling (SEM) are studied. It is concluded that the use of summary statistics in combination with SEM in a time series setting is characterized by certain limitations both with respect to the use of specific models and statistical inference. The application of SEM succeeds only partially and is dependent on the particular models used. Especially for categorical time series, the results are not promising, and there is a need for techniques that use full information. The investigation and evaluation of such procedures is the topic of the second part of this thesis. Towards the end of this thesis, the settting is shifted towards the framework of IRT. The specific focus of each of the chapters is as follows.

Chapter 2 of this thesis concerns an investigation into methods to analyse multivariate normal and categorical time series within the SEM framework. This framework is used because of its familiarity to behavioral researchers and the availability of various standard SEM software packages. In the case of normal time series, it is investigated if the matrix consisting of sample auto- and cross-covariances, referred to as the Toeplitz matrix, can be used to estimate the parameters of various autoregressive models. The use of the Toeplitz matrix and the SEM framework is advantegous, because this matrix is easily computed and can then serve as input for standard SEM software packages to estimate the model. In a simulation study, the performance of a maximum likelihood (ML), weighted least squares (WLS), and, as a reference, Kalman filter (KF) estimation procedure is investigated. For categorical time series, this approach can be used as well. The Toeplitz matrix in this case, however, consists of polychoric auto- and cross-correlations. In a second simulation study, the performance of WLS estimation is investigated in terms of parameter recovery. It was found that the Toeplitz method does not perform properly in all situations, and that, especially for categorical time series, it is advisable to pursue the investigation of filtering methods.

In Chapter 3, univariate categorical time series are analysed within the framework of state space modelling (SSM). In a simulation study, the performance of a Kalman filtering and smoothing procedure is investigated for the estimation of autoregressive models for categorical time series. The discussed procedure is

illustrated by an application to a time series of categorized sleep state measurements.

Chapter 4 of this thesis presents an argumentation in favor of the analysis of intra-individual variation. It is argued that such types of analyses have been neglected in psychology and psychometrics. Having thus set the stage, the second part of this chapter discusses a logistic model for multivariate dichotomous time series that can be seen as a dynamic extension of the ubiquitous Rasch model in IRT. The model is applied to a single-subject multivariate categorical time series consisting of neuroticism scores.

In Chapter 5, the material discussed in Chapter 4 is elaborated on. In particular, extensions to polytomous and multi-subject time series are discussed within the state space framework. Furthermore, it is illustrated by a real data example how this framework can be used for standard applications of IRT as well. The results of applying state space methods are compared with those of standard IRT methods. The chapter ends with an application of the presented models and methods to multi-subject polytomous time series in the form of extraversion scores.

Since the four main chapters comprising this thesis are written with the intention that they form self-contained papers, some material is repeated. Also, abbreviations are introduced in each chapter seperately. The present thesis ends with an epilogue with some general conclusions of the performed investigations, and some guidelines for future research in the field of psychological measurement in the form of categorical time series. The appendix concerns a short note on the application of classical test theory methods in the situation of population heterogeneity.

2

Contributions to the analysis of the block Toeplitz matrix of multivariate stationary time series[1]

2.1 Introduction

The analysis of time series measurements obtained from a single individual has received increasing attention in the behavioral sciences (e.g., Hamaker, Dolan, & Molenaar, 2005; Moskowitz & Hershberger, 2002). Although the majority of research still concerns modelling differences between individuals or groups, it might be argued that the development of models and methods for the study of processes within individuals ought to be pursued with equal effort as to fully grasp the subject matter of the behavioral sciences. However, investigations of models and methods for the analysis of individual time series have concerned univariate rather than multivariate time series, and almost not categorical time series. The purpose of this chapter therefore is to investigate the estimation of models for these types of time series measurements within a framework that is well known to behavioral researchers. Specifically, the fitting of autoregressive moving average (ARMA) models to multivariate time series is examined by means of simulations within the framework of structural equation modelling (SEM). We investigate both procedures for normally distributed time series, and because many observed variables are discrete, a procedure for categorical time series.

ARMA modelling of univariate normally distributed stationary time series may be carried out by a number of methods (Box & Jenkins, 1976; Hamilton, 1994). A distinction can be made between methods which employ the raw data, and methods which make use of summary statistics. At present, the most popular method to estimate the parameters of an ARMA model is maximum likelihood (ML) estimation with the raw data likelihood (Mélard, 1984). This can be performed within the state space framework in a straightforward manner with

[1]This chapter has been conditionally accepted for publication in *Structural Equation Modeling*, and is currently under revision.

Kalman filtering and smoothing techniques, since ARMA models can be formulated as state space models (Durbin & Koopman, 2001; de Jong & Penzer, 2004). Methods using summary statistics generally start out with computing sample auto- and cross-covariances, and then fit the covariance structure of an ARMA model to a matrix of Toeplitz form containing these sample covariances. Instead of using as input for the analyis the matrix containing all auto- and cross-covariances of the observed time series which dimension increases with the length of the time series, one can limit the analysis to no more than the first few lags if one is willing to assume stationarity (Molenaar, 1985). The number of lags analysed is referred to as the window size, and is selected on the basis of the type and the order of the model being fitted. The matrix thus obtained is of Toeplitz or block Toeplitz form, and is therefore often referred to as the Toeplitz matrix. Since the covariance structure of ARMA models can be written as structural equation models (see van Buuren, 1997), such analyses can be performed using software packages designed for SEM, such as LISREL (Jöreskog & Sörbom, 1999). The use of SEM programs enables one to generalize the procedure in a straightforward manner to multivariate data and to multiple-case analyses (Molenaar, 1985; Molenaar, de Gooijer, & Schmitz, 1992). In the present chapter, the performance of model estimation with the Toeplitz matrix is investigated within the SEM framework for several situations pertaining to multivariate stationary time series.

Since most SEM programs are not specifically equipped for the analysis of time series, two issues concerning the special structure of time series and their autocovariances need to be addressed. The first issue concerns the time-dependent structure of time series observations. In standard SEM, where the aims of model fitting include statistical inference, an observed covariance matrix serves as input for the analysis. All estimators in SEM, which allow for statistical inference, require that this matrix is computed with independent and identically distributed observations. This is not the case when using as input a block Toeplitz matrix estimated from an observed multivariate time series. In fact, the very thing of interest in time series analysis is the dependence between sequential observations. In addition, by using only the first few auto- and cross-covariances for the estimation of models, one assumes beforehand that stationarity holds.

Secondly, given the sequential dependence, we need to ascertain the asymptotic properties of sample auto- and cross-covariances in order to fully understand the SEM approach to ARMA modelling. For independent identically distributed multivariate normal data, the maximum likelihood estimator of the covariance

matrix is asymptotically unbiased and efficient, and its distribution is known. For time series observations, unbiased estimation of the theoretical Toeplitz matrix does not pose a problem. However, asymptotic efficiency of the estimator depends on the type and the order of the model. In other words, the Cramér-Rao lower bound is not reached as the length of the time series increases without bound for certain models (Porat, 1987; Kakizawa, 1999). If asymptotic efficiency does not hold, the asymptotic properties of normal theory maximum likelihood estimation cannot be obtained, and this way lead to incorrect statistical inferences from the data.

For univariate ARMA processes, Porat (1987) showed that sample autocovariances are asymptotically efficient only for p-th order autoregressive and q-th order moving average (ARMA(p, q)) processes up to lag p, if $p \geq q$. For vector ARMA processes, Kakizawa (1999) obtained related results in that the sample auto- and cross-covariances of vector AR(p) processes are asymptotically efficient up to lag p. However, regardless of lag, sample auto- and cross-covariances of vector MA processes are asymptotically inefficient. The problem remains unanswered for vector ARMA(p, q) processes, except that it can be shown that if $q > p$, asymptotic efficiency does not hold (Kakizawa, 1999). In the case of multiple indicator ARMA or process factor models (Browne & Nesselroade, 2005), asymptotic results are not available. The absence of full asymptotic results hinders an investigation of the Toeplitz-SEM approach to ARMA modelling, since it cannot be said that the outcomes of such an investigation are due either to the quality of the sample block Toeplitz matrix or to the Toeplitz-SEM approach. For that reason, we restrict ourselves initially to situations in which unbiased and efficient estimation of auto- and cross-covariances is feasible. That is, we focus on autoregressive models. In potential, vector ARMA models can be analysed as well. However, since the necessary window size for this type of models is larger than the lag at which asymptotic efficient estimates are considered possible, it is not pursued here.

The Toeplitz-SEM approach to ARMA modelling is not new, and the available results obtained with this approach are the following. Despite the fact that the Toeplitz-SEM approach uses less information, Hamaker, Dolan, and Molenaar (2002; see also van Buuren, 1997) found that moment estimates asymptotically resemble raw data ML estimates in the case of univariate autoregressive (AR) models, which is consistent with the results of Porat (1987). In the case of moving average (MA) and ARMA models, they found that the moment estimates are not raw data ML estimates. Hamaker et al. (2002) showed that the SEM estimates

are moment estimates, and demonstrated that for AR models the standard errors and χ^2 goodness of fit indices were quite accurate. In ARMA and MA models, the accurary of the results varied somewhat from model to model. Molenaar and Nesselroade (1998) performed a small scale simulation study of the Toeplitz-SEM approach to compare the performance of ML and a weighted least squares (WLS) estimation procedure in the case of multivariate time series (see also Molenaar, 1985). After selecting the maximum lag of the Toeplitz matrix by the Bayesian information criterion computed from fitting vector AR models of increasing order, they fitted a dynamic 1-factor model with autocorrelated errors in LISREL. Although only one instance of the dynamic factor model was used, their results are comparable to Hamaker et al. (2002) in that the parameter estimates are quite accurate for both the ML and WLS method. However, the standard errors and χ^2 goodness of fit indices provide an unclear picture when compared to standard deviations of estimates and expected values, respectively.

Taking account of the above arguments, the Toeplitz method remains a fast and easy method to fit models to multivariate time series. This is mainly because it can be performed in most standard SEM software packages. The software package DyFA deserves attention here, because it is specifically designed for dynamic factor analysis with the Toeplitz method (Browne & Zhang, 2005). DyFA is not used here though, because the present investigation is within the general framework of SEM. An important question to be answered in this paper is how the Toeplitz method performs in a variety of situations in which it can possibly perform well. The first interest then lies in assessing the quality of the parameter estimates of the models used. In the present paper, this assessment is performed by means of simulations in which estimated parameters are compared to their true values in terms of accuracy and precision. In connection to Molenaar and Nesselroade (1998), we use ML and WLS estimation procedures. Since the Toeplitz method uses summary statistics, thereby relying on limited information, it is of general interest to establish the difference between limited and full information methods. A more specific question concerns the possibility of obtaining results with summary statistics that are of comparable quality as results obtained with raw data methods. To adress this question and assess the size of the difference, we compare the outcomes obtained with summary statistics, that is, the results of the Toeplitz method, with the outcomes of a standard normal theory ML based Kalman filter (Harvey, 1989).

Observed variables in the behavioral sciences often have a discrete nature, therefore we also address the Toeplitz method for multivariate categorical time

series. For our situation, this can be achieved by fitting AR models to the matrix of polychoric auto- and cross-correlations. An advantage is that the implementation of this method is relatively easy for those familiar with categorical data analysis within SEM. However, the performance of this method is unknown and therefore studied here. Furthermore, in combination with the results of the normally distributed time series, more insight might be obtained in the Toeplitz-SEM approach.

The outline of this chapter paper is as follows. First, we discuss AR models for multivariate normal time series. Then, the estimation of auto- and cross-covariances, and the construction of the Toeplitz matrix are addressed. This is followed by the representation of the models in the SEM and state space framework. Next, the maximum likelihood, weighted least squares, and Kalman filtering estimation procedures are presented. Subsequently, we discuss a simulation study to the methods presented. Further, the Toeplitz method for categorical time series is discussed, followed by a second simulation study. The chapter ends with a discussion.

2.2 Aspects of autoregressive modelling

We discuss two different types of multivariate autoregressive models. The first type consists of pure vector autoregressive (VAR) models. VAR models can be useful when the dimension of the observed time series is not too large (since the number of parameters increases rapidly with the series dimension). For a more detailed discussion of this type of models, the reader is referred to Hamilton (1994) or Lütkepohl (1991, 2005). The second type of models is multiple indicator vector autoregressive (MI-VAR) models. MI-VAR models consist of latent autoregressions to which the observed time series are related by means of a linear factor model (Hamaker et al., 2005). This type of models is useful when the observed time series can be described by a latent process of smaller dimension (Nesselroade, McArdle, Aggen, & Meyers, 2002).

2.2.1 Multivariate autoregressive models

To ease the presentation, let y_t denote both a multivariate stochastic process and an observed time series thereof. An n-variate zero mean Gaussian autoregressive process of order p can then be imposed on y_t by

$$y_t = \Phi_1 y_{t-1} + \ldots + \Phi_p y_{t-p} + \xi_t, \qquad \xi_t \sim \mathrm{N}(0, \Sigma_\xi) \tag{2.1}$$

where $n \times n$ matrices $\Phi_1, \ldots, \Phi_p$ contain autoregressive parameters, and ξ_t is a multivariate Gaussian white noise sequence with covariance matrix Σ_ξ. The above equation can be rewritten by making use of the backshift operator B as follows

$$(I - \Phi_1 B - \ldots - \Phi_p B^p) y_t = \Phi(B) y_t = \xi_t.$$

Stationarity of the process y_t is obtained if all roots of the polynomial equation $|\Phi(B)| = 0$ lie outside the unit circle (Lütkepohl, 1991, p. 12).

In MI-VAR models, the autoregressive process is not directly observed, yet indicated by an observed time series. An n-variate zero mean stochastic process y_t can be modelled by a latent Gaussian autoregressive process through a linear factor model given by

$$y_t = \Lambda \alpha_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, \Sigma_\varepsilon), \tag{2.2}$$

where Λ is a $n \times m$ matrix of factor loadings and α_t denotes an m-dimensional latent vector autoregressive process of order p given by

$$\alpha_t = \Phi_1 \alpha_{t-1} + \ldots + \Phi_p \alpha_{t-p} + \xi_t, \qquad \xi_t \sim \mathrm{N}(0, \Sigma_\xi). \tag{2.3}$$

Both ε_t and ξ_t are multivariate white noise sequences with covariance matrices Σ_ε and Σ_ξ, respectively. It is noted that the MI-VAR models are not identified without restrictions. What restrictions are necessary for identification and rotational determination is discussed later on together with the SEM representation of the above models. Stationarity holds for the latent process α_t if again the roots of the polynomial equation $|\Phi(B)| = 0$ lie outside the unit circle. It is stressed that stationarity of the observed process y_t is necessary in order to justify the use of only the first few auto- and cross-covariances for parameter estimation purposes. Therefore, we only consider models in which Λ and Σ_ε lack time dependence, in which case stationarity of the observed time series follows from stationarity of the latent process.

2.2.2 Auto- and cross-covariances

The $n \times n$ auto- and cross-covariance matrix of a zero mean stationary Gaussian process y_t at lag u is denoted by Γ_u and given by

$$\Gamma_u = \mathrm{E}(y_t y'_{t-u}), \quad u = 0, 1, 2, \ldots.$$

It can be shown that $\Gamma_u = \Gamma'_{-u}$. We present only the first p auto- and cross-covariances of both types of autoregressive models, since that is sufficient for our

purposes. For a VAR(p), the auto- and cross-covariance matrix at lag u can be written as

$$\Gamma_u = \begin{cases} \Phi_1\Gamma_1' + \ldots + \Phi_p\Gamma_p' + \Sigma_\xi, & \text{if } u = 0, \\ \Phi_1\Gamma_{u-1} + \ldots + \Phi_p\Gamma_{u-p}, & \text{if } u = 1, 2, \ldots, p. \end{cases} \tag{2.4}$$

The above equations are referred to as the Yule-Walker equations (Lütkepohl, 1991, p. 21). For a MI-VAR(p), this can be written as follows

$$\Gamma_u = \begin{cases} \Lambda(\Phi_1\Gamma_1^{*\prime} + \ldots + \Phi_p\Gamma_p^{*\prime} + \Sigma_\xi)\Lambda' + \Sigma_\varepsilon, & \text{if } u = 0, \\ \Lambda(\Phi_1\Gamma_{u-1}^* + \ldots + \Phi_p\Gamma_{u-p}^*)\Lambda', & \text{if } u = 1, 2, \ldots, p, \end{cases}$$

where Γ_u^* is the $m \times m$ auto- and cross-covariance matrix at lag u of the latent VAR(p) process α_t.

An estimate of the auto- and cross-covariance matrix at lag u of a zero mean stationary Gaussian process y_t can be obtained by (Lütkepohl, 1991, p. 79)

$$C_u = \frac{1}{T-u} \sum_{t=u+1}^{T} y_t y_{t-u}'. \tag{2.5}$$

One can also choose to divide by T instead of $T - u$. In that case, the estimator is biased, but has smaller mean square error (Jenkins & Watts, 1968, p. 179). In the above estimator, the mean square error increases as u approaches T. For fixed values of u, however, both estimators have equal asymptotic properties (Hannan, 1970, p. 208). It should also be noted that division by T leads to a non-negative definite block Toeplitz matrix, which in general is a desirable property (Shumway & Stoffer, 2006, p. 30). However, we make use of the above estimator here, because we only discuss situations in which u is small compared to T.

In order to fit a multivariate autoregressive model of order p, the symmetric sample block Toeplitz matrix S_y is constructed as follows

$$S_y = \begin{bmatrix} C_0 & & & \\ C_1 & C_0 & & \\ \vdots & \vdots & \ddots & \\ C_p & C_{p-1} & \cdots & C_0 \end{bmatrix}.$$

For zero mean stationary Gaussian processes, the auto- and cross-covariance estimator has a known normal sampling distribution (Hannan, 1970, p. 208 ff.).

However, as mentioned, this estimator is not asymptotically efficient in all situations. More specifically, the estimator is unbiased, but does not have minimum variance. For instance, a VAR(p) process has an asymptotically efficient sample autocovariance matrix C_u only up to lag p (Kakizawa, 1999). For MI-VAR models, no results concerning the asymptotic efficiency of sample autocovariances are available, and an investigation thereof would require a seperate study. Yet, we discuss MI-VAR models, since they are of practical interest, especially when the observed dimension is large and y_t is amenable to factor analytic modelling. We stress that these asymptotic results twarth the interpretation of outcomes of fitting time series models with auto- and cross-covariances, and are to be kept in mind in the remainder of this chapter.

2.2.3 SEM AND STATE SPACE REPRESENTATION

To represent the discussed models in SEM form, we chose the LISREL Submodel 3B (Jöreskog & Sörbom, 1996). This model is

$$y = \Lambda\eta + \varepsilon, \qquad \varepsilon \sim \mathrm{N}(0, \Theta_\varepsilon), \tag{2.6}$$

$$\eta = B\eta + \zeta, \qquad \zeta \sim \mathrm{N}(0, \Psi). \tag{2.7}$$

The model in the top equation is referred to as the measurement model in which Λ is an $n \times m$ matrix of factor loadings, η is an m-dimensional vector of latent factors, and ε is a multivariate normally distributed random vector with $n \times n$ diagonal covariance matrix Θ_ε. The bottom equation is referred to as the structural relation, where B is an $m \times m$ matrix with zeroes on the diagonal. It is assumed that ε, η, and ζ are uncorrelated. Then, the $n \times n$ covariance matrix of y denoted by Σ_y is modelled by

$$\Sigma_y = \Lambda(I - B)^{-1}\Psi(I - B)^{-1'}\Lambda' + \Theta_\varepsilon.$$

A bivariate VAR of order 1 can now be represented in the SEM framework as follows. Let $\Lambda = I$, $\Theta_\varepsilon = 0$, and

$$y = \begin{bmatrix} y_{1t} \\ y_{2t} \\ y_{1,t-1} \\ y_{2,t-1} \end{bmatrix}, \quad t = 2, \ldots, T.$$

The covariance matrix Σ_y containing the auto- and cross-covariances up to lag 1

is of block Toeplitz form and equals the simplified model structure

$$\Sigma_y = \begin{bmatrix} \Gamma_0 & \\ \Gamma_1 & \Gamma_0 \end{bmatrix} = (I - B)^{-1}\Psi(I - B)^{-1'}, \tag{2.8}$$

where

$$B = \begin{bmatrix} 0 & 0 \\ \Phi_1 & 0 \end{bmatrix} \quad \text{and} \quad \Psi = \begin{bmatrix} \Gamma_0 & \\ 0 & \Sigma_\xi \end{bmatrix}.$$

Other VAR models can be easily put into SEM form by appropriately adjusting the model vectors and matrices. Writing out the above equation results in

$$\begin{bmatrix} \Gamma_0 & \Gamma_0\Phi_1' \\ \Phi_1\Gamma_0 & \Phi_1\Gamma_0\Phi_1' + \Sigma_\xi \end{bmatrix}.$$

The above relations constitute the Yule-Walker equations. It can be shown that all VAR models which are put in SEM form in this manner result in Yule-Walker relations. Regardless of the specific fit function, the resulting estimates are solutions to the Yule-Walker equations, and can be considered Yule-Walker estimates (Hamaker et al., 2002). However, fit functions can differ in their use of auto- and cross-covariances and possibly asymptotic covariances, and these different uses can lead to different solutions. In addition, it should be noted that Yule-Walker estimates are similar, but not identical to least squares estimates (see Lütkepohl, 1991, p. 65). It can be verified that for the above VAR(1) model this approach results in $\hat{\Phi}_1 = C_1C_0^{-1}$ and $\hat{\Sigma}_\varepsilon = C_0 - C_1C_0^{-1}C_1'$. Yule-Walker estimates are known to be biased, especially in the case of strongly autocorrelated processes (Tjøstheim & Paulsen, 1983). It is noted that full VAR models are saturated, an overall likelihood ratio test statistic is zero, and fit statistics can therefore not be obtained. Restricted VAR models, such as models in which Σ_ξ is diagonal, can be represented in an analogous manner, and do provide fit statistics, e.g., likelihood ratio test statistics, i.e., a χ^2 or noncentral χ^2.

As a second example, a latent bivariate VAR(1) with four indicators is put into SEM form in the following manner. For $t = 2, \ldots, T$, let

$$y = \begin{bmatrix} y_{1t} \\ y_{2t} \\ y_{3t} \\ y_{4t} \\ y_{1,t-1} \\ y_{2,t-1} \\ y_{3,t-1} \\ y_{4,t-1} \end{bmatrix}, \Lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} & 0 & 0 \\ \lambda_{21} & \lambda_{22} & 0 & 0 \\ \lambda_{31} & \lambda_{31} & 0 & 0 \\ \lambda_{41} & \lambda_{42} & 0 & 0 \\ 0 & 0 & \lambda_{11} & \lambda_{12} \\ 0 & 0 & \lambda_{21} & \lambda_{22} \\ 0 & 0 & \lambda_{31} & \lambda_{32} \\ 0 & 0 & \lambda_{41} & \lambda_{42} \end{bmatrix}, \eta = \begin{bmatrix} \alpha_{1t} \\ \alpha_{2t} \\ \alpha_{1,t-1} \\ \alpha_{2,t-1} \end{bmatrix}, \text{ and } \quad \Theta_\varepsilon = \operatorname{diag} \begin{bmatrix} \sigma^2_{\varepsilon_1} \\ \sigma^2_{\varepsilon_2} \\ \sigma^2_{\varepsilon_3} \\ \sigma^2_{\varepsilon_4} \\ \sigma^2_{\varepsilon_1} \\ \sigma^2_{\varepsilon_2} \\ \sigma^2_{\varepsilon_3} \\ \sigma^2_{\varepsilon_4} \end{bmatrix}.$$

The block-Toeplitz auto- and cross-covariance matrix Σ_y is modelled by

$$\Sigma_y = \begin{bmatrix} \Gamma_0 & \\ \Gamma_1 & \Gamma_0 \end{bmatrix} = \Lambda(I - B)^{-1}\Psi(I - B)^{-1'}\Lambda' + \Theta_\varepsilon,$$

where

$$B = \begin{bmatrix} 0 & \\ \Phi_1 & 0 \end{bmatrix} \text{ and } \Psi = \begin{bmatrix} \Gamma_0^* & \\ 0 & \Sigma_\xi \end{bmatrix}.$$

The above model is not identified. Scaling of the latent process α_t is necessary, and can be performed by placing restrictions on elements of Λ or Ψ. An additional constraint is required on Λ to fix rotational indeterminacy. In principle, this can be carried out by fixing one of the elements of Λ to be estimated at zero (see Millsap, 2001). Again, other multiple indicator VAR models are easily obtained after appropriate adjustments.

The specification of the above models in state space form can be performed along similar lines. For completeness, a simple linear state space model consisting of an observation model and a transition model is given

$$y_t = Z\alpha_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathrm{N}(0, \Sigma_\varepsilon), \tag{2.9}$$

$$\alpha_t = T\alpha_{t-1} + \zeta_t, \qquad \zeta_t \sim \mathrm{N}(0, \Psi), \tag{2.10}$$

with Z an $n \times m$ design matrix, T an $m \times m$ transition matrix, $\alpha_0 \sim \mathrm{N}(a_0, P_0)$, and disturbances ε_t and ζ_t uncorrelated with each other and α_0. Hamilton (1994), de Jong and Penzer (2004), and Durbin and Koopman (2001) describe state space representations of VAR and other models in greater detail. In closing this section, it should be noted that all autoregressive models considered can be represented in SEM or state space form in various ways. We have chosen the representation in which the latent process is of minimal dimension.

2.2.4 Estimation

Three procedures for the estimation of the parameters of multivariate autoregressive models are compared, namely maximum likelihood (ML) estimation, weighted least squares (WLS) estimation, and Kalman filtering (KF). ML and WLS estimation employ the information contained in the auto- and cross-covariances, whereas KF uses the information of the raw data. All three methods condition on the first p observations, which are sometimes referred to as presample values (Lütkepohl, 1991). With ML and WLS estimation, the difference

between the sample and model block Toeplitz matrix is minimized. This difference is specified in a fit function F which is then minimized by appropriately adjusting the model parameter estimates until convergence. For ML estimation, the fit function that is minimized is given by

$$F_{\mathrm{ML}} = \log|\Sigma_y| + \mathrm{tr}(S_y\Sigma_y^{-1}) - \log|S_y| - n, \tag{2.11}$$

where $|.|$ denotes the determinant and $\mathrm{tr}(.)$ denotes the trace. This is a likelihood ratio. For MI-VAR models, the χ^2 fit statistic can be obtained by computing $(T-p-1)F_{ML}$ at the solution.

The WLS fit function is specified by

$$F_{\mathrm{WLS}} = \mathrm{vech}(S_y - \Sigma_y)'W^{-1}\mathrm{vech}(S_y - \Sigma_y), \tag{2.12}$$

where the vech(.) operator vectorizes the distinct elements for symmetric matrices. The χ^2 statistic in this case can be obtained by computing $(T-p-1)F_{\mathrm{WLS}}$ at the solution. The $\frac{1}{2}n(n+1) \times \frac{1}{2}n(n+1)$ estimated asymptotic covariance matrix of S_y denoted by W is for independent normally distributed variables given by

$$W = 2D^{+}(S_y \otimes S_y)D^{+'}, \tag{2.13}$$

where D^{+} is the Moore-Penrose inverse of the $n^2 \times \frac{1}{2}n(n+1)$ D duplication matrix given by

$$D\mathrm{vech}(S_y) = \mathrm{vec}(S_y),$$

where the vec(.) operator vectorizes the elements for any matrix (see Magnus & Neudecker, 1999). Although this estimator of the asymptotic covariance is generally not suitable for time series, it is used here. The correct asymptotic covariance for any two elements c_{uij} and c_{vkl} of the sample auto- and cross-covariance matrices C_u and C_v is given in Hannan (1970, p. 209) by

$$\mathrm{Cov}(c_{uij}, c_{vkl}) = \frac{1}{T}\sum_{r=-T+1}^{T-1}\left(1 - \frac{|r|}{T}\right)(\gamma_{rik}\gamma_{r+v-u,jl} + \gamma_{r+v,il}\gamma_{r-u,jk}),$$

where $i, j, k, l = 1, \ldots, n$, neglecting fourth cumulants due to the time series being normally distributed. Reiterating that estimates of Γ_u when $u > p$ for VAR(p) models are not asymptotically efficient (Kakizawa, 1999), these estimates are needed to produce an estimate of the asymptotic covariance for WLS estimation. An alternative method to obtain an estimate of the asymptotic covariances is discussed in Molenaar and Nesselroade (1998). They first fit a VAR model selecting the order on the basis of the Bayesian information criterion (BIC), after

which they determine the asymptotic covariances according to the VAR model parameters. In both cases, we know that asymptotic efficiency of the estimates is not guaranteed. For now, therefore, we prefer the much simpler estimator W given in Equation 2.13. The ML and WLS estimation procedure as decribed here are implemented in the LISREL computer program (Jöreskog & Sörbom, 1999).

The fit function minimized with the Kalman filter for the model in Equations 2.9 and 2.10 is given by (Harvey, 1989)

$$F_{\text{KF}} = \tfrac{1}{2}\sum_{t=1}^{T}\log|G_t| + \tfrac{1}{2}\sum_{t=1}^{T} e_t' G_t^{-1} e_t, \tag{2.14}$$

where

$$\begin{aligned} e_t &= y_t - Z a_{t|t-1}, \\ G_t &= Z P_{t|t-1} Z' + \Sigma_\varepsilon. \end{aligned}$$

Now, parameter estimation proceeds iteratively as follows. After starting values for all parameters are specified, the Kalman filter is applied to obtain estimates of the state vector denoted by $a_{t|t-1}$ and associated covariance matrix $P_{t|t-1}$. For $t = 1, \ldots, T$, the Kalman filtering recursions consist of a prediction and correction step respectively given by (Harvey, 1989)

$$\begin{aligned} a_{t|t-1} &= T a_{t-1}, & P_{t|t-1} &= T P_{t-1} T' + \Psi, \\ a_t &= a_{t|t-1} + P_{t|t-1} Z' G_t^{-1}(y_t - Z a_{t|t-1}), & P_t &= P_{t|t-1} - P_{t|t-1} Z' G_t^{-1} Z P_{t|t-1}. \end{aligned} \tag{2.15}$$

The above recursions need to be initialised with values for a_0 and P_0. We chose a practical diffuse prior with initial states fixed at zero and initial variances fixed at a value of 10 (for exact diffuse initialisation, see de Jong, 1991). In the next step, the fit function is minimized to obtain new parameter estimates. The procedure is repeated upon convergence. The KF estimation procedure is implemented in the MKF computer program (Dolan, 2005).

2.3 Simulation study I

In this simulation study, we investigate the Toeplitz method applied to normally distributed time series with the maximum likelihood and weighted least squares estimation procedures described in the previous section. For comparison, the simulated raw data were analysed with the Kalman filter as well.

2.3.1 Set up

Six different autoregressive models are used to simulate time series. These are a MI(2)-AR(1), MI(2)-AR(2), VAR(1), MI(4)-AR(1), MI(4)-AR(2), and MI(4)-VAR(1). Table 2.1 shows the dimension of the observed time series, the dimension of the autoregressive process, the order of the autoregression, and the designations of the six models used. Time series are simulated with two different lengths, $T = 100$ and $T = 1000$. The number of replications in each condition equals 1000. Parameter values are set so that stationarity is obtained and the time series have zero mean and unit variance (at least approximately). Actual parameter values can be found in the tables with results.

Table 2.1: Dimensions and orders of autoregressive models in simulation study

Dim. observed (n)	Dim. process (m)	Order (p)	Model name
2	1	1	MI(2)-AR(1)
2	1	2	MI(2)-AR(2)
2	2	1	VAR(1)
4	1	1	MI(4)-AR(1)
4	1	2	MI(4)-AR(2)
4	2	1	MI(4)-VAR(1)

The main interest of the simulation study lies in the comparison of parameter recovery for the three discussed methods. Parameter estimates are evaluated with respect to accuracy and precision. To this end, mean parameter estimates are compared to true values and mean standard errors of parameter estimates are compared to the standard deviation of parameter estimates over replications.

2.3.2 Results

Since much of the results obtained in different conditions of the simulation study is similar, not all results are displayed. Most noticably, the results obtained with the MI(2)-AR(1) and MI(2)-AR(2) models are so similar to those obtained with MI(4)-AR(1) and MI(4)-AR(2) models, that they are not shown. For the MI(4)-VAR(1) models, the outcomes of $T = 100$ are not shown.

Table 2.2 displays the results obtained for the VAR(1) model for $T = 100$ and $T = 1000$. Results for the ML, WLS, and KF method are almost equal. Parameter estimates are close to true values and mean standard errors resemble

TABLE 2.2: RESULTS FOR VAR(1) MODEL

			ML			WLS			KF		
T	Parameter	Value	Mean[1]	SE[2]	SD[3]	Mean	SE	SD	Mean	SE	SD
100	ϕ^1_{11}	0.80	0.774	0.067	0.069	0.774	0.067	0.069	0.776	0.066	0.070
	ϕ^1_{12}	0.10	0.098	0.066	0.070	0.098	0.066	0.070	0.105	0.066	0.071
	ϕ^1_{21}	0.30	0.303	0.076	0.079	0.303	0.076	0.079	0.312	0.075	0.080
	ϕ^1_{22}	0.60	0.573	0.075	0.074	0.573	0.075	0.074	0.570	0.075	0.076
	$\sigma^2_{\xi_1}$	0.25	0.255	0.036	0.038	0.249	0.036	0.039	0.248	0.035	0.036
	$\sigma^2_{\xi_2}$	0.33	0.330	0.047	0.047	0.322	0.046	0.047	0.323	0.046	0.045
1000	ϕ^1_{11}	0.80	0.797	0.020	0.020	0.797	0.020	0.020	0.797	0.020	0.020
	ϕ^1_{12}	0.10	0.100	0.020	0.021	0.100	0.020	0.021	0.100	0.020	0.021
	ϕ^1_{21}	0.30	0.299	0.023	0.023	0.299	0.023	0.023	0.300	0.023	0.023
	ϕ^1_{22}	0.60	0.598	0.023	0.023	0.598	0.023	0.023	0.598	0.023	0.023
	$\sigma^2_{\xi_1}$	0.25	0.252	0.011	0.012	0.251	0.011	0.012	0.251	0.011	0.012
	$\sigma^2_{\xi_2}$	0.33	0.329	0.015	0.014	0.329	0.015	0.014	0.329	0.015	0.014

[1] Mean estimate

[2] Mean standard error

[3] Standard deviation of estimates

the standard deviations very well for all three methods. The results for $T = 1000$ shown in the bottom half of Table 2.2 indicate the expected overall improvement of the parameter estimates in terms of precision.

The results of the analyses of time series following a MI(4)-AR(1) model are presented in Table 2.3 for series of length $T = 100$ and $T = 1000$. Several observations can now be made for $T = 100$. First, the WLS method performs worst in terms of accuracy, that is, the variances of ε and ξ are not very well recovered compared to the ML and KF method. Second, the standard error estimates of factor loadings λ and variance terms σ^2_ε for the ML and WLS methods seem incorrect when compared to the standard deviations over replications. In contrast, the KF method does seem to produce correct standard errors for these parameters. Finally, the standard deviations of the parameter estimates are comparable for all three methods. The results for a MI(4)-AR(1) model with $T = 1000$ are shown in the bottom half of Table 2.3. The accuracy of the WLS method improves with the increase in T, whereas a difference between mean standard errors and standard deviations remains for factor loadings λ and variance terms σ^2_ε for both the ML and WLS method. Note, however, that this difference is small. Again, the KF method produces standard errors which are comparable to the standard deviation of the estimates.

Table 2.4 shows the results for the MI(4)-AR(2) for $T = 100$ and $T = 1000$.

Table 2.3: Results for MI(4)-AR(1) model

			ML			WLS			KF		
T	Parameter	Value	Mean	SE	SD	Mean	SE	SD	Mean	SE	SD
100	λ_2	0.90	0.902	0.048	0.069	0.902	0.048	0.069	0.900	0.067	0.068
	λ_3	0.90	0.906	0.048	0.068	0.906	0.048	0.069	0.904	0.067	0.068
	λ_4	0.90	0.904	0.048	0.069	0.905	0.048	0.069	0.903	0.067	0.068
	$\sigma^2_{\varepsilon_1}$	0.19	0.188	0.025	0.035	0.168	0.024	0.032	0.188	0.035	0.035
	$\sigma^2_{\varepsilon_2}$	0.19	0.187	0.025	0.036	0.168	0.024	0.034	0.187	0.035	0.036
	$\sigma^2_{\varepsilon_3}$	0.19	0.188	0.025	0.037	0.168	0.024	0.034	0.188	0.035	0.037
	$\sigma^2_{\varepsilon_4}$	0.19	0.189	0.025	0.036	0.169	0.024	0.034	0.190	0.035	0.036
	ϕ_1	0.70	0.677	0.082	0.076	0.698	0.082	0.077	0.677	0.078	0.077
	σ^2_{ξ}	0.51	0.512	0.091	0.098	0.470	0.087	0.094	0.508	0.096	0.097
1000	λ_2	0.90	0.900	0.014	0.021	0.900	0.014	0.021	0.900	0.020	0.020
	λ_3	0.90	0.900	0.014	0.019	0.900	0.015	0.020	0.900	0.020	0.019
	λ_4	0.90	0.900	0.015	0.020	0.900	0.015	0.020	0.900	0.020	0.020
	$\sigma^2_{\varepsilon_1}$	0.19	0.190	0.008	0.011	0.188	0.008	0.011	0.190	0.011	0.011
	$\sigma^2_{\varepsilon_2}$	0.19	0.189	0.008	0.011	0.188	0.008	0.011	0.190	0.011	0.011
	$\sigma^2_{\varepsilon_3}$	0.19	0.190	0.008	0.011	0.188	0.008	0.011	0.190	0.011	0.011
	$\sigma^2_{\varepsilon_4}$	0.19	0.190	0.008	0.011	0.188	0.008	0.011	0.190	0.011	0.011
	ϕ_1	0.70	0.699	0.025	0.023	0.702	0.025	0.023	0.699	0.024	0.023
	σ^2_{ξ}	0.51	0.511	0.029	0.031	0.507	0.029	0.031	0.511	0.030	0.031

The results of the WLS method are not displayed, because the asymptotic covariance matrix W was not positive definite in any replication with $T = 100$ and in a substantial part of the replications with $T = 1000$ (note that W contains 3081 distinct elements). The available results of the ML and KF method are comparable to those obtained for the MI(4)-AR(1) model except that aforementioned differences in standard error and standard deviation are clearer. That is, the standard errors of the ML method appear underestimated.

In Table 2.5, the results are shown for the MI(4)-VAR(1) model obtained with $T = 1000$ and the ML, WLS, and KF method. Results are comparable to those of the MI(4)-AR(1) model in Table 2.3. It seems that the standard error of the estimates of the factor loadings and error variances are incorrect whereas those of the autoregressive parameters and innovation variance are only slightly off. The KF method outperforms both the ML and WLS method in terms of precision.

TABLE 2.4: RESULTS FOR MI(4)-AR(2) MODEL

			ML			KF		
T	Parameter	Value	Mean	SE	SD	Mean	SE	SD
100	λ_2	0.90	0.908	0.040	0.071	0.906	0.068	0.071
	λ_3	0.90	0.905	0.039	0.070	0.903	0.068	0.070
	λ_4	0.90	0.910	0.039	0.075	0.908	0.068	0.075
	$\sigma^2_{\varepsilon_1}$	0.19	0.188	0.020	0.036	0.187	0.035	0.036
	$\sigma^2_{\varepsilon_2}$	0.19	0.189	0.020	0.037	0.189	0.035	0.037
	$\sigma^2_{\varepsilon_3}$	0.19	0.188	0.020	0.036	0.188	0.035	0.036
	$\sigma^2_{\varepsilon_4}$	0.19	0.186	0.020	0.035	0.186	0.035	0.035
	ϕ_1	0.50	0.492	0.112	0.108	0.489	0.108	0.107
	ϕ_2	0.25	0.221	0.112	0.112	0.225	0.108	0.111
	σ^2_ξ	0.52	0.516	0.088	0.098	0.512	0.096	0.097
1000	λ_2	0.90	0.901	0.012	0.021	0.901	0.020	0.021
	λ_3	0.90	0.901	0.012	0.020	0.900	0.020	0.020
	λ_4	0.90	0.900	0.012	0.020	0.900	0.020	0.020
	$\sigma^2_{\varepsilon_1}$	0.19	0.190	0.006	0.011	0.190	0.011	0.011
	$\sigma^2_{\varepsilon_2}$	0.19	0.190	0.007	0.011	0.190	0.011	0.011
	$\sigma^2_{\varepsilon_3}$	0.19	0.190	0.007	0.012	0.190	0.011	0.012
	$\sigma^2_{\varepsilon_4}$	0.19	0.190	0.007	0.011	0.190	0.011	0.011
	ϕ_1	0.50	0.498	0.035	0.035	0.498	0.034	0.035
	ϕ_2	0.25	0.249	0.035	0.035	0.250	0.035	0.035
	σ^2_ξ	0.52	0.523	0.028	0.030	0.523	0.031	0.030

2.4 CATEGORICAL TIME SERIES

We now extend the discussion of methods to estimate autoregressive models to a method for time series that can take on a small number of discrete values, i.e, categorical time series. This method utilizes the WLS procedure to fit models to categorical variables in LISREL that makes use of polychoric correlations for categorical time series. That is, autoregressive models are fitted to a matrix containing polychoric auto- and cross-correlations. The method can be easily implemented analogous to the procedure for continuous time series described in the previous sections. Filtering methods for categorical time series, such as described in Fahrmeir (1992) and Durbin and Koopman (2001), are not employed here, since they are beyond the scope of this chapter.

Table 2.5: Results for MI(4)-VAR(1) model, $T = 1000$

		ML			WLS			KF		
Parameter	Value	Mean	SE	SD	Mean	SE	SD	Mean	SE	SD
λ_{21}	0.90	0.900	0.017	0.023	0.900	0.017	0.023	0.900	0.022	0.023
λ_{42}	0.90	0.900	0.018	0.024	0.900	0.018	0.024	0.900	0.023	0.024
$\sigma^2_{\varepsilon_1}$	0.19	0.189	0.011	0.013	0.187	0.011	0.013	0.189	0.013	0.013
$\sigma^2_{\varepsilon_2}$	0.19	0.190	0.011	0.014	0.188	0.011	0.014	0.190	0.013	0.013
$\sigma^2_{\varepsilon_3}$	0.19	0.190	0.012	0.016	0.188	0.012	0.016	0.190	0.015	0.015
$\sigma^2_{\varepsilon_4}$	0.19	0.190	0.012	0.015	0.188	0.012	0.015	0.190	0.015	0.015
ϕ^1_{11}	0.80	0.794	0.031	0.027	0.794	0.031	0.027	0.794	0.026	0.025
ϕ^1_{12}	0.10	0.100	0.030	0.027	0.101	0.030	0.027	0.101	0.026	0.025
ϕ^1_{21}	0.30	0.302	0.032	0.031	0.302	0.032	0.031	0.303	0.029	0.030
ϕ^1_{22}	0.60	0.595	0.032	0.032	0.595	0.032	0.032	0.594	0.029	0.030
$\sigma^2_{\xi_1}$	0.25	0.252	0.021	0.020	0.251	0.021	0.020	0.252	0.020	0.017
$\sigma^2_{\xi_2}$	0.33	0.329	0.024	0.021	0.328	0.024	0.021	0.329	0.024	0.019

2.4.1 Multivariate autoregressive models for categorical time series

Now, let y_t denote an n-dimensional vector of categorical time series of which each element y_{it}, $i = 1, \ldots, n$, can assume $q + 1$ discrete values ranging from $0, 1, \ldots, q$. To ease the presentation, but without loss of generality, it is assumed that all elements of y_t have an equal number of categories. A continuous latent variable y^*_{it} is assumed to underly each y_{it}, such that

$$y_{it} = k, \quad \text{if} \quad \beta_{ik} < y^*_{it} \leq \beta_{i,k+1}, \tag{2.16}$$

where $\beta_{i0} \equiv -\infty$ and $\beta_{i,q+1} \equiv \infty$ for $i = 1, \ldots, n$ and $k = 0, 1, \ldots, q$. As indicated, the β_{ik}'s are required to be time-invariant thresholds. This requirement is necessary in order for the time series to be stationary.

The specification of VAR and MIVAR models proceeds along the same lines as before, with the notable difference that for VAR models, y^*_t is not observed, and for MIVAR models, both y^*_t and α_t are not observed. This means that additional restrictions are needed in order to provide a metric for both latent processes. Besides time invariant thresholds, the conditions stated in Section 2.2.1 are needed to obtain stationarity.

2.4.2 Polychoric auto- and cross-correlations

Polychoric auto- and cross-correlations are used to estimate multivariate autoregressive models for categorical time series. The estimation of polychoric correlations as implemented in the PRELIS computer program is briefly outlined here for categorical time series. More detailed information on the estimation of the polychoric correlation matrix and its asymptotic covariance matrix can be found in Jöreskog (1994).

Estimation of the thresholds and the polychoric correlation matrix of categorical variables would amount to maximizing the likelihood of the underlying continuous variables with respect to these quantities. Since such an approach rapidly becomes computationally cumbersome with increasing dimension, simpler procedures based on univariate and bivariate marginal distributions are preferred (Song & Lee, 2003). For stationary categorical time series, such a procedure is essentially the same and estimates can be obtained as follows. Consider a stationary standard normal latent process y_t^* which underlies a categorical time series y_t. The univariate marginal distributions can then be given by

$$p(y_{it} = k) = \int_{\beta_{ik}}^{\beta_{i,k+1}} \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2} y_{it}^{*2}\right) dy_{it}^*, \tag{2.17}$$

for $i = 1, \ldots, n$ and $k = 0, 1, \ldots, q$. The bivariate marginal distribution of y_{it} and its lagged version $y_{j,t-p}$, given the first p observations, can be written as

$$p(y_{it} = k, y_{j,t-p} = l) = \int_{\beta_{jl}}^{\beta_{j,l+1}} \int_{\beta_{ik}}^{\beta_{i,k+1}} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{y_{it}^{*2} - 2\rho y_{it}^* y_{j,t-p}^* + y_{j,t-p}^{*2}}{2(1-\rho^2)}\right) dy_{it}^* dy_{j,t-p}^*, \tag{2.18}$$

for $k, l = 0, 1, \ldots, q$ and $i, j = 1, \ldots, n$. In this case, ρ is referred to as the polychoric auto- or cross-correlation of y_{it} and $y_{j,t-p}$ at lag p. With the assumption of stationarity, the parameters of the both marginal distributions do not change under shifts of the time axis, and the latent variables can be integrated out.

The estimation of thresholds and polychoric correlations can now be performed in several ways (Olsson, 1979; Song & Lee, 2003). One manner is to estimate both thresholds and correlations from the joint marginal distributions. A second manner is to estimate the thresholds first from the univariate marginals and then estimate the polychoric correlations from the joint marginals (Jöreskog, 1994). This method has the advantage that it cannot result in different estimates of the same thresholds from one variable computed from different combinations with a second variable (Jöreskog, 1994). This second manner is implemented in PRELIS and is used here.

One can ultimately describe the full marginal likelihood associated with the $n \times (p+1)$-dimensional contingency table and estimate the thresholds and polychoric auto- and cross-correlations under the assumption of stationarity leading to a correlation matrix of Toeplitz form. Again, this likelihood consists of multiple integrals and the estimation procedure rapidly becomes computationally intractable. A method tried by the authors was to first compute the thresholds from the univariate marginals, followed by computing the polychoric auto- and cross-correlation matrix from its full marginal likelihood with given thresholds under the assumption of stationarity. However, since this approach lead to unstable solutions and the PRELIS method can be easily performed, it is preferred for now. It could form an approach in future investigations, because estimates of the polychoric correlations as well as the weight matrix W are then obtained under the assumption of stationarity.

To obtain estimates of the polychoric auto- and cross-correlations with PRELIS, the input data matrix is arranged as follows. In the general case, we have an n-variate categorical time series observed from $t = 1, \ldots, T$, to which we want to fit an autoregressive model. Then, we arrange the data in a data matrix as follows:

$$\begin{bmatrix}
\multicolumn{2}{c}{\text{lag } 0} & & \multicolumn{2}{c}{\text{lag } 1} & & \cdots & & \multicolumn{2}{c}{\text{lag } p} \\
y_{1,p+1} & \cdots & y_{n,p+1} & y_{1,p} & \cdots & y_{n,p} & \cdots & y_{1,1} & \cdots & y_{n,1} \\
y_{1,p+2} & \cdots & y_{n,p+2} & y_{1,p+1} & \cdots & y_{n,p+1} & \cdots & y_{1,2} & \cdots & y_{n,2} \\
\vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\
y_{1,t} & \cdots & y_{n,t} & y_{1,t-1} & \cdots & y_{n,t-1} & \cdots & y_{1,t-p} & \cdots & y_{n,t-p} \\
\vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\
y_{1,T} & \cdots & y_{n,T} & y_{1,T-1} & \cdots & y_{n,T-1} & \cdots & y_{1,T-p} & \cdots & y_{n,T-p}
\end{bmatrix}. \tag{2.19}$$

The observed series of length T is clipped to length $T - p$ so that the time series and lagged versions are of equal length. In order to satisfy the stationarity condition, we need to ensure that the estimated thresholds of different lags of the time series are equal. This is easily resolved in PRELIS, and the polychoric auto- and cross-correlations can then be computed.

A drawback of this procedure is that for shorter time series the estimated

polychoric correlation matrix R_y does not necessarily has the block Toeplitz form:

$$R_y = \begin{bmatrix} R_0 & & & \\ R_1 & R_0 & & \\ \vdots & \vdots & \ddots & \\ R_p & R_{p-1} & \dots & R_0 \end{bmatrix},$$

where the blocks $R_j, j = 0, \dots, p$ contain the polychoric correlations of corresponding columns of Equation 2.19. The blocks that appear repeatedly are computed with different lagged versions of the original time series, and using different lags can lead to different estimates. A simple solution to this problem would be to average the blocks in question. However, some type of pooling procedure has to be applied to the asymptotic covariance matrix W as well, which becomes intractable as the dimension of the series and the order of the autoregressive process increase. In addition, W can lose full rank and the WLS procedure can no longer be applied. Since this problem rapidly disappears as the length of the time series increases, no action is taken in the present study.

2.4.3 SEM REPRESENTATION

Again, the LISREL Submodel 3B is used to represent autoregressive models. The analysis is now performed on the polychoric correlation matrix. This does not influence the representation, which is equal to that in Section 2.2.3. Specifically, the theoretical polychoric correlation matrix $\tilde{R}_y$ is modelled by

$$\tilde{R}_y = \Lambda(I - B)^{-1}\Psi(I - B)^{-1'}\Lambda' + \Theta_\varepsilon. \tag{2.20}$$

An additional constraint is needed since modelling is performed on the latent process y_t^* instead of the observed process. In order to fix the variance of the latent process y_t^* to unity, the constraint can be given by $\text{diag}(\Sigma_\varepsilon) = \text{diag}(I - \Lambda\Lambda')$.

As an example, we restate the latent bivariate VAR(1) model though now with four categorical indicators. For $t = 2, \dots, T$, Let

$$y = \begin{bmatrix} y_{1t} \\ y_{2t} \\ y_{3t} \\ y_{4t} \\ y_{1,t-1} \\ y_{2,t-1} \\ y_{3,t-1} \\ y_{4,t-1} \end{bmatrix}, \Lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} & 0 & 0 \\ \lambda_{21} & \lambda_{22} & 0 & 0 \\ \lambda_{31} & \lambda_{31} & 0 & 0 \\ \lambda_{41} & \lambda_{42} & 0 & 0 \\ 0 & 0 & \lambda_{11} & \lambda_{12} \\ 0 & 0 & \lambda_{21} & \lambda_{22} \\ 0 & 0 & \lambda_{31} & \lambda_{32} \\ 0 & 0 & \lambda_{41} & \lambda_{42} \end{bmatrix}, \eta = \begin{bmatrix} \alpha_{1t} \\ \alpha_{2t} \\ \alpha_{1,t-1} \\ \alpha_{2,t-1} \end{bmatrix}, \text{ and } \Theta_\varepsilon = \text{diag}(I - \Lambda\Lambda').$$

The block-Toeplitz polychoric auto- and cross-correlation matrix $\tilde{R}_y$ is modelled by

$$\tilde{R}_y = \begin{bmatrix} \tilde{R}_0 & \\ \tilde{R}_1 & \tilde{R}_0 \end{bmatrix} = \Lambda(I-B)^{-1}\Psi(I-B)^{-1'}\Lambda' + \Theta_\varepsilon,$$

where $\tilde{R}_0$ and $\tilde{R}_1$ are the polychoric correlation matrices at lag zero and one, respectively, and where

$$B = \begin{bmatrix} 0 & \\ \Phi_1 & 0 \end{bmatrix} \text{ and } \Psi = \begin{bmatrix} \Gamma_0^* & \\ 0 & \Sigma_\xi \end{bmatrix},$$

where Γ_0^* is the variance matrix of the VAR(1) process α_t. Again, identification and rotational constraints are obtained by fixing parameters of Λ or Ψ.

2.4.4 Estimation

Having available an estimate of the polychoric auto- and cross-correlation matrix and its asymptotic covariance matrix W, estimates of the parameters of the autoregressive model can be obtained with the weighted least squares procedure as implemented in LISREL. To this end, the following fit function is minimized

$$F_{\text{WLS}} = \text{vech}(R_y - \tilde{R}_y)'W^{-1}\text{vech}(R_y - \tilde{R}_y). \tag{2.21}$$

For stationary continuous VAR models, the results of the ML and WLS methods are almost exactly the same, as displayed in Tables 2.2 and 2.3. This can be seen as an indication that the WLS procedure in LISREL can work when VAR models are fitted to stationary categorical time series. For that reason, and given the ease of use, the weight matrix W used here is that obtained by PRELIS (for details see Jöreskog, 1994). The derivation of the full information maximum likelihood estimate of the polychoric Toeplitz matrix and its asymptotic covariance matrix would require a separate study.

In summary, the Toeplitz method for categorical time series consists of three steps. First, the thresholds β are estimated from the univariate marginals, which are then used to estimate the polychoric auto- and cross-correlations and their covariances. Finally, the WLS procedure employs the bivariate information of the estimated Toeplitz matrix to obtain parameter estimates of the autoregressive model.

2.5 SIMLATION STUDY II

A second simulation study is performed to investigate the Toeplitz method in the case of categorical time series.

2.5.1 SET UP

The same six autoregressive models are used as in the simulations with continuous time series (see Table 2.1). For all six models, categorical time series are simulated with two and five categories, and lengths $T = 100$ and $T = 1000$. The number of replications equals 1000. The threshold for all time series with two categories is zero leading to $P(y_{it} = 0) = P(y_{it} = 1) = 0.5$. For time series with five categories, the rounded thresholds are -1.65, -0.84, 0.84, and 1.65, leading to the following exact probabilities of the associated ordered categories: 0.05, 0.15, 0.40, 0.15, and 0.05. Parameter values are set so that stationarity holds, and are reported in the tables with results.

The main objective of these simulations is to investigate the performance of the Toeplitz WLS parameter estimation procedure in LISREL for autoregressive models with categorical time series. Parameter estimates are again evaluated on accuracy and precision. Mean parameter estimates are compared to true values and mean estimated standard errors are compared to the standard deviation of parameter estimates over replications.

2.5.2 RESULTS

Only results for models that are illustrative of the methods used are shown and discussed. These models are the VAR(1), MI(4)-AR(1), MI(4)-AR(2), and MI(4)-VAR(1) models. Since a substantial part of the analyses of categorical time series of length $T = 100$ failed due to either the Toeplitz matrix or the W matrix not being positive definite, these analyses are not discussed here. It is not to say that the method cannot be applied for shorter time series, but in order to provide fair comparisons, we have chosen to discuss only situations for which full results are available.

Table 2.6 shows the results for the VAR(1) and MI(4)-VAR(1) models with time series of length $T = 1000$ with two and five categories. For the VAR(1) model, mean parameter estimates closely resemble true values. The results improve as the number of categories increases from two to five in terms of precision, that is, standard errors and standard deviations decrease. Remarkably, estimated

Table 2.6: Results for VAR(1) and MI(4)-VAR(1) models, $T = 1000$

Model	Categories	Parameter	Value	Mean	SE	SD
VAR(1)	2	ϕ_{11}	0.80	0.797	0.057	0.040
		ϕ_{12}	0.10	0.100	0.056	0.044
		ϕ_{21}	0.30	0.299	0.057	0.047
		ϕ_{22}	0.60	0.596	0.056	0.047
	5	ϕ_{11}	0.80	0.794	0.048	0.031
		ϕ_{12}	0.10	0.100	0.042	0.031
		ϕ_{21}	0.30	0.301	0.042	0.032
		ϕ_{22}	0.60	0.595	0.045	0.033
MI(4)-VAR(1)	2	λ_{21}	0.90	0.900	0.022	0.028
		λ_{42}	0.90	0.899	0.026	0.030
		ϕ^1_{11}	0.80	0.797	0.050	0.053
		ϕ^1_{12}	0.10	0.102	0.056	0.049
		ϕ^1_{21}	0.30	0.305	0.055	0.049
		ϕ^1_{22}	0.60	0.594	0.053	0.053
	5	λ_{21}	0.90	0.900	0.016	0.020
		λ_{42}	0.90	0.900	0.018	0.022
		ϕ^1_{11}	0.80	0.797	0.036	0.039
		ϕ^1_{12}	0.10	0.101	0.040	0.037
		ϕ^1_{21}	0.30	0.302	0.039	0.037
		ϕ^1_{22}	0.60	0.598	0.038	0.040

standard errors are larger than standard deviations of the parameter estimates over replications. For the MI(4)-VAR(1) results displayed in the lower part of Table 2.6, the comparability to the continuous time series case increases with the number of categories.

The outcomes of the analyses of categorical time series following MI(4)-AR(1) and MI(4)-AR(2) models are displayed in Table 2.7. Parameters of both models are well recovered by the WLS method in terms of mean values. Again, the larger the number of categories, the more precise the estimates. Yet, for all parameters including the autoregressive parameters, the estimated standard errors are smaller than the standard deviations over replications.

TABLE 2.7: RESULTS FOR MI(4)-AR(1) AND MI(4)-AR(2) MODELS, $T = 1000$

Model	Categories	Parameter	Value	Mean	SE	SD
MI(4)-AR(1)	2	λ_2	0.90	0.901	0.015	0.020
		λ_3	0.90	0.900	0.015	0.020
		λ_4	0.90	0.901	0.015	0.020
		ϕ_1	0.70	0.712	0.027	0.032
	5	λ_2	0.90	0.900	0.010	0.015
		λ_3	0.90	0.900	0.010	0.015
		λ_4	0.90	0.900	0.010	0.015
		ϕ_1	0.70	0.705	0.021	0.027
MI(4)-AR(2)	2	λ_2	0.90	0.901	0.012	0.021
		λ_3	0.90	0.901	0.012	0.021
		λ_4	0.90	0.901	0.012	0.020
		ϕ_1	0.50	0.517	0.053	0.057
		ϕ_2	0.25	0.250	0.055	0.059
	5	λ_2	0.90	0.901	0.008	0.015
		λ_3	0.90	0.900	0.008	0.015
		λ_4	0.90	0.901	0.008	0.015
		ϕ_1	0.50	0.510	0.038	0.044
		ϕ_2	0.25	0.251	0.039	0.045

2.6 DISCUSSION

In this paper, we studied the performance of the Toeplitz method for the fitting of autoregressive models to multivariate stationary time series. Since this method is fast and easy to use for those familiar with SEM, it has the potential to become a popular tool to analyse time series if it performs according to expectations. The results of this study indicate, however, that only in certain situations the Toeplitz method works well. If the purpose of the analysis is to estimate parameters, then the Toeplitz method is useful, purely as a method of moments. In contrast, if the analyses are intended for inferential statements making use of statistics such as standard errors, one should be wary of making these statements based on results obtained with the Toeplitz method.

For normally distributed time series, the results of the present study indicate that the Toeplitz method only provides correct estimates and standard errors for pure VAR models, and not for MI-VAR models. For VAR models, this is to be

expected, because the Toeplitz method is then equal to the multivariate extension of the Yule-Walker method. This method is known to perform badly when the process tends to non-stationarity or is strongly periodical (Tjøstheim & Paulsen, 1983; de Hoon, van der Hagen, Schoonewelle, & van Dam, 1996). For these situations, the ordinary least squares or Burg estimator outperform the Yule-Walker estimator (Lütkepohl, 1991). However, procedures for these estimators cannot be easily constructed within the realm of SEM and implemented in existing SEM software.

To establish the difference with raw data methods, we applied a Kalman filter to all studied models for normally distributed time series as well. It was found that the main difference between the Toeplitz method and the Kalman filtering method existed in the estimation of standard errors. That is, the Kalman filtering method produced correct standard errors for all models studied, while the Toeplitz method did not. The mean and standard deviation of the parameter estimates were, however, very much alike for both methods. For parameter estimation purposes, the Toeplitz method therefore seems usable, for statistical inference, on the other hand, it is not. Since we took into account the fact that the necessary auto-covariances are only efficient for certain models, we conclude that it is likely that for other multivariate models, the Toeplitz method will not work well for inference. However, it might still have practical utility. For the models studied here, one can for instance multiply the standard errors by a factor two in order to be on the safe side for the signifance of parameter estimates.

For categorical time series, the results are comparable or worse. That is, the parameters are recovered well in terms of accuracy, but estimated standard errors are incorrect for all models used in this study. Part of these results might be traced to problems associated with the estimation of the matrix of polychoric auto- and cross-correlations and its asymptotic covariance matrix under the condition of stationarity. The task of improving the estimation procedure has more cons than pros, however. The complexity of this task is high and based on the results obtained with continuous time series, the merit is questionable. The effort might be better spend on developing or improving filtering techniques for categorical time series. In future investigations, the results of the present study can be linked to those obtained with a filter based method for categorical time series. Up till now, such methods have seen little application within the domain of behavioral research.

An important shortcoming of the Toeplitz method is that it can only be used for stationary time series. Since it is more likely that nonstationarity is the rule

rather than the exception in behavioral research practice, the Toeplitz method is no match for filtering techniques. Despite the results of our simulation studies, the Toeplitz method is a useful point of departure in the analysis, e.g., by providing sensible starting values for computer intensive filtering techniques.

3

State space analysis of univariate categorical time series

3.1 Introduction

In time series analysis, the interest lies in the modelling of sequential observations obtained from some aspect of a dynamic system in order to explain or predict its course. The general approach considered in this paper to modelling time series, without making reference to a specific class of models, is the state space approach. In this approach, the observed time series is associated with a time series of unobserved states through an observation equation. The dynamics of the system are modelled at the level of the unobserved time series. Specifically, this dynamic part is represented by a stochastic difference equation, which is referred to as the transition equation. The combination of the observation and transition equation results in the canonical state space model. In this chapter, we limit our attention to the analysis of univariate time series that can only take on a restricted number of discrete values, that is, we consider univariate categorical time series models.

In general, state space modelling proceeds by specifying an appropriate observation and transition equation, and estimating the unobserved states, and the unknown parameters of the state space model. Although selecting an appropriate state space model can be a complicated process, the primary interest in this chapter lies in its estimation. The problem of estimating unobserved states is dependent on the phase of the data collection process. If estimates of future states are required, it is referred to as the prediction problem. If estimates are required as the data are recorded, the problem is one of filtering (online). Thirdly, if estimates are required only after the whole time series has been recorded, the problem is one of smoothing (offline). In this paper, filtering and smoothing techniques are considered given the complete time series (offline). One commonly used technique for normally distributed time series is the Kalman filter (Kalman, 1960), which proceeds forwards in time by sequently updating the available state estimates as observations are being gathered. The Kalman filter

represents a least-squares solution to the problem of minimizing prediction error. However, it makes efficient use of the normal distribution, which allows the linear transformations that are necessary for the solution (Jullier & Uhlmann, 1997). A commonly used smoothing technique is the classical fixed interval smoother that moves backwards in time and adjusts the estimates of the Kalman filter by utilising information unknown at the time the filter was applied (see Anderson & Moore, 1979).

A great deal of the research in the field of time series modelling is based on the normal distribution (Harvey, 1989; Hamilton, 1994; Durbin & Koopman, 2001). For time series following a normal distribution, observations can be related directly to the states, and when combined with a linear transition, filtering and smoothing procedures can be readily applied. However, ever since the advent of filtering and smoothing techniques, many applications have been challenged by difficulties concerning non-normally distributed observations, non-linear transitions, or both. Several methods have been developed over the years to account for these problems, which can essentially be divided into two categories. The first category consists of exact solutions, which are available only for some specific non-linear cases (see Elliott, Aggoun, & Moore, 1995; Vidoni, 1999). The second and largest category is formed by solutions in which some kind of approximation procedure is used. In turn, this category can roughly be subdivided into two types. The first type is formed by solutions that replace non-normal and non-linear elements of a state space model by normal and linear approximations. In the extended Kalman filter (Jazwinski, 1970), for example, Taylor expansions of non-linear transitions around provisional state estimates are simply inserted in the usual Kalman filtering recursions. Although many approximate methods of this type work well for minor departures of normality and linearity, it is known that they break down in more intricate situations (Schnekenburger, 1988). The second type, which is computationally more intensive, employs simulation techniques to approximate non-normal and non-linear elements numerically. These techniques are generally referred to as sequential Monte Carlo methods (for an overview, see Doucet, de Freitas, & Gordon, 2001). With the increases in computing power, this type of solutions has become a powerful tool for estimating non-normal and non-linear state space models.

The special case of non-normality considered in this paper arises when the time series can be characterised as sequential categorical observations. The discrete nature of these observations precludes the direct use of the normal distribution, and so modelling categorical time series has to be performed by different

means. It is noted that categorical time series are a special case of discrete variate time series (see McKenzie, 2003). Whereas realizations of the latter exist of (positive) integer values, categorical time series only assume a limited number of integer values. The class of time series models considered here is that of state space models for exponential family time series (see, e.g., Durbin & Koopman, 1997). In this class, it is assumed that the time series observations follow a distribution belonging to the exponential family. Since the binomial and multinomial distributions are generally used for categorical observations and are members of the exponential family, this approach allows for a general frame of reference.

As for the available filtering and smoothing techniques that can be applied in this case, a number of them deserve special attention here. Kitagawa (1987) presented a general approach to non-normal state space modelling in which the probability densities or distributions are numerically approximated directly by first order splines. However, this approach is computationally intractable for higher dimensional models. Fahrmeir (1992) presented a generalized extended Kalman filter and smoother in which posterior modes of a penalized likelihood are numerically approximated. An iterative version of this filter and smoother is discussed in Fahrmeir and Wagenpfeil (1997), and is used here. Durbin and Koopman (1997) present a simulation technique in which essentially the same recursions as those of Fahrmeir and Wagenpfeil (1997) are used. It is the purpose of this paper to discuss and investigate a unified approach to estimate both states and unknown parameters of models for univariate categorical time series. A further purpose is to assess the performance of the methods by means of simulation.

In practice, state space models include not only unobserved states, but also certain other parameters, depending on the time series model used. Such parameters are often called hyperparameters, and are generally unknown. Estimating these parameters is a difficult problem, since state estimates and parameter estimates are confounded. Kitagawa (1998) presented a self-organizing state space model, in which states and parameters are estimated simultaneously. However, most other procedures are iterative in that they first estimate states with given parameters, and then estimate parameters with given states until convergence. We discuss such a technique for the estimation of the parameters of categorical time series models which is based on the penalized likelihood criterion of Fahrmeir and Wagenpfeil (1997).

This chapter is set out as follows. First, the state space modelling approach for univariate categorical time series is discussed. In the next section, filtering and smoothing techniques are presented. This is followed by a discussion of a

method to estimate parameters. Then, the results of a simulation study are presented, which was conducted to assess the performance of the presented methods. Finally, the methods are illustrated by an application to a categorical time series consisting of sleep state measurements. The chapter ends with a discussion of the results.

3.2 Categorical time series models

We discuss our approach to modelling categorical time series within the framework of state space models. To ease the presentation, we assume that the distributions are members of the exponential family, and that the latent process is linear and normally distributed (see also Durbin & Koopman, Ch. 10, 2001; Fahrmeir & Tutz, Ch. 8, 2001; Klein, 2003). This approach to modelling can also be fitted into the framework of dynamic generalized linear models (West, Harrison, & Migon, 1985; Fahrmeir & Tutz, 2001). However, we adhere to the state space framework, because the filtering and smoothing techniques that we use are explained more naturally within this framework. For dynamic regression models for categorical time series, the reader is referred to Kedem and Fokianos (2002).

As a prelude to the discussion of categorical time series, we shortly discuss the state space approach to modelling exponential family time series. In general, an n-dimensional multivariate time series y_t is assumed to follow an exponential family distribution. Dropping the dispersion parameter for simplicity, this is

$$p(y_t|\theta_t, y_{t-1}^*) = \exp(y_t'\theta_t - b_t(\theta_t) + c_t(y_t)), \tag{3.1}$$

where θ_t is referred to as the natural parameter, y_{t-1}^* denotes the history of y_t given by $(y_{t-1}', \ldots, y_1')'$, $b_t(\theta_t)$ is a twice differentiable function, and $c_t(y_t)$ is a function of y_t only. Equation 3.1 is referred to as the observation equation. The natural parameter θ_t is related to the linear predictor η_t as follows

$$\begin{aligned}\theta_t &= v(\eta_t) = \mu_t^{-1}(h(\eta_t)),\\ \eta_t &= u(\theta_t) = g(\mu_t(\theta_t)),\end{aligned}$$

where $g(.)$ is the link function, and $h(.) = g(.)^{-1}$ is the response function. If the function $g(.)$ is the natural link function, the mean value mapping $\mu(.)$ equals $g(.)^{-1}$, and consequently, the natural parameter and the linear predictor coincide, i.e., $\theta_t = \eta_t$. We limit our attention in this paper to models, which allow for the

use of the natural link function, and hereafter, we only refer to the linear predictor η_t. In Equation 3.1, it is assumed that the present observation is independent of the history of the process η^*_{t-1}, that is

$$p(y_t|\eta^*_t, y^*_{t-1}) = p(y_t|\eta_t, y^*_{t-1}). \tag{3.2}$$

In addition, it is assumed that the process η_t is first-order Markovian

$$p(\eta_t|\eta^*_{t-1}, y^*_{t-1}) = p(\eta_t|\eta_{t-1}). \tag{3.3}$$

Two standard results of exponential family distributions are that the mean and variance functions are the first and second derivatives of the function $b_t(\eta_t)$ with respect to η_t, that is

$$\mathrm{E}(y_t|\eta_t, y^*_{t-1}) = \frac{\partial b_t(\eta_t)}{\partial \eta_t} = \mu_t, \tag{3.4}$$

$$\mathrm{Var}(y_t|\eta_t, y^*_{t-1}) = \frac{\partial^2 b_t(\eta_t)}{\partial \eta_t \partial \eta_t'} = \Sigma_t. \tag{3.5}$$

The structural relation is formed by the specific construction of the linear predictor η_t. To this end, an $n \times m$ design matrix Z_t with known elements and an m-dimensional series of unobserved states α_t are related to the mean by

$$\mu_t = h(\eta_t) = h(Z_t\alpha_t). \tag{3.6}$$

In general, Z_t can consist of fixed values, covariates and past values of y_t. The dynamic part of the model is captured in the sequential dependence of α_t. A normal linear transition equation can be expressed for $t = 1, 2, \ldots, T$, as

$$\alpha_t = F_t\alpha_{t-1} + R_t\xi_t, \qquad \xi_t \sim \mathrm{N}(0, Q_t), \tag{3.7}$$

where F_t is an $m \times m$ transition matrix, R_t is an $m \times p$ selection matrix with fixed elements, and ξ_t is a p-dimensional normally distributed white noise sequence with associated covariance matrix Q_t. An advantage of this representation is that the state vector α_t can consist of time-varying and time-constant elements, and by using the selection matrix appropriately Q_t remains positive definite. The time-varying elements of the initial state α_0 are assumed to be normally distributed with mean the corresponding elements of a_0 and $p \times p$ covariance matrix Q_0.

A wide variety of models is obtained by the specific choice of the distribution, the link function, the structural relation, and the transition equation. For normally distributed time series, for example, autoregressive moving average

(ARMA) models and structural models with trends and seasonal effects can be specified in this manner. We now turn to the discussion of the models that we use for univariate categorical time series in which we make a distinction between categorical time series with two categories (dichotomous) and more than two categories (polytomous).

3.2.1 Dichotomous time series

A special case of categorical time series arises when the time series can take on only two values at each time point, generally scored as either 0 or 1. In this case, the binomial distribution can be assumed, which is a member of the exponential family. Since a single realisation of the time series is considered in the present situation, the binomial coefficient is discarded. Such a time series can be considered to be a dependent sequence of Bernoulli trials. The probability of y_t is then given by

$$p(y_t|\pi_t, y_{t-1}^*) = \pi_t^{y_t}(1-\pi_t)^{(1-y_t)}, \tag{3.8}$$

where π_t is the probability that y_t takes on the value 1 at time t. The natural link and response function, and the variance function associated with the binomial distribution are respectively given by

$$\eta_t = \log\left(\frac{\pi_t}{1-\pi_t}\right), \tag{3.9}$$

$$\pi_t = \frac{\exp(\eta_t)}{1+\exp(\eta_t)}, \tag{3.10}$$

$$\sigma_t^2 = \pi_t(1-\pi_t). \tag{3.11}$$

A complete model is obtained by the specific construction of the linear predictor and the transition equation.

As an example of how model specification proceeds, we discuss this in detail for a dichotomous time series which is governed by an unobserved stationary normally distributed first-order autoregressive (AR(1)) process θ_t. Then, the linear predictor is inserted in the logistic response function given by Equation 3.10, and is constructed as follows

$$\eta_t = \theta_t + \beta,$$

where β is a time-invariant threshold. The latent AR(1) process θ_t can be expressed as

$$\theta_t = \phi_1\theta_{t-1} + \xi_t, \qquad \xi_t \sim \mathrm{N}(0, \sigma_\xi^2), \tag{3.12}$$

where ϕ_1 is the autoregressive parameter, and ξ_t is a white noise sequence. The process θ_t is said to be stationary if it holds that $|\phi_1| < 1$. Since θ_t has no scale, we fix it to have zero mean and unit variance by setting $\sigma_\xi^2 = 1 - \phi_1^2$ (see Hamilton, 1994, p. 53). Now, the time indices of the state space model matrices in Equations 5.9 and 5.10 are discarded, and to represent the above model in state space form their elements are filled in as follows

$$Z = \begin{bmatrix} 1 & 1 \end{bmatrix}, \quad \alpha_t = \begin{bmatrix} \theta_t \\ \beta \end{bmatrix}, \quad F = \begin{bmatrix} \phi_1 & 0 \\ 0 & 1 \end{bmatrix}, \quad R = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad Q = \begin{bmatrix} \sigma_\xi^2 \end{bmatrix}.$$

Note that the innovations ξ_t are simply inserted in the transition equation. Other dynamic models can be formed as well, and are amenable to state space form by analogous reasoning.

3.2.2 Polytomous time series

The multinomial distribution, which is also a member of the exponential family, is adopted for polytomous time series. The number of categories is denoted by q. It is common practice to dummy code each polytomous response by $k = q - 1$ dummy variables y_{jt}, $j = 1, \ldots, k$ (see Fahrmeir & Tutz, 2001). The dummies are coded 1 if category j is observed, and 0 otherwise. In this manner, the polytomous time series is represented by a multivariate dichotomous dummy time series $y_t = (y_{1t}, \ldots, y_{kt})'$. Again, the multinomial coefficient is discarded, because we only consider a single realisation of the time series. Then, the multinomial distribution function is given by

$$p(y_t|\pi_t, y_{t-1}^*) = \pi_{1t}^{y_{1t}} \ldots \pi_{kt}^{y_{kt}} (1 - \pi_{1t} - \ldots - \pi_{kt})^{1 - y_{1t} - \ldots - y_{kt}}. \tag{3.13}$$

A distinctive feature of a polytomous time series is whether it consists of nominal or ordinal categories. In the case of nominal categories, the category specific probabilities can be modelled without reference of one category to another, i.e., the only feature is that the categories are distinct. For ordinal categories, however, it is desirable that the ordinal information be represented in the modelling of the category specific probabilities. This can be achieved in various manners, and different ordering representations lead to different interpretations of the model and its parameters. For some representations, however, the natural link function can no longer be used and the advantages of the exponential family are lost. For a more thorough description of different link functions and ordering representations for models with categorical variables, the reader is referred to Agresti (2002) and Fahrmeir and Tutz (2001).

The natural link and response function, and the variance function associated with the multinomial distribution are given by

$$\eta_{jt} = \log\left(\frac{\pi_{jt}}{1 - \sum_{i=1}^{k} \pi_{it}}\right), \tag{3.14}$$

$$\pi_{jt} = \frac{\exp(\eta_{jt})}{1 + \sum_{v=1}^{k} \exp(\eta_{vt})}, \quad j = 1, \ldots, k, \tag{3.15}$$

$$\Sigma_t = \text{diag}(\pi_t) - \pi_t \pi_t'. \tag{3.16}$$

Our interest lies in ordinal categories, and we discuss models for this type only. As an example, we completely specify the model for a polytomous time series with three categories following a latent stationary Gaussian first order autoregressive process. Writing out the response function for this situation results in

$$\pi_{0t} = \frac{1}{1 + \exp(\eta_{1t}) + \exp(\eta_{2t})},$$
$$\pi_{1t} = \frac{\exp(\eta_{1t})}{1 + \exp(\eta_{1t}) + \exp(\eta_{2t})},$$
$$\pi_{2t} = \frac{\exp(\eta_{2t})}{1 + \exp(\eta_{1t}) + \exp(\eta_{2t})}.$$

In order to retain the ordinal information, the linear predictor is constructed as follows

$$\begin{bmatrix} \eta_{1t} \\ \eta_{2t} \end{bmatrix} = \begin{bmatrix} \theta_t + \beta_1 \\ 2\theta_t + \beta_1 + \beta_2 \end{bmatrix},$$

where θ_t is an AR(1) process as given in Equation 3.12 and β_1 and β_2 are threshold parameters. The model can be expressed in terms of conditional probabilities, which emphasizes the resemblance with the model for dichotomous time series. For the two conditional dichotomies that y_t is either 0 or 1, and either 1 or 2, the conditional probabilities are given by

$$\pi_{1t|0,1} = \frac{\pi_{1t}}{\pi_{0t} + \pi_{1t}} = \frac{\exp(\theta_t + \beta_1)}{1 + \exp(\theta_t + \beta_1)},$$
$$\pi_{2t|1,2} = \frac{\pi_{2t}}{\pi_{1t} + \pi_{2t}} = \frac{\exp(\theta_t + \beta_2)}{1 + \exp(\theta_t + \beta_2)}.$$

So, the model provides a natural extension of the model for dichotomous time series. This extension resembles the extension of the Rasch model to the partial credit model used in psychometrics (see Masters, 1982). The state space

representation of this model is given by

$$Z = \begin{bmatrix} 1 & 1 & 0 \\ 2 & 1 & 1 \end{bmatrix}, \quad \alpha_t = \begin{bmatrix} \theta_t \\ \beta_1 \\ \beta_2 \end{bmatrix}, \quad F = \begin{bmatrix} \phi_1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad R = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad Q = \begin{bmatrix} \sigma_\xi^2 \end{bmatrix}.$$

Again, it is acknowledged that the ordinal information in polytomous time series can be represented in different manners. However, we find this representation convenient, because of its natural extension of dichotomous models. Other types of dynamic models can be obtained by appropriate adjustments.

3.3 Estimation

We discuss the estimation of categorical time series models in two parts. First, the filtering and smoothing procedures for the estimation of the latent process α_t are discussed in the situation of known parameters. Second, we discuss the estimation of these parameters. Both procedures are based on the log-posterior distribution of the underlying states α_T^* given an observed stretch $y_1, y_2, \ldots, y_T$. This log-posterior is a penalized log-likelihood criterion (Fahrmeir & Tutz, 2001, p. 351), and can be obtained from the state space model and the assumptions described in the previous section. This results in

$$\begin{aligned} L(\alpha_T^*) = &\sum_{t=1}^{T} l_t(\alpha_t) - \tfrac{1}{2}(\alpha_0 - a_0)' R_0 Q_0^{-1} R_0'(\alpha_0 - a_0) \\ &- \tfrac{1}{2} \sum_{t=1}^{T} (\alpha_t - F_t \alpha_{t-1})' R_t Q_t^{-1} R_t'(\alpha_t - F_t \alpha_{t-1}), \end{aligned} \tag{3.17}$$

where the contributions of the categorical time series observations are given by

$$l_t(\alpha_t) = \sum_{j=1}^{k} y_{jt} \log(\pi_{jt}) + (1 - \sum_{j=1}^{k} y_{jt}) \log(1 - \sum_{j=1}^{k} \pi_{jt}). \tag{3.18}$$

3.3.1 Estimating latent states

In order to obtain estimates of the latent process α_t, we make use of the iteratively weighted Kalman filter and smoother (IWKFS) developed by Fahrmeir and Wagenpfeil (1997). The resulting estimates are numerical approximations to posterior modes and curvatures of the log-posterior in Equation 3.17. The filter consists of a prediction and correction step of which the modes and curvatures

are denoted by $a_{t|t-1}$, $V_{t|t-1}$, $a_{t|t}$, and $V_{t|t}$, respectively. The smoother steps are denoted by $a_{t|T}$ and $V_{t|T}$. In discussing the filter and smoother, the parameters in a_0, Q_0, F_t, and Q_t are considered as fixed and known. The estimation of these parameters is discussed in the next paragraph.

Since the considered distributions are members of the exponential family and we consider natural link and response functions only, the filtering recursions can be expressed in a more simplified form than described in Fahrmeir and Wagenpfeil (1997). Essentially, both the Jacobian matrix of the response function and the conditional covariance function Σ_t are needed in the recursions. However, these coincide when the natural link and response functions are used, which leads to the simplifications. Durbin and Koopman (2000) describe a similar procedure derived on different grounds. However, their recursions can be rewritten into the recursions of Fahrmeir and Wagenpfeil (1997) in the case of natural link functions.

At each iteration i, the IWKFS needs to be invoked with evaluation values for the latent process, given by $\tilde{a}^i = (\tilde{a}_1^{i\prime}, \tilde{a}_2^{i\prime}, \ldots, \tilde{a}_T^{i\prime})'$. We continue by defining the filtering recursions for $t = 1, \ldots, T$, as follows

1. Prediction:

$$\begin{aligned} a_{t|t-1} &= F_t a_{t-1|t-1}, \quad a_{0|0} = a_0, \\ V_{t|t-1} &= F_t V_{t-1|t-1} F_t' + R_t Q_t R_t', \quad V_{0|0} = R_0 Q_0 R_0'. \end{aligned} \tag{3.19}$$

2. Correction:

$$\begin{aligned} V_{t|t} &= (V_{t|t-1}^{-1} + B_t)^{-1}, \\ a_{t|t} &= a_{t|t-1} + V_{t|t} b_t, \end{aligned} \tag{3.20}$$

   where B_t and b_t are given by

$$\begin{aligned} B_t &= Z_t' \Sigma_t Z_t, \\ b_t &= Z_t'(y_t - h(Z_t \tilde{a}_t^i)) - B_t(a_{t|t-1} - \tilde{a}_t^i), \end{aligned}$$

   and Σ_t is evaluated at $\tilde{a}_t^i$.

In the first iteration, the state evaluation vector consists of the filter predictions, i.e., $\tilde{a}_t^1 = a_{t|t-1}$, and the last term in the expression of b_t disappears. The filter then simplifies to the generalized extended Kalman filter developed by Fahrmeir (1992).

The fixed interval smoother for $t = T, \ldots, 2$ is given by the recursions

$$\begin{aligned} a_{t-1|T} &= a_{t-1|t-1} + G_t(a_{t|T} - a_{t|t-1}) \\ V_{t-1|T} &= V_{t-1|t-1} + G_t(V_{t|T} - V_{t|t-1})G_t', \end{aligned} \tag{3.21}$$

where

$$G_t = V_{t-1|t-1}F_t'V_{t|t-1}^{-1}. \tag{3.22}$$

The smoother is started with filter estimates $a_{T|T}$ and $V_{T|T}$. After each iteration, the state evaluation vector is updated with the smoother estimates, that is, $\tilde{a}^{i+1} = (a_{1|T}', a_{2|T}', \ldots, a_{T|T}')'$. The procedure is repeated upon convergence of the state estimates. It is noted that in our implementation of the above recursions, we have used the Moore-Penrose generalized inverse in the case that any of the matrices to be inverted is singular.

3.3.2 Estimating parameters

Having available estimates of the latent process α_T^*, we can now maximize the penalized likelihood in Equation 3.17 with respect to the parameters. The course chosen here is to perform this maximization numerically. To this end, we employ the NPSOL software package which consists of a number of FORTRAN routines with which the penalized likelihood is maximized by means of a quasi-Newton optimization procedure (Gill, Murray, Saunders, & Wright, 1986). Other optimization routines can be used as well.

Recall the example of an AR(1) process underlying a polytomous time series with three ordered categories. The parts of the penalized likelihood can be expressed as follows

$$\begin{aligned} l_t(\alpha_t) &= y_{1t} \ln \left(\frac{\exp(\theta_t + \beta_1)}{1 + \exp(\theta_t + \beta_1) + \exp(2\theta_t + \beta_1 + \beta_2)} \right) \\ &+ y_{2t} \ln \left(\frac{\exp(2\theta_t + \beta_1 + \beta_2)}{1 + \exp(\theta_t + \beta_1) + \exp(2\theta_t + \beta_1 + \beta_2)} \right) \\ &+ (1 - y_{1t} - y_{2t}) \ln \left(\frac{1}{1 + \exp(\theta_t + \beta_1) + \exp(2\theta_t + \beta_1 + \beta_2)} \right) \\ &= y_{1t}(\theta_t + \beta_1) + y_{2t}(2\theta_t + \beta_1 + \beta_2) \\ &- \ln(1 + \exp(\theta_t + \beta_1) + \exp(2\theta_t + \beta_1 + \beta_2)). \end{aligned}$$

The contribution of the latent process reduces to

$$\tfrac{1}{2} \sum_{t=1}^{T} \frac{(\theta_t - \phi_1\theta_{t-1})^2}{\sigma_\xi^2},$$

where σ_ξ^2 is fixed to $1 - \phi_1^2$, so that the total variance of the process is equal to one. The initial state and variance are given by

$$a_0 = \begin{bmatrix} \theta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \qquad Q_0 = \left[\sigma_{\theta_0}^2\right],$$

where θ_0 and $\sigma_{\theta_0}^2$ are constrained to be zero and one, respectively, for identification purposes. With this constraint, the initial state contribution to the likelihood can be discarded. So, β_1, β_2, and ϕ_1 are the only parameters to be estimated. Estimates of the threshold parameters β_1 and β_2 can be obtained by application of the IWKFS. So, our purpose in this example is to estimate ϕ_1, which can be performed by numerically maximizing the penalized likelihood. Maximization is performed by NPSOL which returns an estimate and a (numerically approximated) standard error. After a maximum is found, the IWKFS is applied once more to obtain final estimates of the latent process θ_t and the threshold parameters β_1 and β_2.

3.4 Simulation study

3.4.1 Set up

A series of simulations was carried out to investigate the performance of the IWKFS for the estimation of the models described in the previous section. In order to provide insight into the methods, the following variables are used: the number of categories and the length of the time series. The number of categories is chosen to be two, three, and five. The lengths of the time series equal $T = 100$ and $T = 1000$. All simulated categorical time series follow a latent Gaussian AR(1) with $\phi_1 = 0.7$ and consequently $\sigma_\xi^2 = 0.51$, and the number of replications is set at 1000.

The distribution of the thresholds is set so that the probabilities at $\theta_t = 0$ are symmetrically distributed over the categories. For dichotomous time series, threshold β_1 is set at 0. For polytomous time series, thresholds β_1 and β_2 are set at 0.75 and -0.75 for three categories, and $\beta_1, \beta_2, \beta_3$, and β_4 are set at 1.50, 0.50, -0.50, and -1.50 for five categories, respectively. It is noted that the threshold parameters need to be ordered for simulating the time series in the polytomous case. Otherwise some categories never come to be the most probable category regardless of the value of the latent process. Since the latent process is the only

source of variaton, not all categories can be reached as a consequence. This is not to say that the threshold parameters need to be ordered in the estimation procedure.

The performance of the parameter estimation procedure with the IWKFS is evaluated by comparing mean parameter estimates over replications with true parameter values, and mean estimated standard errors with standard deviations of parameter estimates over replications.

3.4.2 Results

The results of the simulation study are shown in Tables 3.1 and 3.2. For the autoregressive parameter ϕ_1, Table 3.1 displays mean estimates over replications, mean estimated SEs as produced by the NPSOL optimization routine, and SDs of estimates taken over replications. The estimates of ϕ_1 are reasonably close to their true values. In addition, the fact that in all studied cases mean standard errors are larger than the standard deviations of the estimates can be seen as an indication that the AR parameter can be estimated consistently.

Table 3.1: Results for autoregressive parameter

T	Categories	Parameter	Value	Mean[1]	SE[2]	SD[3]
100	2	ϕ_1	0.70	0.692	0.129	0.091
1000	2	ϕ_1	0.70	0.744	0.046	0.043
100	3	ϕ_1	0.70	0.692	0.112	0.081
1000	3	ϕ_1	0.70	0.724	0.032	0.024
100	5	ϕ_1	0.70	0.626	0.111	0.072
1000	5	ϕ_1	0.70	0.670	0.031	0.024

[1] Mean estimate

[2] Mean standard error

[3] Standard deviation of estimates

Table 3.2 shows mean threshold estimates over replications, mean estimated SEs as produced by IWKFS, and SDs of threshold estimates over replications. The estimates of the threshold parameters β do not closely resemble true values, and are not very consistent when comparing mean standard errors to standard deviations over replications. This result is most likely due to the IWKFS and the fact that all parameters are estimated together.

TABLE 3.2: RESULTS FOR THRESHOLD PARAMETERS

T	Categories	Parameter	Value	Mean	SE	SD
100	2	β_1	0.00	0.009	0.325	0.445
1000	2	β_1	0.00	-0.007	0.108	0.147
100	3	β_1	0.75	1.223	0.373	0.510
		β_2	-0.75	-1.265	0.376	0.502
1000	3	β_1	0.75	1.186	0.118	0.151
		β_2	-0.75	-1.188	0.118	0.148
100	5	β_1	1.50	1.716	0.454	0.961
		β_2	0.50	0.810	0.357	0.500
		β_3	-0.50	-0.793	0.356	0.479
		β_4	-1.50	-1.658	0.441	0.974
1000	5	β_1	1.50	1.745	0.143	0.737
		β_2	0.50	0.784	0.114	0.199
		β_3	-0.50	-0.792	0.114	0.197
		β_4	-1.50	-1.729	0.142	0.750

Two options can be considered in adjusting the model. The first option is to use a different link function, for example the logistic function with a different parameterization or the probit function (see e.g., Song, 2000). However, order restrictions on the threshold parameters are needed in most cases (Fahrmeir & Tutz, 2001, p. 348). In addition, the benefits of using the canonical link function are lost when using a different link function. A second option is to assume a logistic distribution for the sequence ξ_t (see Arnold & Robertson, 1989; and Durbin & Koopman, 2001, Ch. 10), but one then loses the benefits of the normal distribution. Either way, the penalized likelihood criterion and the IWKFS recursions need to be rewritten accordingly, and the parameters have a different interpretation.

3.5 REAL DATA EXAMPLE

We now discuss an application of the method to real data consisting of sleep state measurements of an infant recorded over one night as discussed in Kedem and Fokianos (2002, §3.5.3). The data are EEG coded into one of four ordered categories: awake (0), quiet sleep (1), indeterminate sleep (2), and active sleep (3). The purpose of this analysis is illustrate how the discussed methods can be

Table 3.3: Results for sleep data

Model	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	D	AIC	BIC	%
AR(1)	0.94	-	1.34	-2.05	-0.92	953.73	961.73	981.46	87.36
	(0.01)	-	(0.20)	(0.22)	(0.23)				
AR(2)	0.66	0.31	0.83	-3.13	-2.44	796.74	806.74	831.40	93.15
	(0.03)	(0.03)	(0.56)	(0.57)	(0.58)				

used to find a suitable model that can describe the sleep state process with fair accuracy. To this end, we have fitted an AR(1) and AR(2) model. The results of the analysis are displayed in Table 3.3. To compare the fit of the two models, we investigated several goodness of fit measures (see Fahrmeir & Tutz, 2001, §3.4.3; Kedem & Fokianos, 2002, §3.4.3). One of these is the scaled deviance or likelihood ratio statistic, and can be expressed as

$$D = -2 \sum_{t=1}^{T} l_t(\alpha_t),$$

where $l_t(\alpha_t)$ is given by Equation 3.18. The asymptotic distibution of D approaches a χ^2 distribution with $kT - p$ degrees of freedom, where p is the number of parameters. However, the approximation may fail and large values cannot be seen as an indication of lack of fit (Fahrmeir & Tutz, 2001, p. 51). We also investigated Akaike's information criterion (AIC) and the Bayesian information criterion (BIC), which can be given by

$$\begin{aligned} \text{AIC} &= D + 2p, \\ \text{BIC} &= D + p \log T. \end{aligned}$$

Since the variance of the latent process is fixed to one for scaling purposes, the percentage of variance explained by the model can also be used for inspecting model fit. This percentage is calculated easily from the model parameters in this situation, and is given in the last column in the Table 3.3. All four fit measures indicate that in this analysis the AR(2) model provides a better explanation of the data than the AR(1) model. Figure 3.1 depicts the sleep state measurements and the predicted values of the AR(1) and AR(2) models. It can be seen that the AR(2) model fits closer to the data. For a description of analyses in which auxiliary information in the form of predictors such as heart rate and temperature are used, see the discussion in Kedem and Fokianos (2002, §3.5.3).

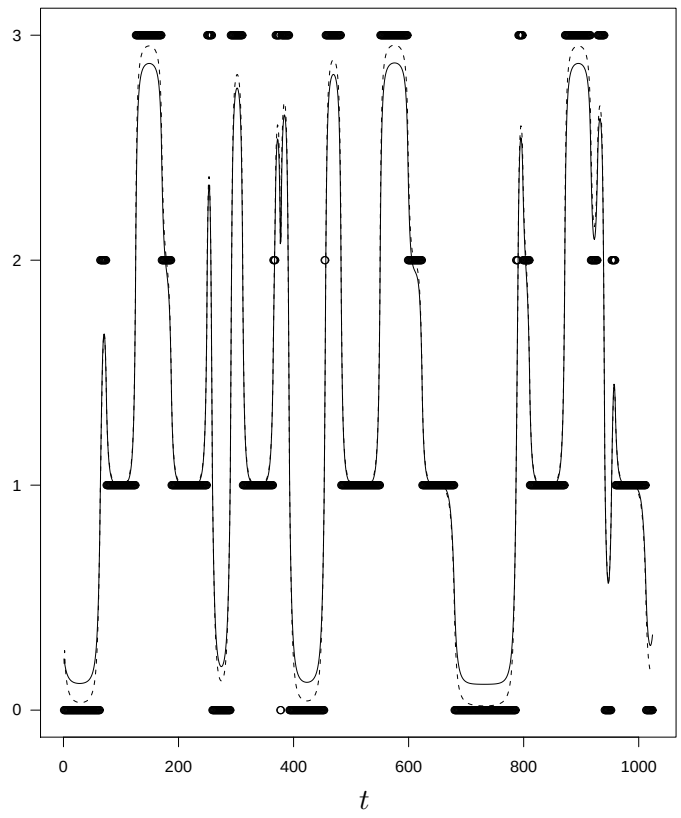


FIGURE 3.1: SMOOTHED EXPECTED VALUES FOR AR(1)(-) AND AR(2)(- -) MODELS WITH DOTTED OBSERVATIONS

3.6 DISCUSSION

The IWKFS developed by Fahrmeir and Wagenpfeil (1997) was implemented to fit models to categorical time series as well as a method to numerically optimize a penalized likelihood to obtain estimates of parameters of these models. The results of the simulation study indicate that the autoregressive parameter of an AR(1) model can be estimated with reasonable consistency. However, the estimates of threshold parameters obtained with the IWKFS proceeded are biased and inconsistent. An application to sleep state measurements illustrate the use of the modelling and estimation techniques.

The approach to modelling categorical time series presented in this paper can be extended naturally to multivariate categorical time series, which remains an interesting topic for further research and is largely untouched. The performance of the estimation of thresholds is expected to improve, particularly when the latent process is of smaller dimension than the observed time series. This situation

of a dynamic factor model (see, e.g., Molenaar, 1985) for a multivariate categorical time series would therefore be of special interest. The performance of the presented methods in this situation is a topic for future investigations. In addition, the methods can also be extended to the analyses of multiple cases. Such extensions can provide worthwile comparisons between latent processes of e.g., sleep stages for several infants. In particular, thresholds can then be fixed over time and over cases, whereafter different latent processes can be fitted seperately to each case.

In closing, it is stressed that all time series models considered in this paper are stationary. For many applications, stationarity cannot be assumed. The performance of the methods for nonstationary time series models remains to be assessed.

4

Logistic models for single-subject time series[1]

4.1 Introduction

Statistical methods in psychology are mostly applied to a collection of individuals rather than to a single one (Kratochwill, 1978, p. 3). The development of methods for psychological testing in the first half of the twentieth century put the individual on the background, because the initial objective of psychological testing was to differentiate among individuals. Keeping this in mind, the focus in the remainder of the twentieth century on advancement of statistical methods based on variation between individuals (inter-individual variation, IEV) instead of variation within a single individual, seems understandable. However, models for time-dependent variation of a single individual (intra-individual variation, IAV) have been widely available for some time. The discovery of the intrinsically stochastic time-dependent behavior within grains of pollen suspended in air (Brownian motion) led to the development of appropriate statistical models for single systems in the beginning of the 20th century. In this regard, the lack of interest in a pure $N = 1$ perspective in psychometrics seems remarkable.

It is not to say that examples of analyses of IAV are wholely absent in the psychometric literature. The measurement of (individual) change, for example, is a branch of psychometrics with a relatively long history. An early overview of problems encountered in measuring change can be found in Harris (1963). In that book, a single-subject analysis of multivariate time series is described by Holtzmann, who stressed that psychologists should study this type of analysis, in

[1] An earlier draft of this chapter appeared as van Rijn, P.W. & Molenaar, P.C.M. (2005). Logistic models for single-subject time series. In L.A. van der Ark, M.A. Croon, & K. Sijtsma(Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (pp. 125-146). London: Erlbaum.

view of the increasing importance of time series in other branches of science such as econometrics and biometrics (Holtzmann, 1963, p. 199). More recently, Nesselroade and Schmidt McCollam (2000) advocated analysing IAV in the context of developmental processes in psychology, and Collins and Sayer (2001) provided an overview of newly developed methods for the analysis of change.

Apart from the historical development, it is difficult to find an explicit and convincing rationale for the one-sided focus on IEV in contemporary psychometrics. The restriction to IEV appears to be considered to be an almost self-evident consequence of the scientific ideal to strive for general nomothetic knowledge. The science of psychology should involve theories and laws that apply to all human subjects. Such nomothetic knowledge would seem to be poorly served by intensive study of single subjects, because results thus obtained may not be generalizable in the intended sense. Despite its possible appeal, we will argue that this kind of rationale is incorrect in many instances by making use of a set of well-known mathematical theorems.

In our criticism of the one-sided focus on IEV, we do not take issue with the ideal of nomothetic knowledge, i.e., the search for psychological theories and laws that apply to all human subjects. Our criticism only concerns the assumption that theories and laws based on analysis of IEV apply to each human subject, and thus, would hold for IAV. To obtain valid theories and laws about IAV, one cannot generalize results derived from IEV, but one has to study IAV in its own right. That is the implication of the mathematical theorems to which we will refer. Having available the results of a sufficient number of individual analyses of IAV, one then can search for general characteristics by means of standard inductive techniques. If successful, this will yield valid nomothetic knowledge about the structure of IAV, i.e., nomothetic theories and laws about idiographic (individual) processes.

This chapter is divided into two parts. The first part starts out with a description of analyses for IEV and IAV, and the condition under which there exists a relationship between the two types of analysis, namely ergodicity. This condition is explained in the context of psychometrics.

Having thus set the stage for serious consideration of IAV, the second part of this chapter discusses latent variable models for single-subject time series data. Special attention is given to the logistic model for multivariate dichotomous time series, which can be seen, in its simplest form, as a dynamic variant of the Rasch model, where the person parameter is replaced by a person process. It must be stated that logistic models for repeated measurements have been discussed

by various authors (e.g., Kempf, 1977; Fischer 1983, 1989; Verhelst & Glas, 1993; Agresti, 1997). In most of these applications, however, the IAV nature of single-subject data is not emphasized as much as in the present chapter. The presented modelling approach is illustrated by examples with simulated data and an application to real data.

4.2 The ergodic notion in psychology

In its basic form, standard statistical analysis in psychology proceeds by drawing a sample of individuals, assessing their scores on selected measurement instruments, and then computing statistics by taking appropriate averages over the scores of all available individuals. If all individuals would yield the same score, statistical analysis would be severely reduced. Hence, it is the manner in which scores vary across subjects, IEV, which provides the information for the analysis. In contrast, in time series analysis the same individual subject is repeatedly measured, and statistics are computed by taking appropriate averages over the scores obtained at all measurement occasions. Hence, it is the manner in which an individual's scores vary across measurement occasions, IAV, which provides the information for time series analysis.

We already indicated that psychometricians are mainly interested in analyses of IEV. A vivid illustration of this tendency can be found in the classic treatise of test theory by Lord and Novick (1968). They define the concept of true score of a person as the mean of the distribution of scores obtained by independent repeated measurement of this person. This is obviously a definition in terms of IAV. Lord and Novick then remark that repeated measurement of the same person will affect this person's state and give rise to fatigue, habituation, or other confounding effects. They conclude that therefore, instead of measuring one person a large number of times, test theory has to be based on the alternative paradigm in which a large number of persons is measured once or twice. The shift to the latter alternative paradigm implies that test theory is based on analysis of IEV.

Notwithstanding that confounding factors such as habituation and fatigue might complicate the implementation of a purely IAV based test theory, a reference to such contingent states of affairs cannot be taken as the reason for the impossibility of this whole paradigm. In addition, Lord and Novick (1968, p. 32) state that the definition of true score in terms of IAV would be better suited for individual assessment than an IEV based test theory that is meant to differen-

tiate among individuals. It might therefore be expected that psychological tests constructed on the basis of analysis of IEV perform suboptimally when applied for the purpose of individual prediction. However, a task that still awaits further elaboration is the assessment of situations in which such test performance is suboptimal.

The urgency to determine the performance of standard tests in the context of individual assessment and prediction is all the more pressing given the strong justification for the conjecture that the differences between analysis of IEV and IAV go deeper than a mere difference in degree of success in the context of individual prediction. The reasons we have in mind are of two kinds: the implications of ergodic theorems, and results from mathematical biology suggesting the presence of substantial heterogeneity in human populations. Ergodic theory concerns the characterization of stochastic processes for which analysis of IEV and IAV yield the same results (e.g., Petersen, 1983). The classic ergodic hypothesis originates from statistical physics, and states that the average of a stochastic process over time is equal to the average of the ensemble of stochastic processes at a single point in time. By an ensemble is meant the possibly infinite number of hypothesized copies of a system. An ergodic hypothesis can also be stated for the variance or distribution of a stochastic process. In this sense, for ergodic processes the one-sided psychometric focus on analysis of IEV does not present any problem, because results thus obtained also are valid for individual assessment and prediction of IAV. Unfortunately, however, the criteria for ergodicity are very strict, and involve the absence of any time-dependent changes in the distributional characteristics of a stochastic process. Therefore, all developmental, learning and adaptive processes do not obey the criteria for ergodicity. For these classes of non-ergodic processes, there may not exist any lawful relationship between IEV and IAV.

Related to ergodicity is the notion of stationarity, which concerns the distributional characteristics of a single realization of a stochastic process. Stationarity amounts to the absence of time-dependent changes in distributional characteristics and, except for certain special cases of non-stationarity, is a condition for ergodicity.[2] For Gaussian processes, stationarity is a necessary condition for ergodicity. An example of a stationary process can be given by the notion of general intelligence in normal adults.[3] Now, it can be assumed that, under nor-

[2]Note that strict stationarity is meant here, see, e.g., Hamilton (1994, p. 45-46).

[3]For the sake of argument, let us neglect all problems associated with the theoretical status and operationalization of general intelligence.

mal circumstances, a normal adult's level of general intelligence does not change structurally over his lifespan. Administration of several intelligence tests over the lifespan will result in scores that vary, but probably only slightly. The distribution of intelligence scores in the first half of the measurements is likely to be equal to that of the second half, and the process can be considered to be stationary. However, if circumstances change drastically due to, for example, illness or excessive training on intelligence tests, this distribution is bound to change as well. The process then is no longer stationary.

Even if the distributional characteristics of a stochastic process are invariant in time, that is, the process is stationary, it still may be non-ergodic. The key difference between stationarity and ergodicity concerns the uniqueness of the so-called equilibrium distribution of a stochastic process, i.e., the distribution of the values of a stochastic process as time increases without bound. Each stationary process gives rise to an equilibrium distribution, but this equilibrium distribution may not be unique. Only if the process is ergodic, then this is necessary and sufficient for its equilibrium distribution to be unique (cf. Mackey, 1992, Theorem 4.6). Hence stationary processes are non-ergodic if they display a moderate kind of heterogeneity: their equilibrium distribution is not unique. Notice that this is the kind of non-ergodicity known from Markov chain theory (e.g., Kemeny, Snell, & Knapp, 1966). Already the presence of this moderate form of heterogeneity with respect to the equilibrium distribution implies the possibility of a lack of lawful relationships between IEV and IAV.

Now, let us return to our intelligence example. The general level of intelligence of an ensemble of normal adults is not likely to change structurally over time under normal circumstances. Yet, it is unlikely that all adults have the same distribution of intelligence scores over time, that is, there does not exist a unique equilibrium distribution. Thus, the ensemble of human adults in this case is non-ergodic, although the individual intelligence processes in this example can be considered stationary. If the ensemble of adults would be ergodic, the following odd statement would hold: "Five percent of the people score 125 or higher on an intelligence test, therefore five percent of the time your intelligence score is higher than 125".

There are strong indications that heterogeneity in human populations may be much more pervasive, transcending the moderate forms associated with non-ergodicity. Mathematical theory about biological pattern formation (e.g., Murray, 1993) and nonlinear epigenetics (Edelman, 1987) shows that growth processes are severely underdetermined by genetic and environmental influences. Conse-

quently, growth processes have to be self-organizing in order to accomplish their tasks. In particular the maturation of the central nervous system results from self-organizing epigenetic processes. Self-organization, however, gives rise to substantial endogeneous variation that is independent from genetic and environmental influences (Molenaar, Boomsma, & Dolan, 1993; Molenaar & Raijmakers, 1999). For instance, homologous neural structures on the left-hand and right-hand side of the same individual (IAV) can differ as much as the left-hand side of this neural structure in different individuals (IEV). Insofar as the activity of such heterogeneous neural structures is associated with the performance on psychological tests, this performance can be expected to be heterogeneous in much stronger forms than is the case with non-ergodicity.

It has been shown by means of simulation experiments as well as mathematical proof (Molenaar, Huizenga, & Nesselroade, 2003; Kelderman & Molenaar, 2007) that standard factor analysis of IEV is insensitive to the presence of substantial heterogeneity. For instance, it is an assumption of the standard factor model that factor loadings are invariant (fixed) across subjects. If, however, these factor loadings in reality varied randomly across subjects (a violation of the assumption of fixed factor loadings), then the standard factor model still fits satisfactorily. There appears to be only one principled way in which the presence of such heterogeneity can be detected, namely by carrying out replicated factor analyses of IAV (dynamic factor analysis of multivariate time series; Molenaar, 1985) and then compare the solutions thus obtained for distinct subjects.

In closing this section, it is reiterated that in general one cannot expect lawful relationships to exist between the structure of IEV and the structure of IAV. Such relationships can only be obtained under the restrictive condition that the processes concerned are ergodic. For non-ergodic processes, and in cases where human subjects are heterogeneous in even more pervasive ways (e.g., each subject having its personal factor model with its own distinct number of factors, factor loading pattern and/or specific variances), the use of IAV paradigms is mandatory. To accomplish this, appropriate time series analysis extensions of standard statistical techniques are required. Brillinger (1975) presents a rigorous derivation of time series analogues of all standard multivariate techniques (analysis of variance, regression analysis, principal component analysis, canonical correlation analysis). In the next section, we present an overview of time series analogues of latent variable models.

4.3 Latent variable models

From a general point of view, a stochastic process can be interpreted as a random function. That is, as an ensemble of time-dependent functions on which a probability measure is defined (e.g., Brillinger, 1975, section 2.11). Each time-dependent function of this ensemble is called a trajectory (or realization). Even if information is available about the entire past of a stochastic process up to some time t, then exact prediction for the next time point still is impossible. Each trajectory in an ensemble extends over the entire time axis. An observed time series, i.e., the particular stretch of values obtained by repeated measurement of a single subject, constitutes a randomly drawn trajectory from the ensemble, where this trajectory is clipped by a time window with width equal to the period of repeated measurement. In what follows we will denote a stochastic process by y_t and an observed time series thereof by y_t, $t = 1, \ldots, T$. We acknowledge that this notation is not entirely correct, but it is convenient and customary.

A subset of latent variable models for IAV is obtained by replacing all random variables in a standard latent variable model by stochastic processes. Bartholomew (1987) has given a useful classification of standard latent variable models based on two features: whether the observed variable is continuous or discrete and whether the common latent variable is continuous or discrete. This classification will be followed in our overview of latent variable models for IAV. There is an additional third feature which has to be considered for latent variable models for IAV, namely whether the time dimension is continuous or discrete. We will, however, restrict attention to models in discrete time only, as this is sufficient for our present purposes.

If both the observed variable and the common latent variable are continuous, the latent variable model is classified as a factor model. Replacement of all random variables in the linear factor model by continuous stochastic processes yields the linear state-space model: $y_t = Z_t\alpha_t + \epsilon_t$, wherein y_t is the observed continuous n-variate process, Z_t is a matrix of factor loadings, α_t is a common m-variate latent process (also called state process), and ϵ_t is a n-variate measurement error process. Statistical analysis of the state-space model is well developed and is treated in several text books (e.g., Durbin & Koopman, 2001). Hamaker, Dolan, and Molenaar (2003) discuss applications of the state-space model in psychological research.[4]

[4]Software for the fit of state-space models can be downloaded from: `http://users.fmg.uva.nl/cdolan/`.

If both the observed variable and the common latent variable are discrete, the latent variable model is classified as a latent class model. Replacement of all random variables in the latent class model by discrete stochastic processes yields the hidden Markov model (e.g., Elliott, Aggoun, & Moore, 1995). Visser, Raijmakers, and Molenaar (2000) present applications of hidden Markov modelling in psychological research.[5] If the observed variable is continuous and the common latent variable is discrete, the latent variable model is classified as a latent profile model (Bartholomew, 1987; Molenaar & von Eye, 1994). Replacement of the observed variable by a continuous stochastic process and the common latent variable by a discrete stochastic process yields a variant of the hidden Markov model (Elliott et al., 1995).

If the observed variable is discrete and the common latent variable is continuous, the latent variable model is classified as a generalized linear model. Replacement of all random variables in the generalized linear model by discrete (observed) and continuous (latent) stochastic processes yields a dynamic generalized linear model as described in Fahrmeir and Tutz (2001).

We will focus on a subset of dynamic generalized linear models. That is, models in which the observed process is dichotomous, related to a continuous latent process through the logistic response function.

4.4 A LOGISTIC MODEL FOR DICHOTOMOUS TIME SERIES

Dichotomous (or binary) time series can be modelled in various ways. If auxiliary information is available, regression models can be used. For dichotomous time series, such models are discussed in detail in Kedem and Fokianos (2002) and in Fahrmeir and Tutz (2001). Our focus is on modelling dichotomous time series using latent variables which is comparable to the modelling of dichotomous variables in item response theory (see also, Mellenbergh, 1994; Mellenbergh & Van den Brink, 1998). Since the latent variable is replaced by a stochastic process, this approach can be seen as a dynamic extension of item response modelling. As stated before, this approach is not new, although the emphasis on the modelling of IAV in this sense is novel. Modelling is pursued following the dynamic generalized linear modelling approach as described in Fahrmeir and Tutz (2001), that is, by specifying a distributional model, response function, linear predictor, and transitional model.

[5] Appropriate software can be found at: `http://users.fmg.uva.nl/ivisser/hmm`.

4.4.1 General outline

Consider the situation in which we have a dichotomously scored, multivariate time series, i.e., an n-dimensional observation vector y_t such that $y_t \in \{0,1\}^n$, at each time point $t = 1, \ldots, T$. Each single univariate observation y_{it}, $i = 1, \ldots, n$, follows a Bernoulli distribution with parameter π_{it} as the probability of obtaining a score one, given by

$$y_{it} \sim B(\pi_{it}) = \pi_{it}^{y_{it}}(1-\pi_{it})^{1-y_{it}}, \quad 0 < \pi_{it} < 1. \tag{4.1}$$

The probability π_{it} is modelled by inserting a linear predictor η_{it} into the logistic response function, resulting in

$$\pi_{it} = \frac{\exp(\eta_{it})}{1+\exp(\eta_{it})}. \tag{4.2}$$

Next, the n-dimensional linear prediction vector η_t is constructed by linking the $n \times m$ design matrix Z_t with the m-dimensional latent state vector α_t

$$\eta_t = Z_t \alpha_t. \tag{4.3}$$

A linear transition equation is assumed, which relates states at $t-1$ to t through the $m \times m$ transition matrix F_t, and is given by

$$\alpha_t = F_t \alpha_{t-1} + R_t \xi_t, \quad \xi_t \sim \mathrm{N}(0, Q_t). \tag{4.4}$$

The state vector α_t is allowed to contain time-invariant elements. The $m \times p$ selection matrix R_t is assumed to be a subset of the columns of the m-dimensional identity matrix I_m, so that it associates the elements of the p-dimensional disturbance vector ξ_t with the p time-varying elements of the state vector (Durbin & Koopman, 2001, p. 38).[6] The elements of ξ_t are generally referred to as innovations. In general, the state process is started up by an initial state α_0. The initialization is dependent on the type of process that is used. If all elements of the state vector are time-varying, the process can be initialized by α_0, which is normally distributed as follows

$$\alpha_0 \sim \mathrm{N}(a_0, Q_0). \tag{4.5}$$

In the above representation, the covariance matrices Q_t are nonsingular, which is somewhat more advantageous in estimation procedures (Durbin & Koopman,

[6] Notice that $m = p + n$ does not necessarily follow.

2001, p. 38). Note that a_0, Q_0, Z_t, F_t, R_t, and Q_t can contain parameters to be estimated. Methods for the estimation of these parameters are not well developed. However, Fahrmeir and Wagenpfeil (1997) describe an estimation procedure for the estimation of a_0, Q_0, and Q_t.

The following three assumptions are stated to completely specify the model in terms of densities. The first assumption is that current observations are dependent on current states only:

$$p(y_t|\alpha_t, \alpha_{t-1}, \ldots, \alpha_0, y_{t-1}, y_{t-2}, \ldots, y_1) = p(y_t|\alpha_t, y_{t-1}, y_{t-2}, \ldots, y_1).$$

The second assumption is that the state process is first order Markovian:

$$p(\alpha_t|\alpha_{t-1}, \ldots, \alpha_0) = p(\alpha_t|\alpha_{t-1}).$$

Finally, and in addition to assumption one, it is assumed that the multivariate observations are independent given the current state:

$$p(y_t|\alpha_t, y_{t-1}, y_{t-2}, \ldots, y_1) = \prod_{i=1}^{n} p(y_{it}|\alpha_t, y_{t-1}, y_{t-2}, \ldots, y_1).$$

Since the specific contents of the state and disturbance vector can be freely chosen, a variety of latent processes can be captured with the current representation. Depending on the hypothesized dynamic constellation of the latent process, one can choose between, for instance, autoregressive processes, moving average processes, and random walks (e.g., Hamilton, 1994). In addition, trends and cyclic change parameters can be included in the current representation. For now, we restrict the latent process to be either a white noise process, an autoregressive process, or a random walk.

The model can be extended to more than one person ($N > 1$), more than one latent process ($p > 1$), and also to multi-categorical or polytomous time series. However, the interest here lies in $N = 1$ and since results of analyses of dichotomous time series with this type of models are not widely available, we next consider some simple, yet illustrative modelling examples.

4.4.2 A DYNAMIC LOGISTIC MODEL

We now illustrate how a dynamic variant of the Rasch model can be obtained. We begin by contructing the state vector α_t, which consists of two parts. The first part describes a person's univariate latent process θ_t. The second part consists of n time-invariant threshold parameters, each associated with the corresponding

element of y_t, denoted by β. We consider three different processes, namely, a white noise process, a first-order autoregressive process, and a first-order random walk. These processes are given by, respectively,

$$\theta_t = \mu_\theta + \xi_t, \qquad \xi_t \sim \mathrm{N}(0, q), \tag{4.6}$$

$$\theta_t = \mu_\theta + \phi_1 \theta_{t-1} + \xi_t, \qquad \xi_t \sim \mathrm{N}(0, q), \tag{4.7}$$

$$\theta_t = \mu_\theta + \theta_{t-1} + \xi_t, \qquad \xi_t \sim \mathrm{N}(0, q), \tag{4.8}$$

where μ_θ is a time-invariant mean and ϕ_1 is the autoregressive parameter. Note that if $|\phi_1| < 1$, the autoregressive process is stationary. The random walk in Equation 4.8 can be perceived of as the discrete time analogue of Brownian motion (Klebaner, 1998, p. 80). It should be noted that the random walk process is nonstationary since $\mathrm{Var}(\theta_t) \to \infty$ as $t \to \infty$, and therefore non-ergodic. For each of the three processes, we have $\alpha_t = (\theta_t, \mu_\theta, \beta')'$ and $m = p + n + 1$. For simplicity and sufficiency for present purposes, the following model parameters and matrices are considered time invariant as well: the design matrix (Z), the transition matrix (F), the selection matrix (R), and the covariance matrix of the state disturbances (Q).

Consider now the situation in which we have four dichotomous variables. Modelling is pursued as follows. We have a 4-dimensional vector of observations y_t, a 4-dimensional probability vector π_t, and a 4-dimensional linear prediction vector η_t, related to each other as stated in Equations 4.1 and 4.2. The time invariant elements of the state vector do not need to be initialized, and the white noise process in Equation 4.6 does not either. Only the autoregressive process and the random walk have to be initialized with θ_0. The initial state and the 6-dimensional state vector then have the following form

$$\alpha_0 = R\theta_0 = \begin{bmatrix} \theta_0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \qquad \alpha_t = \begin{bmatrix} \theta_t \\ \mu_\theta \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}.$$

The specification of the design matrix defines the relation between the person

process and the threshold parameters, and is given by

$$Z = \begin{bmatrix} 1 & 1 & -1 & 0 & 0 & 0 \\ 1 & 1 & 0 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 & -1 & 0 \\ 1 & 1 & 0 & 0 & 0 & -1 \end{bmatrix}.$$

The logistic response function relates the linear predictor $\eta_t = Z\alpha_t$ to the probability vector π_t with elements

$$\pi_{it} = \frac{\exp(\theta_t - \beta_i)}{1 + \exp(\theta_t - \beta_i)}. \tag{4.9}$$

Equation 4.9 can be seen as a dynamic variant of the Rasch model. Note that this is the form of the Rasch model without the so-called item-invariant discrimination parameter (see Hambleton & Swaminathan, 1985, p. 47). For the random walk process, the transition matrix F is simply the 5×5 identity matrix, I_5. For the white noise process, F is the same except that its first element $F_{1,1}$ is zero. For the autoregressive process, the first element $F_{1,1}$ is equal to ϕ_1. The selection matrix R reduces to a vector r, because we have only a single time-varying parameter (θ_t), and is given by

$$r = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Estimation of the latent process is discussed next.

4.5 ESTIMATION

The iteratively weighted Kalman filter and smoother (KFS) as described in Fahrmeir and Wagenpfeil (1997) is used to obtain estimates of the latent process α_t. The KFS procedure maximizes the following log-posterior distribution

of the states α_t, $t = 1, \ldots, T$,

$$\begin{aligned}\log p(\alpha) = &\sum_{t=1}^{T}\sum_{i=1}^{n}\left(y_{it}\log(\pi_{it}) + (1-y_{it})\log(1-\pi_{it})\right)\\ &-\frac{1}{2}(\alpha_0 - a_0)'Q_0^{-1}(\alpha_0 - a_0)\\ &-\frac{1}{2}\sum_{t=1}^{T}(\alpha_t - F_t\alpha_{t-1})'R_tQ_t^{-1}R_t'(\alpha_t - F_t\alpha_{t-1}),\end{aligned} \tag{4.10}$$

For the white noise process, the initial state contribution can be eliminated from the log-posterior. Also, in the third part of Equation 5.11, only the time-varying elements (θ_t) of the state vector contribute to the log-posterior. However, this contribution differs for the three models that are used here, and therefore, the above representation is retained. In the presentation of the KFS procedure, the parameters or values in a_0, Q_0, Z, F, R, and Q are considered to be either known or fixed. In each iteration i of the KFS procedure, evaluation values for the state process are needed, which are denoted by $\tilde{a}^i = (\tilde{a}_1^{i\prime}, \tilde{a}_2^{i\prime}, \ldots, \tilde{a}_T^{i\prime})'$. Filtering and smoothing then proceeds as follows.

4.5.1 Filtering

First, the filter is initialized by

$$a_{0|0} = a_0 \quad \text{and} \quad V_{0|0} = R_0Q_0R_0'.$$

The extended Kalman filter consists of two recursive steps, a prediction and correction step, which are taken consecutively. The prediction step is described as follows

$$\begin{aligned}a_{t|t-1} &= F_t a_{t-1|t-1},\\ V_{t|t-1} &= F_t V_{t-1|t-1}F_t' + R_tQ_tR_t'.\end{aligned}$$

The correction step is given by the following equations

$$\begin{aligned}V_{t|t} &= (V_{t|t-1}^{-1} + B_t)^{-1},\\ a_{t|t} &= a_{t|t-1} + V_{t|t}b_t,\end{aligned}$$

where b_t and B_t are the so-called working score function and expected information matrix given by

$$\begin{aligned}B_t &= Z_t'D_t\Sigma_t^{-1}D_t'Z_t,\\ b_t &= Z_t'D_t\Sigma_t^{-1}(y_t - h(Z_t\tilde{a}_t^i)) + B_t(a_{t|t-1} - \tilde{a}_t^i),\end{aligned}$$

where $h(.)$ is the logistic response function and, for the modelling example considered in Section 4.4.2, D_t is the symmetric Jacobian matrix of the response function given by

$$D_t = \frac{\partial h(\eta_t)}{\partial \eta_t} = \begin{bmatrix} \pi_{t1}(1-\pi_{t1}) & & & \\ 0 & \pi_{t2}(1-\pi_{t2}) & & \\ 0 & 0 & \pi_{t3}(1-\pi_{t3}) & \\ 0 & 0 & 0 & \pi_{t4}(1-\pi_{t4}) \end{bmatrix},$$

and Σ_t is the covariance matrix of y_t given by

$$\Sigma_t = \begin{bmatrix} \pi_{t1}(1-\pi_{t1}) & & & \\ 0 & \pi_{t2}(1-\pi_{t2}) & & \\ 0 & 0 & \pi_{t3}(1-\pi_{t3}) & \\ 0 & 0 & 0 & \pi_{t4}(1-\pi_{t4}) \end{bmatrix}.$$

Both D_t and Σ_t are evaluated at $\tilde{a}_t^i$. Note that D_t and Σ_t are equal for our modelling example, although, in general, this is not necessarily the case.

4.5.2 SMOOTHING

The fixed interval smoother is a backward procedure to obtain state estimates $a_{t-1|T}$ utilizing the information of the complete time series. For $t = T, \ldots, 2$, we obtain

$$\begin{aligned} a_{t-1|T} &= a_{t-1|t-1} + G_t(a_{t|T} - a_{t|t-1}), \\ V_{t-1|T} &= V_{t-1|t-1} + G_t(V_{t|T} - V_{t|t-1})G_t', \end{aligned}$$

where

$$G_t = V_{t-1|t-1}F_t'V_{t|t-1}^{-1}.$$

After each application of the filter and smoother, the state evaluation values are updated with the results of the smoother, that is, $\tilde{a}^{i+1} = (a_{1|T}', \ldots, a_{t|T}')'$. The filter and smoother are applied repeatedly until some convergence criterion is reached. The stopping criterion of the KFS procedure used in the present investigation is $\max|\tilde{a}^i - \tilde{a}^{i-1}| < 1^{-12}$.

Table 4.1: Results on parameter estimation for simulated data

		WN		AR		RW	
Parameter	Value	Est.[1]	SE[2]	Est.	SE	Est.	SE
μ_θ	0.00	0.068	(0.086)	0.066	(0.148)	0.134	(0.276)
β_1	-1.50	-1.289	(0.148)	-1.268	(0.152)	-1.447	(0.160)
β_2	-0.50	-0.511	(0.134)	-0.532	(0.139)	-0.500	(0.146)
β_3	0.50	0.559	(0.136)	0.461	(0.140)	0.504	(0.153)
β_4	1.50	1.241	(0.151)	1.339	(0.158)	1.442	(0.182)

[1] Estimate

[2] Standard error

4.6 Examples

4.6.1 Simulated data

Data were simulated using the three models described in Section 4.4.2 with the following parameter settings. Each simulated time series has length $T = 200$. The threshold parameters for all three models are the same and given by

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} -1.5 \\ -0.5 \\ 0.5 \\ 1.5 \end{bmatrix}.$$

For identification purposes, μ_θ is kept fixed at zero. This parameter can then be deleted from the state vector, and associated model matrices Z, F, and R are of reduced dimension. The variance of the white noise process θ_t is fixed at one, so $q = 1$. The autoregressive parameter is $\phi_1 = 0.7$, and the variance of the autoregressive process θ_t is also fixed at one, which means that $q = 1 - \phi_1^2 = 0.51$. The variance of the innovations of the random walk process is $q = 0.01$. So we have four items, a single latent factor, a single person, and a series of length $T = 200$.

The results on parameter estimation for the three simulated data examples are given in Table 4.1. The estimated β's are rescaled, so that their mean is zero, and an estimate of μ_θ is obtained. In this example, the item parameters are recovered best with the RW process in terms of bias, but the standard errors of the item parameters are slightly larger than for WN and AR processes. The

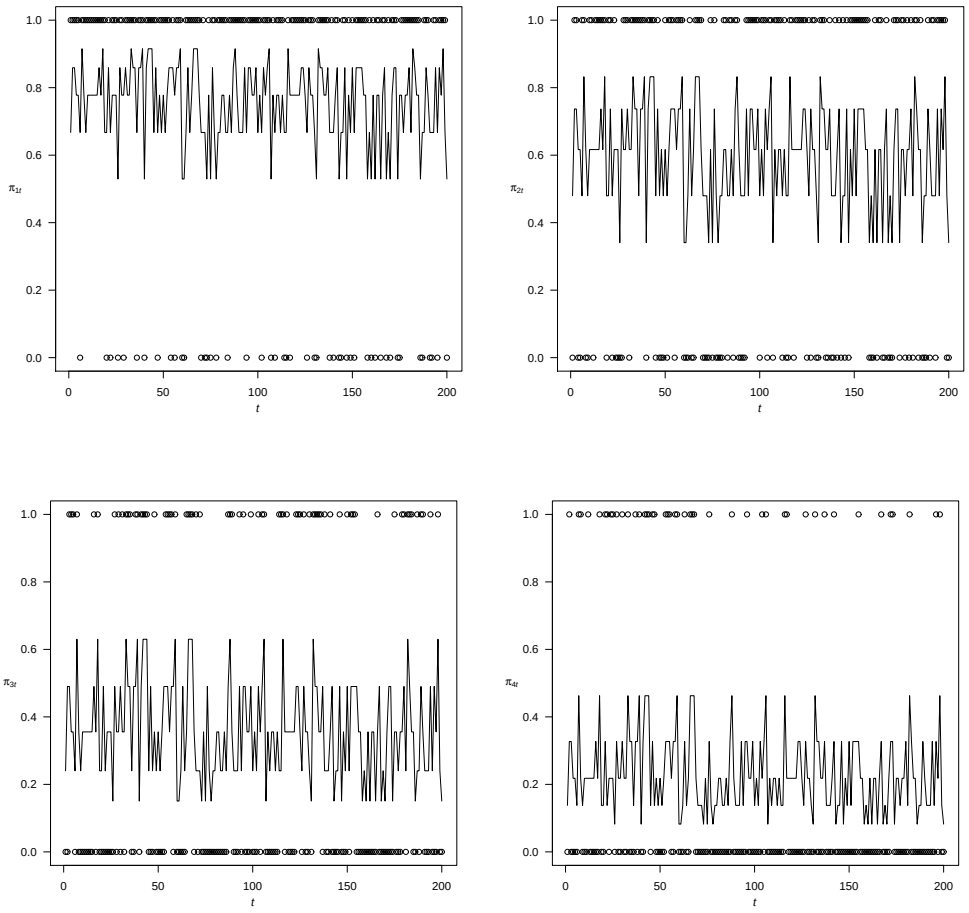


FIGURE 4.1: ESTIMATED PROBABILITIES WITH DOTTED OBSERVATIONS FOR WHITE NOISE PROCESS

differences in standard error of μ_θ for the three processes are remarkably large compared to those for the item parameters.

The results of filtered and smoothed probabilities, and filtered and smoothed states are presented in figures. Figure 4.1 displays the smoothed probabilities of the white noise process and Figure 4.2 displays the true and estimated latent process. Since the process is sequentially independent, the smoothed process can take on only five different values (the number of possible sumscores on the four items). This is seen in Figure 4.1, where each item has five probabilities and the probability values are dependent on the β's. Because the estimated latent process can take on only five different values, the true process is not well tracked, which can be seen in Figure 4.2.

Figure 4.3 displays the true and estimated probabilities for the AR example.

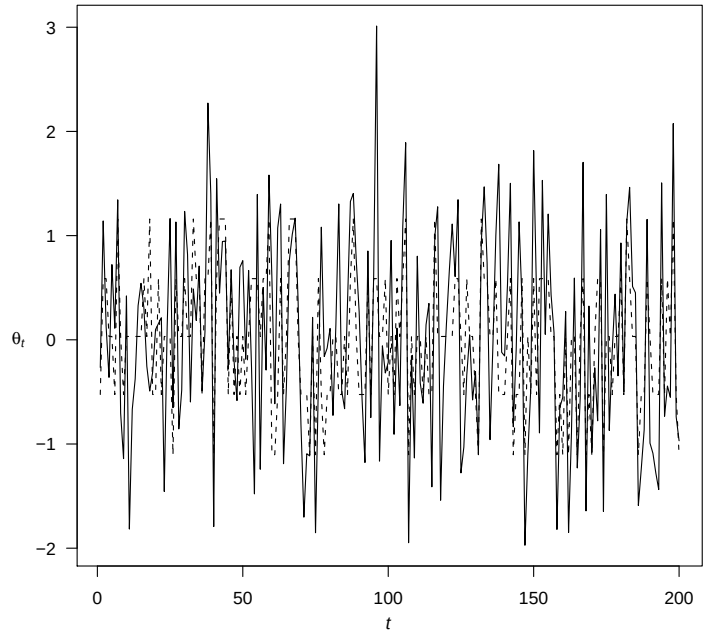


Figure 4.2: True (-) and estimated (- -) white noise process

Since the process is dependent on its first lag, the estimated latent process can take on many more values than the number of sumscores. The estimated probabilities track the true probability with reasonable accuracy. Figure 4.4 shows the true and estimated process paths. The true process is roughly tracked, although sometimes peaks or jumps seem difficult to recover.

Figures 4.5 and 4.6 show the results obtained with the random walk. Since a relatively small variance was selected for the innovations, the process is smoother than the first two examples. The estimated probabilities follow the true probabilities, although they tend to be too smooth. The divergence of the estimated latents process, however, is more severe, which is clearly seen in the middle of the time series in Figure 4.6.

4.6.2 Real data

Real data were analysed with the described techniques. We selected a single subject and a single subscale (Neuroticism) containing six items of a data set consisting of personality questionnaires containing 30 items scored on seven-point scales, administered to 22 psychology students on 90 consecutive days (Borkenau

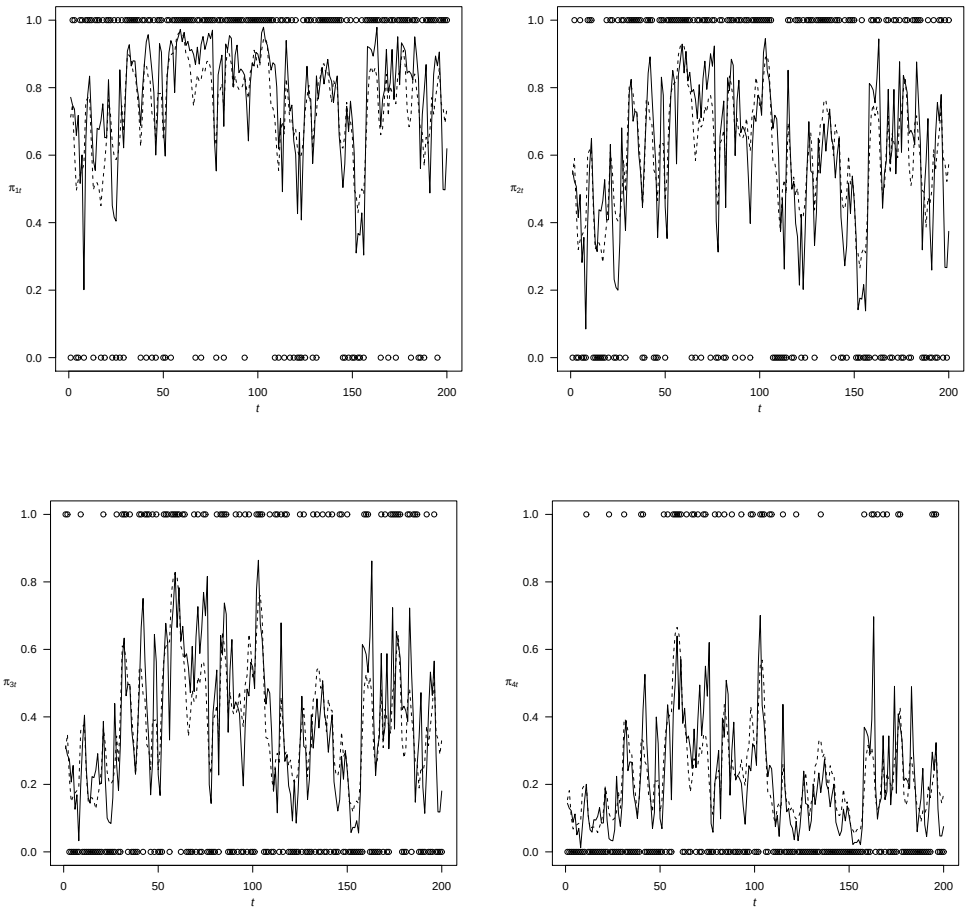


FIGURE 4.3: TRUE (-) AND ESTIMATED (- -) PROBABILITIES WITH DOTTED OBSERVATIONS FOR AUTOREGRESSIVE PROCESS

& Ostendorf, 1998).[7] The questionnaires were constructed as to measure the Big Five personality factors (McCrae & John, 1992). The data were dichotomized for illustrative purposes only in order to apply the dynamic logistic model.

The WN, AR, and RW processes were fitted to the observed time series. Since this example is for illustrative purposes, the parameters q and ϕ_1 were fixed at the same values as in the simulated data examples. The results on the estimation of the other parameters is displayed in Table 4.2. The estimates of μ_θ display the same pattern as in the simulated data. The item parameter estimates show a different pattern in that with the WN process, the estimates are somewhat more spread out than with the AR process and the RW process. The standard errors

[7]Data were kindly made available by professor Borkenau.

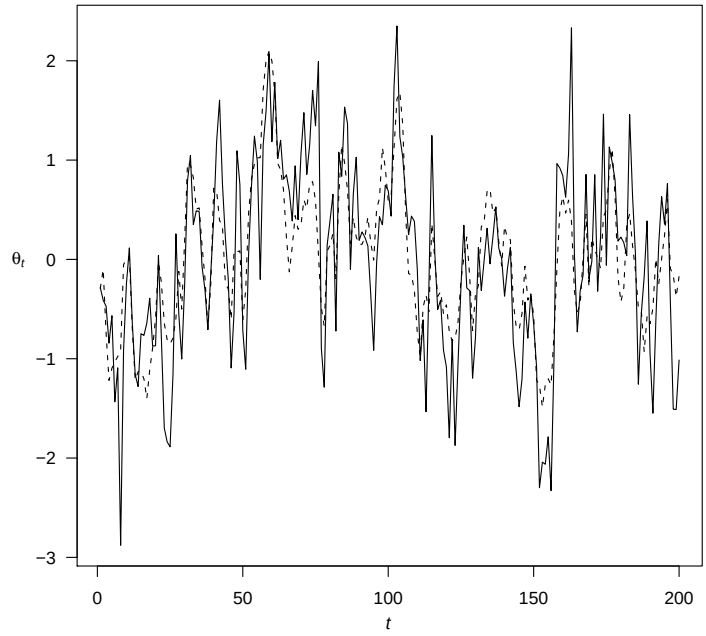


Figure 4.4: True (-) and estimated (- -) autoregressive process

of the item parameters do not differ much between the three processes.

Figure 4.7 shows the estimated latent WN, AR, and RW processes. On the left side, the results of the filter are displayed, and on the right side, the results of the smoother. Some observations can be made. For the WN process, there are as many values for the process as there are sumscores, but only after the smoother has been applied. The effect of smoothing is much larger for the RW process than for the other two processes in this example. The smoothed RW process displays an interesting swelling pattern which is difficult to unveil with the WN process. In the AR process, the pattern of the RW process is becoming visible.

4.7 Discussion

In the present chapter we took a closer look at the rationale for the emphasis in psychometrics on the analysis of IEV. It was found that this rationale is weak and that arguments for analysis of IAV are too easily brushed aside. We provided arguments for the development of models based on IAV. The question of the existence of any lawful relationship between analysis of IEV and IAV was addressed

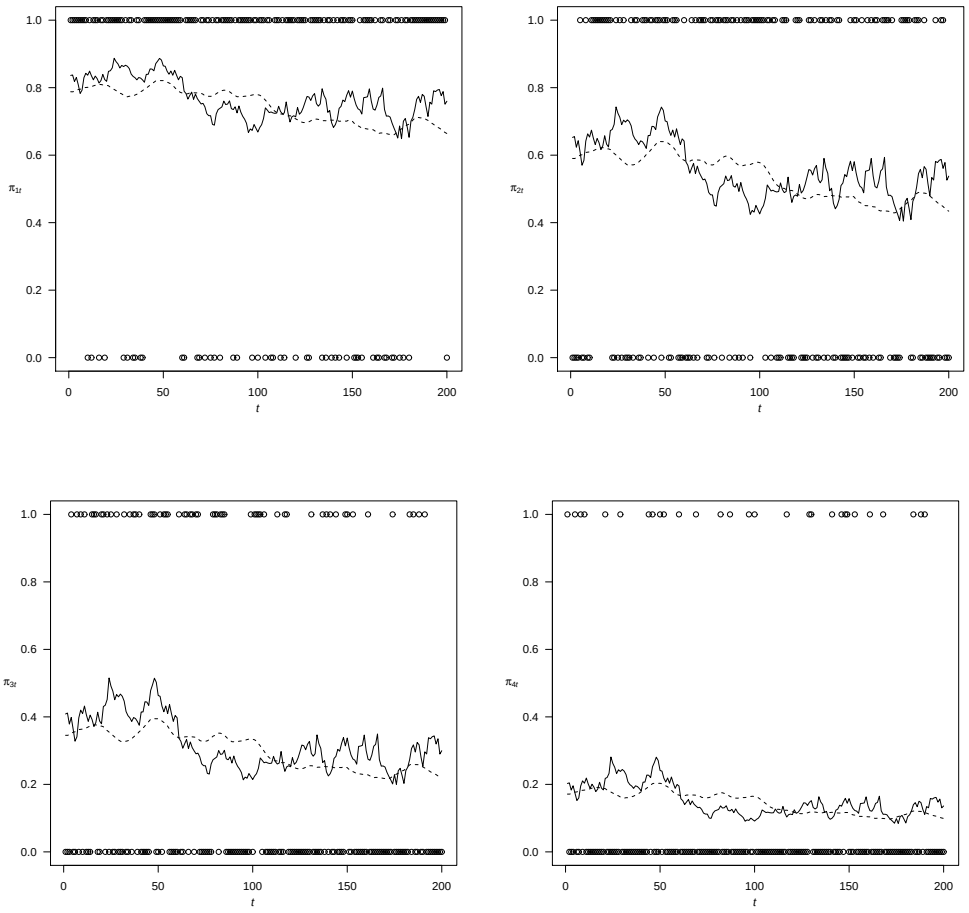


FIGURE 4.5: TRUE (-) AND ESTIMATED (- -) PROBABILITIES WITH DOTTED OBSERVATIONS FOR RANDOM WALK PROCESS

and it was argued that there are criteria for the existence of such a relationship. These criteria, however, are very strict and are met only when the processes concerned are ergodic. Since, in practice, little is known about the relation between analysis of IEV and IAV, and thus about ergodicity of the processes concerned in psychometrics, investigation of this relation is important. First, however, reliable methods have to be developed for analysis of IAV. The present chapter attempted to provide an outline of methods for analysing single-subject dichotomous time series.

An advantage of the presented modelling approach is that models for polytomous responses can be easily obtained after appropriate adjustments. In addition, it can be investigated if several persons can be analysed with the same model with equal parameter settings, i.e., if measurement invariance holds. How-

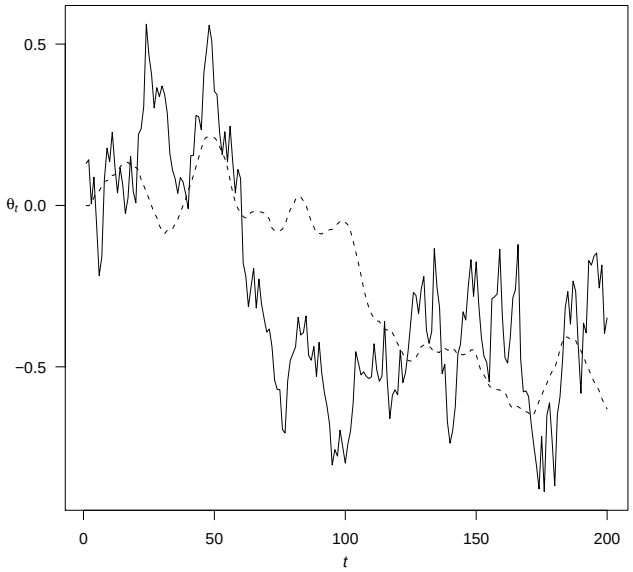


Figure 4.6: True (-) and estimated (- -) random walk process

ever, the results of the simulated and real data examples indicate that the discussed modelling outline requires further investigation. Important topics in this investigation are the development of estimation methods for variances and autoregressive parameters. Fahrmeir and Wagenpfeil (1997) discuss a procedure to estimate a_0, Q_0, and Q, but little is known about its behavior. Finally, in order to perform a full analysis of real data, methods to evaluate the fit of the discussed types of models become a necessary tool. Such methods await development and investigation.

Table 4.2: Results on parameter estimation for single subject neuroticism data

		WN		AR		RW	
Item	Parameter	Est.	SE	Est.	SE	Est.	SE
	μ_θ	0.276	(0.123)	0.319	(0.222)	0.637	(0.319)
Irritable (+)	β_1	-0.117	(0.214)	-0.109	(0.207)	-0.100	(0.202)
Emotionally stable (-)[1]	β_2	0.211	(0.217)	0.193	(0.210)	0.176	(0.205)
Calm (-)	β_3	0.046	(0.215)	0.041	(0.208)	0.037	(0.203)
Bad-tempered (+)	β_4	0.799	(0.230)	0.738	(0.224)	0.679	(0.219)
Resistant (-)	β_5	-0.714	(0.216)	-0.655	(0.210)	-0.601	(0.204)
Vulnerable (+)	β_6	-0.225	(0.214)	-0.208	(0.207)	-0.191	(0.202)

[1] Items with a minus are negatively formulated and therefore recoded.

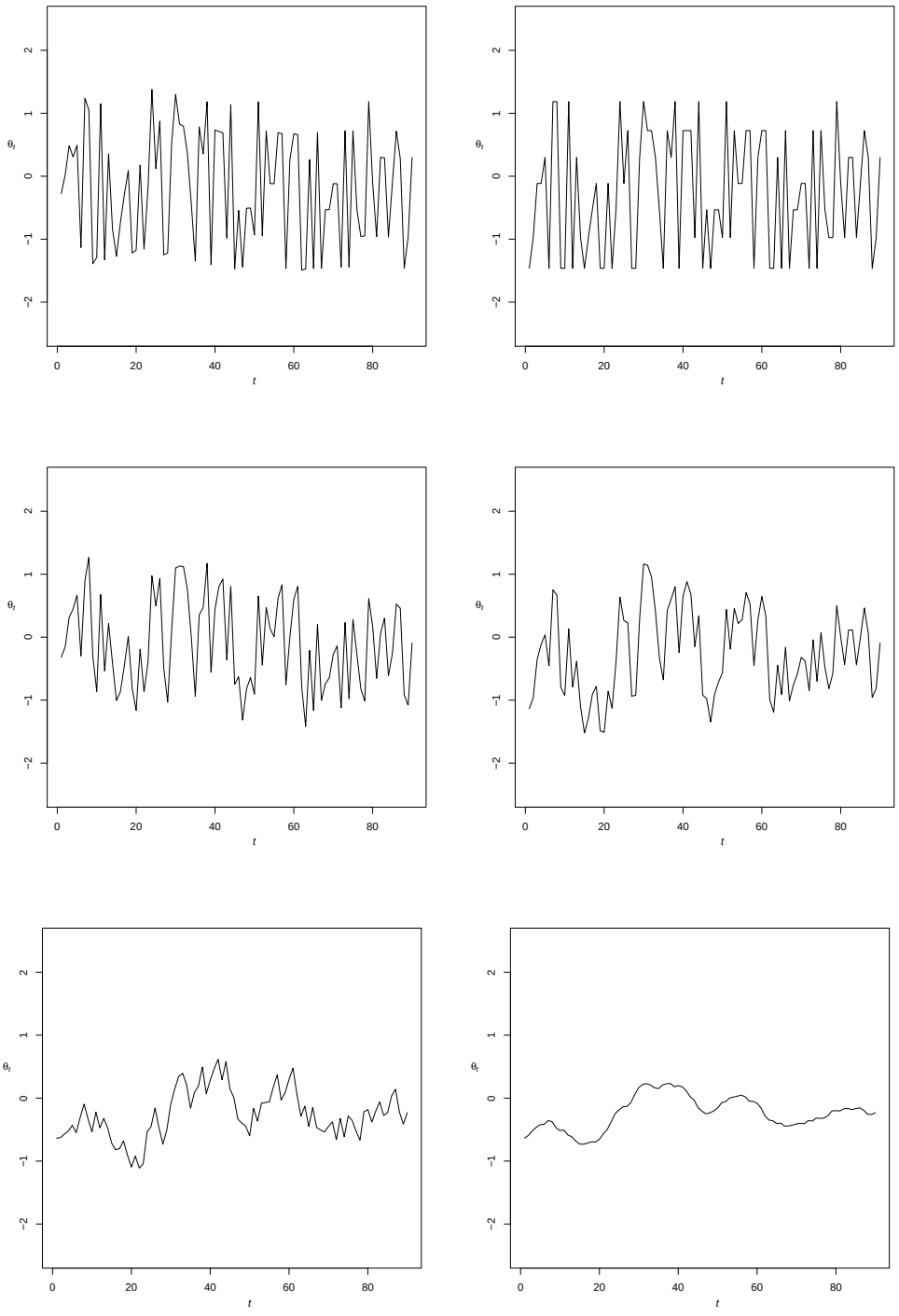


FIGURE 4.7: FILTER (LEFT) AND SMOOTHER (RIGHT) RESULTS OF FITTING WHITE NOISE (TOP), AUTOREGRESSIVE (MIDDLE), AND RANDOM WALK (BOTTOM) PROCESSES TO SINGLE SUBJECT NEUROTICISM DATA

5

State space methods for item response modelling

5.1 Introduction

Item response theory (IRT) is an important area of research in psychometrics. It provides a theoretical framework for the construction, application, and evaluation of tests in psychological and educational measurement. Within contemporary IRT, probabilistic models are used to describe the relations between observable item responses and unobservable psychological traits or abilities (Hambleton & Swaminathan, 1985). A primary purpose of IRT, and of test theory in general, is to describe the differences in item responses and test scores between individuals, because most applications are more suited for describing groups than individuals (Lord & Novick, 1968, p. 32). The purpose of this paper is to describe IRT models that are based on differences in item responses within individuals to whom is administered the same test repeatedly over time, that is, to describe dynamic IRT models. For the analysis of this type of measurements, the state space framework is used (Durbin & Koopman, 2001). As demonstrated, the state space methodology is capable of handling the analysis of both standard and dynamic IRT models in a straightforward manner. In addition, this framework can simultaneously handle the analysis of differences between and within individuals.

Most IRT models are developed to account for the variation in item responses that arises when items are administered to different individuals. The argumentation for the development of such models is to explain the differences between, rather than within, individuals. It can be argued that, at least on a theoretical level, this focus is one-sided, and that it is of interest to develop IRT models that account for the variation within individuals. For a thorough argumentation in favor of developing models for describing variation within individuals, see Molenaar (2004) and the subsequent discussion. One of the arguments for pursuing the development of such models is that a model that provides an informative account of inter-individual differences, is not necessarily valid at the level of intra-individual differences, as observed in repeated measures obtained

from a single individual (Kelderman & Molenaar, 2007; Borsboom, Mellenbergh, & van Heerden, 2003). While the ultimate goal of a standard application of IRT is to make inferences about a person's position on one or more latent dimensions of inter-individual differences, such inferences are not necessarily correct at the intra-individual level. More specifically, the constitution of the latent space of inter-individual traits or abilities that one intends to measure may be either incomplete or over-complete at the intra-individual level. That is, while ignoring a latent dimension contributing to an individual's development on the one hand, too much importance might be attached to certain factors which can be relevant for only some individuals on the other. In addition, a person's latent position can change over time, and individual differences can exist in the structure of such changes.

The above argumentation can be related to a discussion of two commonly used interpretations of probability in IRT (see Fischer & Molenaar, 1995; Holland, 1990). In describing the two views, a single administration of a single specific test is considered. The first view is the so-called random-sampling view, in which the only source of variation is the random sampling of individuals. In this view, individuals conceivably possess fixed, yet unknown response patterns. By sampling individuals at random from a population with a certain latent position, variation in response patterns is likely to occur. The probability of a response pattern can then be viewed as the proportion of individuals from that population that would show that pattern under repeated random sampling. The second and more popular view is the stochastic-subject view. In this view, each individual possesses a certain latent position, and probability is defined by the proportion of response patterns obtained by hypothesized repeated and independent administrations of the same test under the exact same circumstances.

This view is adopted when defining IRT models from variation within individuals over time, except that the administrations are not independent and circumstances are subject to change. In defining a dynamic IRT model in this sense, an individual can at first instance be considered as fixed, and is therefore no source of variation. Variation in response patterns may arise from different administrations of the same test, and an analogous argumentation can be advanced in this case. An immediate practical problem occurs, since these administrations can only be performed sequentially. As a consequence, they cannot be considered to be independent, and this sequential dependence should be accounted for in the model. As such, we are dealing with a time series, and methods to model the dependence and changes explicitly are widely available (Hamilton, 1994). Stan-

dard IRT models can be built from the random-sampling view (Holland, 1990), but when it comes to the interpretation of the characteristics that desribe an individual, one often relapses to the stochastic-subject view (Fischer & Molenaar, 1995). Ellis and van den Wollenberg (1993) have argued that this shift is justified only if the condition of local homogeneity holds. Local homogeneity amounts to invariance of the distribution of the item responses conditional on the latent variables accross all possible subpopulations (Ellis & van den Wollenberg, 1993, Theorem 2).

IRT models in which intra-individual variation is explicitly accounted for provide a different view on matters, because they start out from the smallest possible subpopulation, that is, a single individual. An important topic that arises when one is interested in models for intra-individual variation is that of stationarity. A second topic which is important when one is concerned with more than one individual, is that of ergodicity. Stationarity concerns the lack of time dependence of the distribution of a single time series (Hamilton, 1994). Ergodicity concerns the distributional properties of an ensemble of time series. The notion of ergodicity, to which we adhere, can be formulated as follows: If the model is stationary and the same for each individual, the collection of individuals is considered ergodic. This means that individuals are interchangeable and that an analysis of a collection of individuals on a single time point provides the same results as an analysis of a randomly selected single individual on a collection of time points. Generalizations can then be conducted either way. Consider the following example, if a model with a univariate latent process can be used to explain the response patterns obtained by repeated administrations of the same test to a single individual, and the model parameters and the distribution of this process does not change over time, then the condition of stationarity is met. Now, if the same model and univariate latent process holds for the collection of individuals, the condition of ergodicity is said to be met.

Thus, when the collection of individuals is ergodic, the two views on probability in item response modelling coincide. In this sense, it is agreed that local homogeneity is necessary to abide by the stochastic-subject view. In contrast, if ergodicity holds, individual predictions based on interindividual analysis are justified as well as interindividual statements based on a single intra-individual analysis. However, human ensembles are likely to be non-ergodic in many aspects (see Chapter 4 of this thesis).

In comparing the two types of variation within IRT, two implications of the ergodic notion are pertinent. First, the number of latent dimensions best describing

the item responses can differ between individuals. In this case, ergodicity does not hold, and each individual or cluster of comparable individuals needs to be analysed seperately. Second, even if the latent processes of two persons are stationary and of the same dimension, the specific time dependencies can be different. For example, an autoregressive and a white noise process can both be stationary and univariate. If this combined model proves to be best fitting on the data of two individuals, can the interpretation of the two latent processes be equated? In terms of the psychological content of the processes, there is no definitive answer. This issue complicates matters substantially in comparing individual differences in variation over time.

Many models for dynamic testing and measuring change have been described (e.g., Fischer, 1989; Embretson, 1994), and some of which are of special interest here. Within the framework of IRT, Kempf (1977) dropped the assumption of local independence. He derived sufficient statistics for the person parameters by conditioning on the sumscore of previous responses. Verhelst and Glas (1993) circumvented dropping the assumption of local independence by deftly manipulating the concept of incomplete designs. They develop a dynamic Rasch model and a marginal maximum likelihood estimation procedure. In the present chapter, we describe a dynamic model for responses to repeatedly administered items in which latent variables are generalized to latent processes. An extended version of the Kalman filter is used to estimate the parameters, which is based on the posterior distribution of so-called states of a state space model (Fahrmeir, 1992; Fahrmeir & Wagenpfeil, 1997).

The outline of this chapter is as follows. First, the Rasch model and the partial credit model are discussed followed by a description of commonly used parameter estimation methods and the assessment of model fit. Next, the dynamic versions of the Rasch and partial credit models are introduced. Hereafter, a description follows of how to specify both standard and dynamic IRT models as a state space model. Then, the estimation by means of Kalman filtering and smoothing is discussed. The methods are illustrated by two example data sets, one standard IRT analysis and one dynamic IRT analysis. This chapter ends with a discussion.

5.2 Standard IRT

5.2.1 Models

In a basic IRT setting, a unidimensional latent variable θ is assumed to be related to the probability of the responses y on a test of n items by a monotonically increasing function. This function is referred to as the item response function (IRF). The latent variable θ refers to some unobserved psychological trait of an individual, e.g., math ability, which the items are supposed to measure, and describes the differences between individuals. The characteristics of an item can be described by one or more parameters. Rasch (1960) developed such a model for dichotomous item responses, that is, when the elements of y can each be scored as either 0 or 1. In this model, the probability that a person with a certain θ responds to item i with threshold parameters β_i with 1 is determined by the logistic function as follows

$$p(y_i = 1|\theta) = p_i(\theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)}, \qquad i = 1, \ldots, n. \tag{5.1}$$

If a person's θ exceeds the items threshold β_i, the response 1 is more likely, and if β_i exceeds θ, the response 0 is more likely. This model has been used extensively in educational settings in which dichotomous item responses can often be scored as incorrect (0) or correct (1). The parameter β_i has therefore acquired the interpretation as item difficulty. The model in Equation 5.1 has come to be known as the Rasch model (RM), although it is also referred to as the one-parameter logistic model (Hambleton & Swaminathan, 1985). The RM possesses the property of specific objectivity which states that the comparison of items is only dependent on the difference in difficulty (β), and in turn, the comparison of persons is only dependent on the difference in ability (θ). For two different items and two different persons, specific objectivity is obtained by simplifying the following log odds ratios,

$$\log\left(\frac{\frac{p_1(\theta)}{1-p_1(\theta)}}{\frac{p_2(\theta)}{1-p_2(\theta)}}\right) = \beta_2 - \beta_1 \qquad \text{and} \qquad \log\left(\frac{\frac{p_i(\theta_1)}{1-p_i(\theta_1)}}{\frac{p_i(\theta_2)}{1-p_i(\theta_2)}}\right) = \theta_1 - \theta_2.$$

The Rasch model for dichotomous items can be extended to allow for more differences between items than the difficulty alone. For instance, differences in steepness of the IRFs can be represented in the model by the inclusion of a discrimination parameter. It can also be extended to allow for guessing if ability items with a closed response format are concerned. For the logistic model, these

extensions were developed by Birnbaum (1968). Rasch and others extended the models to allow for items with more than two responses, that is, for polytomous items (Samejima, 1969; Andersen, 1995). Since these extensions not only allow for polytomously scored ability items, but also for the use of Likert scale formats, IRT models have found their way outside the realm of ability testing. Apart from educational settings, applications can nowadays be found in clinical psychology, personality testing, and attitude measurement (see Embretson & Reise, 2000; Van der Linden & Hambleton, 1997).

One polytomous extension of the dichotomous Rasch model is the rating scale model discussed by Andrich (1978a, 1978b). Masters (1982) obtained a somewhat more general extension for items with ordered categories scored from $k = 0, \ldots, q$. This model is known as the partial credit model (PCM), and it is used here. Without loss of generality, it is assumed throughout that all items have an equal number of categories, that is, $q + 1$. Then, the conditional probability that a person with a certain θ responds to item i with threshold parameters $\beta_i = \{\beta_{i1}, \ldots, \beta_{iq}\}$ with response k is determined by the logistic function as follows

$$p(y_i = k|\theta) = p_{ik}(\theta) = \frac{\exp \sum_{v=0}^{k}(\theta - \beta_{iv})}{\sum_{c=0}^{q} \exp \sum_{v=0}^{c}(\theta - \beta_{iv})}, \quad k = 0, 1, \ldots, q, \quad i = 1, \ldots, n, \tag{5.2}$$

where $\sum_{v=0}^{0}(\theta - \beta_{iv}) \equiv 0$. The above function is referred to as the k-th item category response function (ICRF). The following representation can be used as well

$$p_{ik}(\theta) = \frac{\exp(k\theta - \sum_{v=0}^{k} \beta_{iv})}{1 + \sum_{c=1}^{q} \exp(c\theta - \sum_{v=1}^{c} \beta_{iv})},$$

in which case we define $\beta_{i0} \equiv 0$. For $k = 1, \ldots, q$ it holds that, given that a person responded to item i in category k or $k - 1$, the probability of response k is governed by the Rasch model, that is,

$$p(y_i = k|y_i = k \text{ or } k - 1, \theta) = \frac{\exp(\theta - \beta_{ik})}{1 + \exp(\theta - \beta_{ik})}.$$

The item category parameters $\beta_{i1}, \ldots, \beta_{iq}$ are interpreted as threshold locations on the θ-scale (see Masters & Wright, 1997). If θ exceeds β_{ik}, then category k becomes more probable than category $k - 1$. If not, category $k - 1$ is more probable than category k. The location on the θ-scale at which $\theta = \beta_{ik}$ indicates that the adjacent ICRFs intersect. Note that the item category parameters do not need to be ordered as is the case in other polytomous item response models

such as the graded response model developed by Samejima (1969). The RM and its extensions such as the PCM are exponential family models, and therefore have certain favorable properties such as sufficient statistics for their parameters.

5.2.2 Estimation

In the estimation of item parameters, the person parameters θ and item parameters β are referred to as incidental and structural parameters, respectively (Fischer & Molenaar, 1995; Neyman & Scott, 1948). This is due to the fact that the number of person parameters increases as the sample size N increases, whereas the number of item parameters does not. With the increase in sample size, the item parameter estimates improve in terms of accuracy and precision, whereas the person parameters do not. Reiterating that the final object of any psychological or educational test is to make inferences about persons, the sometimes used term nuisance parameter for θ is only meaningful in the phase of item parameter estimation. In fact, van der Linden and Hambleton (1997, p. 5) rather refer to the person parameter as structural parameters, and the item parameters as nuisance parameters. Needless to say, if test length is increased, person parameter estimates improve, and item parameter estimates do not. Item and person parameters are usually estimated seperately, and, consequently, item parameter estimation methods are distinguished from person parameter estimation methods. We discuss some commonly used methods in the following.

Generally, both item and person parameter estimation methods are based on the log-likelihood function of the observed responses, possibly in combination with some appropriate prior distribution of the item and person parameters. Given a set of n dichotomous items administered to N persons, the log-likelihood function of $\theta = (\theta_1, \ldots, \theta_N)$ and $\beta = (\beta_1, \ldots, \beta_n)$ for the RM can be written as

$$\begin{aligned} \log L(y; \theta, \beta) &= \sum_{j=1}^{N} \sum_{i=1}^{n} y_{ij} \log(p_i(\theta_j)) + (1 - y_{ij}) \log(1 - p_i(\theta_j)) \\ &= \sum_{j=1}^{N} \sum_{i=1}^{n} y_{ij}(\theta_j - \beta_i) - \log(1 + \exp(\theta_j - \beta_i)). \end{aligned} \tag{5.3}$$

For polytomous items, the log-likelihood function is obtained as follows. Let the polytomous response y_{ij} be replaced by a dummy vector of length q of which each element y_{ijk}, $k = 1, \ldots, q$ is scored 1 if category k is observed and 0 otherwise.

Then, the log-likelihood function for polytomous PCM items can be given by

$$\log L(y;\theta,\beta) = \sum_{j=1}^{N}\sum_{i=1}^{n}\left(\sum_{k=1}^{q} y_{ijk}\log(p_{ik}(\theta_j)) + (1-\sum_{k=1}^{q} y_{ijk})\log(1-\sum_{k=1}^{q} p_{ik}(\theta_j))\right). \tag{5.4}$$

The three most commonly used item parameter estimation methods for the RM and the PCM are joint maximum likelihood (JML), conditional maximum likelihood (CML) and marginal maximum likelihood (MML) estimation (see Molenaar, 1995; Andersen, 1995). JML estimation is an iterative procedure in which both item and person parameters are estimated. First, starting values for the item parameters are considered as fixed and the person parameters are estimated by maximizing the log-likelihood function. In turn, the obtained person parameter estimates are considered as fixed, and the item parameters are estimated again by maximizing the log-likelihood. This procedure is repeated until the estimates of both sets of parameters have converged. The JML procedure cannot be applied when either persons or items obtain minimum or maximum scores, because the log-likelihood has no finite maximum in that case. For polytomous PCM items, JML estimation also poses problems when categories have zero frequencies, so-called null categories (see Wilson & Masters, 1993). Bertoli-Barsotti (2005) stated necessary and sufficient conditions for existence and uniqueness of JML item parameter estimates for the PCM for this case.

CML estimation is a procedure that makes use of the fact that the Rasch model and its extensions are exponential family models, in which case a persons sumscore is a sufficient statistic for θ. By conditioning on this sufficient statistic, θ can be removed from the likelihood and the item parameters are estimated by maximizing the thus obtained conditional likelihood (Andersen, 1972). Again, estimation problems arise in cases of minimum or maximum item scores and null categories.

In the MML estimation procedure, the likelihood is multiplied by a probability density function for θ that is assumed over persons. The normal density is an obvious choice, yet other distributions can be used as well (see Thissen, 1982). The MML procedure generally starts out by integrating θ out of the product of the likelihood and the assumed density, and the thus obtained marginal likelihood is maximized to find item parameter estimates. Since the integration cannot be performed exactly, one usually resides to some kind of numerical approximation such as Gauss-Hermite quadratures. An advantage of this procedure is that item parameter estimates can be found for extreme item scores as well as null

categories.

In comparing the three methods, there is no obvious advantage, although some comparisons can be made (see also, Holland, 1989). The CML and MML procedures are known to produce consistent estimates, whereas the JML procedure does not necessarily do so. CML estimation has attractive asymptotic features, and does not need to make assumptions about the distribution of θ. On the other hand, some information can be lost in CML estimation (see Holland, 1989; Eggen, 2000). On the whole, MML estimation is most widely applicable, whereas CML estimation possesses the most attractive properties when applicable. However, note that the advantages of MML are largely due to additional assumptions, which are not necessary to apply the JML and CML estimation procedures.

Except for the JML procedure, item parameters are generally estimated first. Then, these estimates are considered as fixed and θ is estimated by some appropriate procedure. It should be noted that CML estimation is also applicable to the estimation of person parameters for Rasch models, although this is unusual and computationally cumbersome. Commonly used methods for the estimation of θ with fixed item parameters are maximum likelihood (ML), weighted maximum likelihood (WML), or a Bayesian estimation procedure such as Bayesian modal (BM) estimation (see also, Hoijtink & Boomsma, 1995). In ML estimation θ, the log-likelihoods in Equations 5.3 and 5.4 are simply maximized while keeping item parameters fixed. Again, no estimates are available for extreme response patterns. Warm (1989) developed an alternative procedure named WML which overcomes this problem, and in which the likelihood is multiplied by a weight function involving the test information function and then maximized with respect to θ. In BM estimation, the likelihood function is multiplied by a distribution function for θ and the mode of the resulting posterior is used as an estimate for θ. If a standard normal distribution is used, the log posterior can be written as (Hambleton & Swaminathan, 1985, p. 94)

$$f(\theta|y) \propto \log L(y;\theta,\beta) - \frac{1}{2}\sum_{j=1}^{N}\theta_j^2.$$

Tsutakawa and Johnson (1989) discussed a person parameter estimation procedure which accounted for the uncertainty in the item parameter estimates.

5.2.3 EVALUATION

In general, a thorough evaluation of the fit of an IRT model consists of several stages, and many aspects of the model and estimation procedure can be tested. Glas and Verhelst (1995) distinguished three aspects of evaluating model fit, which are the assumptions and properties of the selected model and estimation procedure, the type of statistic, and the mathematical refinement of the method. One can select appropriate fit measures on the basis of which aspects are deemed important for a particular application of IRT.

The basic assumptions such as the (uni)dimensionality of the model and local independence are checked first. For RMs combined with CML estimation, sufficiency of the sumscore can be tested. If MML is used, fit measures testing the appropriateness of the assumed distribution for θ can be used. In addition, properties pertaining to specific models can be investigated. If the RM holds, there exist no differences in discrimination between the items, and, in ability testing, the probability of guessing the correct answer is minimal. Such properties can for instance be tested with likelihood ratio tests. Furthermore, the fit of certain elements of the model can be studied seperately. For example, the fit of a specific item can be investigated, and person fit statistics can be constructed to detect aberrant response behavior.

Since a full discussion of model fit procedures is beyond the purpose of this chapter, we only discuss procedures to evaluate overall fit and to inspect item and person fit that we use in our analyses later on. Measures of overall goodness of fit are often based on differences between observed and expected response patterns. Usually, some approximation to the χ^2 distribution is calculated from the n-dimensional contingency table and the model used. However, such contingency tables rapidly become very large as the number of items and categories increase. That is, the table rapidly becomes sparse, and the χ^2 approximations are only valid if the sample size is very large which in turn increases the power of the test. In addition, the implications of rejecting the null hypothesis depend on the model and estimation procedure.

Within the realm of MML, for each possible response pattern, an expected frequency can be calculated on the base of the marginal probabilities. These expected frequencies can be obtained with the following marginal probability of a given response vector y

$$p_{\text{MML}}(y) = \int_{-\infty}^{\infty} \left(\prod_{i=1}^{n} \prod_{k=1}^{q} p_{ik}(\theta)^{y_{ik}} \left(1 - \sum_{k=1}^{q} p_{ik}(\theta) \right)^{1-\sum_{k=1}^{q} y_{ik}} \right) f(\theta) d\theta, \qquad (5.5)$$

where the integral can be approximated numerically by, say, Gauss-Hermite quadrature. A simple χ^2 approximation such as the Pearson X^2 or the likelihood-ratio statistic G^2 is easily computed by

$$X^2 = N \sum_y \frac{(p_{\text{OBS}}(y) - p_{\text{MML}}(y))^2}{p_{\text{MML}}(y)},$$
$$G^2 = 2N \sum_y p_{\text{OBS}}(y) \log \left(\frac{p_{\text{OBS}}(y)}{p_{\text{MML}}(y)} \right),$$

where $p_{\text{OBS}}(y)$ is the observed proportion of response pattern y, and the summation occurs over all possible response patterns, that is, over all cells in the n-dimensional contingency table. Note that the observed proportion is required to be larger than zero in order to compute G^2.

Masters and Wright (1997) discuss infit and outfit measures for inspecting item and person fit. For each response, a standardized residual can be calculated from the observed and expected response by

$$z_{ij} = \frac{y_{ij} - \text{E}(y_{ij}|\theta)}{\sqrt{\text{Var}(y_{ij}|\theta)}}, \tag{5.6}$$

where $\text{E}(y_{ij}|\theta)$ and $\text{Var}(y_{ij}|\theta)$ are the expected response and variance given by

$$\text{E}(y_{ij}|\theta) = \sum_{k=0}^{q} k p_{ik}(\theta_j), \quad \text{Var}(y_{ij}|\theta) = \sum_{k=0}^{q} (k - \text{E}(y_{ij}|\theta))^2 p_{ik}(\theta_j).$$

Note that the probability $p_{ik}(\theta_j)$ can be calculated in different ways depending on which estimation procedure was used to estimate the item and person parameters. Item and person fit indices can be obtained by

$$u_i = \sum_{j=1}^{N} z_{ij}^2 \quad \text{and} \quad u_j = \sum_{i=1}^{n} z_{ij}^2,$$

respectively. The asymptotic distribution of the above measures is unknown and no rules of thumb are available for the interpretation of their values. Yet, it is safe to say that items and persons with relatively large residuals require close inspection if overall fit measures like the X^2 and G^2 indicate a bad fit.

5.3 Dynamic IRT

5.3.1 Models

The dynamic item response models used here are straightforward extensions of the RM and the PCM. The relations and parameters, however, are interpreted

at the level of a single individual. It is assumed throughout that time unfolds in equidistant discrete steps, item parameters are constant over time, and θ is unidimensional. These assumptions are not strictly necessary, but simplify the presentation and illustration of the models and methods considerably. The dichotomous or polytomous n-variate time series y_t is observed from $t = 1, 2, \ldots, T$. We do not distinguish between a time series and a realization thereof. The conditional probability of response $y_{it} = 1$ of a person with a certain θ_t to dichotomous item i at time t is defined by

$$p(y_{it} = 1|y^*_{t-1}, \theta_t, \beta_i) = p_i(\theta_t) = \frac{\exp(\theta_t - \beta_i)}{1 + \exp(\theta_t - \beta_i)}, \tag{5.7}$$

where y^*_{t-1} denotes the complete history of responses of a person, that is, $(y_{t-1}, \ldots, y_1)$, and θ_t is a latent process. The above dynamic Rasch model (DRM) can be extended in the same manner as the RM to allow for polytomous time series. This dynamic partial credit model (DPCM) is then determined by the conditional probability that a person with a certain θ_t at time t responds to item i with threshold parameters $\beta_i = (\beta_{i1}, \ldots, \beta_{iq})$ with response k as follows

$$p(y_{it} = k|y^*_{t-1}, \theta_t, \beta_i) = p_{ik}(\theta_t) = \frac{\exp(k\theta_t - \sum_{v=0}^{k} \beta_{iv})}{1 + \sum_{c=1}^{q} \exp(c\theta_t - \sum_{v=1}^{c} \beta_{iv})}, \tag{5.8}$$

where $\beta_{i0} \equiv 0$. The specification of the latent person process and further assumptions are discussed in the next section.

5.3.2 State space representation

In general, a state space model concerns the relations between an observed time series and an unobserved time series (for an overview, see Sage & Melsa, 1979, or Durbin & Koopman, 2001). The relation between the observed time series y_t and a series of unobserved states α_t is specified in an observation model. In addition, a transition model describes the evolution of the unobserved states over time. The observation and transition model together form what is referred to as the state space model. The state space modelling framework is comparable to the framework of structural equation modelling (SEM) often used in behavioral research (McCallen & Ashby, 1984). Whereas the state space framework generally pertains to within system variation, for example, the tracking of the position, direction, and speed of an aeroplane, the SEM framework usually concerns between systems variation, that is, individual differences on some psychological variables of interest. The simplest of state space models is that in which all variables are

normally distributed and all relationships are linear. The generality of state space modelling lies in the fact that the same estimation methods can be applied to a very wide range of time series models. For the normal linear case, dynamic regression models, structural time series models, and auto-regressive moving average (ARMA) models can all be analyzed within this framework after being represented in state space form. This is analogous to the procedure in SEM in which a variety of regression and factor models can be estimated after being put in SEM form.

The specific state space model considered here concerns non-normally distributed variables and non-linear relationships. Now, the observation models for the dynamic Rasch and partial credit model are already given in Equations 5.7 and 5.8, respectively. The approach to the modelling of dichotomous and polytomous time series presented here resembles the discussion of dynamic generalized linear models for categorical time series in Fahrmeir and Tutz (2001, Chapter 8). The relations between the observed and latent time series are specified by the construction of a linear predictor η_t. To this end, an $(n \times q) \times m$ design matrix Z_t with known elements and the m-dimensional series of unobserved states α_t are related to the mean of the observed time series by

$$\mu_t = h(\eta_t) = h(Z_t \alpha_t), \tag{5.9}$$

where the function $h(.)$ is referred to as the response function. The covariance matrix of the observed time series is denoted by Σ_t. The design matrix Z_t can consist of fixed values, covariates and past values of y_t. For the evolution of α_t, the following linear transition equation is used

$$\alpha_t = F_t \alpha_{t-1} + R_t \xi_t, \qquad \xi_t \sim \mathrm{N}(0, Q_t), \tag{5.10}$$

where F_t is an $m \times m$ transition matrix, R_t is an $m \times p$ selection matrix, and ξ_t is a p-dimensional white noise sequence with associated covariance matrix Q_t. The state vector α_t can consist of time-varying and time-constant elements, which are selected by R_t. The initial state α_0 is normally distributed with mean a_0 and covariance matrix Q_0. In our present situation, the state vector can consist of multiple person processes, person means, and item parameters. Examples of specifying IRT models in state space form are discussed in the next section.

In addition to the observation and transition equation, the following assumptions are made. Let y_{t-1}^* and α_t^* denote the complete histories of the observed and unobserved time series, that is, $y_{t-1}^* = (y_{t-1}, \ldots, y_1)$ and $\alpha_t^* = (\alpha_t, \ldots, \alpha_0)$.

Then, it is assumed that y_t is independent of α^*_{t-1}, that is,

$$p(y_t|y^*_{t-1}, \alpha^*_t) = p(y_t|y^*_{t-1}, \alpha_t).$$

Additionally, it is assumed that the state process α_t is first-order Markovian, i.e.,

$$p(\alpha_t|\alpha^*_{t-1}, y^*_{t-1}) = p(\alpha_t|\alpha_{t-1}).$$

The final assumption resembles the assumption of local independence in IRT and is given by

$$p(y_t|y^*_{t-1}, \alpha_t) = \prod_{i=1}^{n} p(y_{it}|y^*_{t-1}, \alpha_t).$$

5.3.3 Examples of model specification

In order to illustrate the generality of the state space modelling framework, two examples of how the discussed IRT models can be specified in state space form are given. The first example is the specification of the RM for which we use a data set to estimate its parameters with state space methods and compare with the estimation methods commonly used in IRT in the next section. We consider a Rasch model for five items and for persons that are tested on only one occasion. Although it is an atypical application of the state space framework, this model can be specified in state space form as follows. The observation equation has already been given by Equation 5.1. The transition equation consists of an independent normally distributed process. However, it does not describe the variation over time, but the variation between persons. In the specification, the index t can now be replaced by j to emphasize that persons are considered instead of time points. It is stressed that by specifying the transition equation in this manner, a distribution for θ is assumed which is generally not necessary in a Rasch model. However, if MML estimation is used, a (standard) normal distribution is usually assumed in the estimation of item parameters. The design matrix specifies the relation between the person parameters θ and the item parameters β, which are stacked in the state vector α_j. For the model under consideration, this results in

$$Z = \begin{bmatrix} 1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & -1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & -1 \end{bmatrix} \quad \text{and} \quad \alpha_j = \begin{bmatrix} \theta_j \\ \mu_\theta \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}.$$

For identification purposes, the mean μ_θ is fixed at zero, and all corresponding elements of the model vectors and matrices can be deleted. The response function $h(.)$ is the logistic, that is, $\frac{\exp(.)}{1+\exp(.)}$. It can be easily verified that by inserting the above specification into Equation 5.9, the relation of Equation 5.1 is obtained. Note that the vector containing the n item probabilities $p(\theta_j)$ is equal to $h(Z\alpha_j)$. We can then denote the covariance matrix by making use of the assumption of local independence as follows

$$\Sigma_j = \begin{bmatrix} p_1(\theta_j)q_1(\theta_j) & & & & \\ 0 & p_2(\theta_j)q_2(\theta_j) & & & \\ 0 & 0 & p_3(\theta_j)q_3(\theta_j) & & \\ 0 & 0 & 0 & p_4(\theta_j)q_4(\theta_j) & \\ 0 & 0 & 0 & 0 & p_5(\theta_j)q_5(\theta_j) \end{bmatrix},$$

where $q_i(\theta_j) = 1 - p_i(\theta_j)$, $i = 1, \ldots, 5$. Keeping in mind that μ_θ is fixed, the transition, selection and error covariance matrices for this model are specified by

$$F = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad R = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \text{and} \quad Q = \sigma_\theta^2.$$

The initial state and covariance a_0 and Q_0 can be disregarded in this situation, because it can be assumed that the observations are independent, and therefore the Markov assumption can be dropped. The model can be extended in a straightforward manner for instance to analyze multiple groups and inspect group differences and uniform differential item functioning.

As a second example, consider the time series obtained from the scores of three persons measured from $t = 1, \ldots, T$ on two items with each three categories following a partial credit model. The observed time series vector y_t consists of the stacked dummy coded response vectors of each person and is of length $N \times n \times q = 3 \times 2 \times 2 = 12$. Assume that the first person follows a zero mean independent normally distributed process, the second latent process is a zero mean first order autoregression, and the third latent process obeys a zero mean first order random walk. Then, the design matrix and the state vector are

specified as follows

$$Z = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 2 & 0 & 0 & -1 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 & 0 \\ 2 & 0 & 0 & 0 & 0 & -1 & -1 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 2 & 0 & -1 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 2 & 0 & 0 & 0 & -1 & -1 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 2 & -1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 2 & 0 & 0 & -1 & -1 \end{bmatrix} \quad \text{and} \quad \alpha_t = \begin{bmatrix} \theta_{1t} \\ \theta_{2t} \\ \theta_{3t} \\ \beta_{11} \\ \beta_{12} \\ \beta_{21} \\ \beta_{22} \end{bmatrix}.$$

The covariance matrix for this model is a block matrix given by

$$\Sigma_t = \begin{bmatrix} \Sigma_{1t} & & \\ 0 & \Sigma_{2t} & \\ 0 & 0 & \Sigma_{3t} \end{bmatrix},$$

where each block $j = 1, \ldots, 3$ is given by

$$\Sigma_{jt} = \begin{bmatrix} \text{diag}(p_1(\theta_{jt})) - p_1(\theta_{jt})p_1(\theta_{jt})' & \\ 0 & \text{diag}(p_2(\theta_{jt})) - p_2(\theta_{jt})p_2(\theta_{jt})' \end{bmatrix}$$

where $\text{diag}(p_i(\theta_{jt}))$, $i = 1, 2$ is a 2×2 diagonal matrix with the elements of the vector containing the item category probabilities $p_i(\theta_{jt})$ on the diagonal. The transition, selection, and state error covariance matrix can be formed by

$$F = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \phi_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad R = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \text{and} \quad Q = \begin{bmatrix} \sigma^2_{\theta_1} & & \\ 0 & \sigma^2_{\theta_2} & \\ 0 & 0 & \sigma^2_{\theta_3} \end{bmatrix}.$$

The process of the first person is stationary. In order for the process of person two to be stationary, we constrain $|\phi_1|$ to be < 1. The process of the third person is not stationary. It should be noted that the Kalman filtering and smoothing procedure discussed next does not require the process under scrutiny to be stationary.

5.3.4 Estimation

The estimation of the discussed IRT models represented in state space form is performed by an iterative Kalman filtering and smoothing procedure as described in Fahrmeir and Wagenpfeil (1997). In this procedure, the mode of the posterior distribution of the states α_t is found by numerical approximations. The log-posterior for the state space model described above in the polytomous case is given by

$$\begin{aligned}\log L(y;\theta,\beta) = &\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{i=1}^{n}\sum_{k=1}^{q} y_{ijkt}\log(p_{ik}(\theta_{jt}))\\ &+(1-\sum_{k=1}^{q} y_{ijkt})\log(1-\sum_{k=1}^{q} p_{ik}(\theta_{jt})\\ &-\frac{1}{2}(\alpha_0-a_0)'RQ_0^{-1}R'(\alpha_0-a_0)\\ &-\frac{1}{2}\sum_{t=1}^{T}(\alpha_t-F\alpha_{t-1})'RQ^{-1}R'(\alpha_t-F\alpha_{t-1}),\end{aligned} \tag{5.11}$$

where all individual latent processes θ_{jt}, $j = 1,\ldots,N$, individual means, and item parameters are stacked in the state vector α_t. The procedure consists of several steps which are now described in detail.

In discussing the steps of the Kalman filter and smoother (KFS), it is assumed that the elements of Z, F, a_0, Q_0, R, and Q are either fixed or known. Estimates of α_t and associated covariance matrices are denoted by $a_{t|t}$ and $V_{t|t}$ for the filter, and $a_{t|T}$ and $V_{t|T}$ for the smoother. Each iteration i of the KFS needs evaluation values for the complete latent state process which are denoted by $\tilde{a}^i = (\tilde{a}_1^{i\prime}, \tilde{a}_2^{i\prime}, \ldots, \tilde{a}_T^{i\prime})'$. The filtering recursions consist of a prediction and correction step defined for $t = 1,\ldots,T$ by

1. Prediction:

$$\begin{aligned}a_{t|t-1} &= Fa_{t-1|t-1}, \quad a_{0|0} = a_0,\\ V_{t|t-1} &= FV_{t-1|t-1}F' + RQR', \quad V_{0|0} = RQ_0R'.\end{aligned} \tag{5.12}$$

2. Correction:

$$\begin{aligned}V_{t|t} &= (V_{t|t-1}^{-1} + B_t)^{-1},\\ a_{t|t} &= a_{t|t-1} + V_{t|t}b_t,\end{aligned} \tag{5.13}$$

where B_t and b_t are given by

$$B_t = Z'\Sigma_t Z,$$
$$b_t = Z'(y_t - h(Z\tilde{a}_t^i)) - B_t(a_{t|t-1} - \tilde{a}_t^i),$$

and Σ_t is evaluated at $\tilde{a}_t^i$.

The filter predictions $a_{t|t-1}$ are a natural choice to start up the iterations, i.e., $\tilde{a}_t^1 = a_{t|t-1}$. If the procedure is terminated after a single iteration, it is equal to the generalized extended Kalman filter described in Fahrmeir (1992).

The fixed interval smoother is initialized with the final estimates of the Kalman filter $a_{t|t}$ and $V_{t|t}$. For $t = T, \ldots, 2$, the smoother can be given by

$$\begin{aligned} a_{t-1|T} &= a_{t-1|t-1} + G_t(a_{t|T} - a_{t|t-1}), \\ V_{t-1|T} &= V_{t-1|t-1} + G_t(V_{t|T} - V_{t|t-1})G_t', \end{aligned} \tag{5.14}$$

where

$$G_t = V_{t-1|t-1}F'V_{t|t-1}^{-1}. \tag{5.15}$$

After the smoother is applied the evaluation values are updated with the smoother estimates, that is, $\tilde{a}^i = (a_{1|T}^{'}, \ldots, a_{t|T}^{'})'$. The procedure is repeated until some convergence criterion is reached. The stopping criterion used in the present study is $\max|\tilde{a}^i - \tilde{a}^{i-1}| < 1^{-12}$. The KFS procedure discussed in the above resembles the procedure described in Durbin and Koopman (2001, Chapter 10).

When the KFS procedure is applied in a standard IRT setting, some comparisons can be made with the estimation procedures described in the previous section. The KFS procedure resembles MML in that a distribution is assumed for θ, and so estimates can be obtained for extreme score patterns. However, the distributional assumption is part of the model and not of the estimation procedure (see also, Holland, 1990). It differs from MML in that both item and person parameters are estimated simultaneously. This resembles JML estimation procedure, yet, that procedure iteratively estimates person and item parameters. The procedure shows similarities with that described in Tsutakawa and Johnson (1989) in which uncertainties in item parameter estimates are incorporated in the person parameter estimation procedure.

5.3.5 EVALUATION

The same aspects of evaluating the fit of standard IRT models are involved in the case of dynamic IRT models. An important additional aspect is the dependence

of the observations over time. In particular, the time dependence is explicitly accounted for by the specification of the model, and the extent to which this is successfully performed can be checked. A customary method is to inspect the autocorrelations or spectra of the residuals. Any substantial residual dependencies can then be interpreted as a misspecification of the transition equation. This inpection can be performed at the level of the individual. Together with the fit indices discussed earlier which are now to be interpreted conditionally on y^*_{t-1}, an indication of the appropriateness of the model can be obtained at different levels.

The standardized residuals z_{ij} as defined in Equation 5.6 can now be extended with the time index t. The lag l $n \times n$ auto- and cross-correlation matrix of the standardized residuals of person j is denoted by C_{jl} and can be computed[1] from the standardized residual z_{ijt} by

$$C_{jl} = \sum_{t=l+1}^{T} \frac{z_{jt} z'_{j,t-l}}{T}.$$

The lag zero correlation matrix can be used for inspection of any residual dependencies not accounted for by the model. If the model provides an accurate description, off-diagonal elements should be close to zero. Lagged correlation matrices can be used to inspect any residual time dependencies.

5.4 Examples

In order to provide an illustration of how the discussed models can be applied by making use of the state space framework, we analyse two data sets: one obtained in a standard IRT setting and the other in a longitudinal setting. The first analysis is performed in order to compare the KFS estimation procedure with standard IRT estimation methods for the estimation of both person and item parameters. The second analysis illustrates the generality of the state space framework in a longitudinal setting where the number of persons is relatively small and the number of time points is relatively large.

5.4.1 Standard IRT: LSAT-6 data

The first data set has been used frequently to illustrate and compare parameter estimation methods for the RM (see e.g., Baker, 1991; Thissen, 1982; Andersen

[1]Note that the denominator that we use here is T, as opposed to $T - l$, which was used in Chapter 2.

& Madsen, 1977). It consists of the responses of 1000 persons to five figure classification items which formed Section 6 of the Law School Admission Test (LSAT) as described in Bock and Lieberman (1970). With these data, the item parameter estimation methods CML and MML are compared with the KFS. In addition, we compared BM estimation of θ with the output of the KFS. It is stressed that with the KFS, both person and item parameter estimates are obtained at the same time. The RM is used in the state space representation that was discussed in the previous section.

All analyses are performed with the free software package R (R Development Core Team, 2008). For CML and MML estimation of the item parameters, the R packages extended Rasch modelling (eRm; Mair & Hatzinger, 2006, 2007) and latent trait modelling (ltm; Rizopoulos, 2006) are used, respectively. The ltm package is also used to produce BM estimates of the person parameter. The KFS estimation procedure was implemented in R by the present authors.[2] In applying the MML and KFS procedures, a standard normal distribution is assumed for θ. In comparing the methods for item parameter estimation, the mean item difficulty is fixed to zero.

Table 5.1 displays the five estimated item difficulties and standard errors (SEs) of the CML, MML, and KFS estimation procedures. The CML and MML point estimates are very close to each other compared to those obtained with the KFS. The differences between the point estimates of the KFS and those of CML and MML are not large, yet not negligible, larger at the extremes, and can be interpreted as bias. The differences are most likely due to the fact that the KFS procedure utilizes a posterior and CML and MML a likelihood. The standard errors of all three methods are very close, although those of the KFS are slightly smaller.

In Table 5.2, the sumscores and associated person parameter estimates and SEs are shown for the BM and KFS estimation procedures. BM estimation of θ is performed with fixed item parameters estimated with MML, whereas the KFS estimates of θ are obtained simultaneously with the item parameter estimates. The point estimates of the two methods have to be compared in view of the differences between MML and KFS for the estimation of item parameters and especially μ_θ. Keeping these differences in mind, the two methods can be considered to yield comparable point estimates. The standard errors of the KFS are substantially smaller than the SEs obtained with BM estimation. This might be due to the

[2]The source code is available upon request.

Table 5.1: Item parameter estimates of LSAT data set

	CML		MML		KFS	
Parameter	Est.[1]	SE[2]	Est.	SE	Est.	SE
β_1	-1.256	(0.104)	-1.255	(0.104)	-1.223	(0.102)
β_2	0.475	(0.070)	0.476	(0.070)	0.465	(0.069)
β_3	1.236	(0.069)	1.235	(0.069)	1.194	(0.065)
β_4	0.168	(0.073)	0.168	(0.073)	0.168	(0.071)
β_5	-0.623	(0.086)	-0.625	(0.086)	-0.604	(0.084)
μ_θ	-	(-)	1.475	(0.052)	1.417	(0.051)
σ_θ	-	(-)	0.755	(0.069)	-	(-)

[1] Estimate

[2] Standard error

fact that the KFS simultaneously estimates item and person parameters.

Goodness of fit of the Rasch model was evaluated with the Pearson X^2 and the likelihood ratio statistic G^2 as discussed earlier. For the LSAT data, the results of the MML estimation procedure were used for computing the fit statistics which resulted in $X^2 = 18.33$ with df $= 25$ ($p = 0.83$), and $G^2 = 21.80$ also with df $= 25$ ($p = 0.65$), indicating a good fit. Since the fit is satisfactory, checking further item and person fit diagnostics is not deemed necessary for now.

In order to compare the results of the KFS estimation procedure in this situation with the CML and MML estimation procedures, a small study is conducted in which data are simulated on the basis of the LSAT analysis. That is, 1000 replications of 1000 responses to a test with 5 items are simulated. The param-

Table 5.2: Person parameter estimates of LSAT data set

	BM		KFS	
Sumscore	Est.	SE	Est.	SE
0	-0.432	(0.790)	-0.526	(0.704)
1	0.038	(0.793)	-0.038	(0.696)
2	0.516	(0.801)	0.448	(0.701)
3	1.007	(0.816)	0.950	(0.720)
4	1.519	(0.836)	1.490	(0.754)
5	2.058	(0.862)	2.095	(0.804)

TABLE 5.3: RESULTS OF SIMULATION WITH RESPECT TO ESTIMATION OF ITEM PARAMETERS

		CML			MML			KFS		
Parameter	Value	Mean[1]	SE[2]	SD[3]	Mean	SE	SD	Mean	SE	SD
β_1	-1.25	-1.263	0.104	0.103	-1.262	0.106	0.106	-1.231	0.104	0.100
β_2	-0.50	-0.502	0.084	0.082	-0.504	0.084	0.092	-0.484	0.083	0.079
β_3	0.00	0.001	0.073	0.072	0.001	0.075	0.077	0.005	0.074	0.070
β_4	0.50	0.506	0.072	0.072	0.507	0.070	0.073	0.494	0.069	0.069
β_5	1.25	1.257	0.069	0.069	1.258	0.069	0.072	1.216	0.065	0.064
μ_θ	1.50	-	-	-	1.507	0.053	0.060	1.446	0.051	0.044
σ_θ	0.75	-	-	-	0.754	0.071	0.103	-	-	-

[1] Mean estimate

[2] Mean standard error

[3] Standard deviation of estimates

eters in this simulation are rounded off for ease of comparison. The results of this study with respect to the item parameters are displayed in Table 5.3. It can be seen that the item parameters produced by KFS are biased. The pattern of the bias is the same as in Table 5.1. The differences in SEs between the three methods are very small and can be neglected. The mean and standard deviations of the fit measures were 27.32 and 11.26 for the G^2, and 25.88 and 10.76 for the X^2, respectively, both with 25 degrees of freedom.

Table 5.4 shows the results of the simulations with respect to the person parameter estimates. Again, the results should be compared in view of the differences in the estimates of item parameters and the mean μ_θ. In this respect, the differences are not too large, except perhaps for sumscore 5.

TABLE 5.4: RESULTS OF SIMULATION WITH RESPECT TO ESTIMATION OF PERSON PARAMETERS

	BM			KFS		
Sumscore	Mean	SE	SD	Mean	SE	SD
0	-0.402	0.791	0.128	-0.509	0.703	0.050
1	0.061	0.794	0.102	-0.027	0.692	0.051
2	0.542	0.803	0.090	0.461	0.701	0.047
3	1.027	0.818	0.067	0.962	0.721	0.049
4	1.544	0.838	0.067	1.507	0.756	0.046
5	2.082	0.864	0.067	2.112	0.807	0.046

5.4.2 Dynamic IRT: Borkenau data

The second example consists of an application of a dynamic PCM as discussed in Section 5.3 to repeated administrations of a personality questionnaire. We use a selection of the data set that has already been used to illustrate dynamic models and associated estimation methods (Molenaar, 2004; Hamaker, Dolan, & Molenaar, 2005). The data have been collected by Borkenau and Ostendorf (1998) and consist of the responses of 22 persons to a 30 item personality questionnaire on 90 consecutive days. The questionnaire was designed to measure the Big Five personality factors and the items were scored on a seven-point Likert scale. Since our interest lies in the illustration of a dynamic PCM, we only used the responses to the six items that are indicative of the factor Extraversion. This scale consisted of three positively formulated and three negatively formulated items.

Since we have no knowledge about the individual latent processes for this type of analysis, modelling starts out by assuming that each person follows a unidimensional independent normally distributed process with possibly different means. In other words, a measurement invariant model is assumed over persons to start the analysis. Each individuals variance is fixed at one for now, and to reiterate, the item category parameters are assumed to be constant over time.

The specification of the dynamic PCM in state space form for this situation proceeds along similar lines as described in Section 5.3. A full description of the state space representation is given. If we define

$$Z_1 = \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ \vdots & \vdots \\ 6 & 6 \end{bmatrix}, \qquad \text{and} \qquad Z_2 = \begin{bmatrix} -1 & 0 & \cdots & 0 \\ -1 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & -1 \end{bmatrix},$$

we can write the 792×80 design matrix Z for this example by

$$Z = \begin{bmatrix} I_{22} \otimes (1_6 \otimes Z_1) & 1_{22} \otimes (I_6 \otimes Z_2) \end{bmatrix},$$

where $I_{\cdot}$ is an identity matrix of indicated dimension, $1_{\cdot}$ indicates an identity vector of indicated dimension, and $\otimes$ is the kronecker product. The state vector

containing the person and item parameters is given by

$$\alpha_t = \begin{bmatrix} \theta_{1t} \\ \mu_{\theta_1} \\ \vdots \\ \theta_{22,t} \\ \mu_{\theta_{22}} \\ \beta_{11} \\ \beta_{12} \\ \vdots \\ \beta_{66} \end{bmatrix}$$

The covariance matrix Σ_t can be build in an analogous manner as described in Section 5.3.3. If we continue by defining

$$F_1 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},$$

then the 80×80 transition matrix can be written as

$$F = \begin{bmatrix} I_{22} \otimes F_1 & 0 \\ 0 & I_{36} \end{bmatrix}.$$

Finally, let us define

$$R_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

then the 80×22 selection matrix can be written as

$$R = \begin{bmatrix} I_{22} \otimes R_1 \\ 0 \end{bmatrix}.$$

The model in the above representation is not identified and at least one of the person means or item parameters needs to be fixed. We choose to fix the first person mean at zero for the present situation. The dimensions of all associated model vectors and matrices are then reduced by one. The resulting dimensions of the model vectors and matrices are displayed in Table 5.5.

Table 5.6 displays the results of the KFS estimation procedure with respect to the item category parameters. The values of the category thresholds are fairly spread out over the scale. Their overall mean is rescaled at zero, so that the individual latent processes can be related to the response scale. Except for the

Table 5.5: Model dimensions for state space representation of dynamic PCM for Extraversion data

Model vector	Dim.	Model matrix	Dim.
y_t	792	Z	792×79
α_t	79	F	79×79
ξ_t	22	R	79×22

first two category thresholds of item four, all item category parameters are ordered within items. It is observed that the parameters of the outer categories have the largest standard errors, especially the lowest categories.

Table 5.7 displays the estimated person means, standard errors, and person fit output of the KFS procedure. It is clear that the differences in individual means are quite large, ranging from -0.47 for person 3 up till 2.72 for person 6. The overall mean is equal to 0.68, indicating that the persons responded on the positive side of the scale of the extraversion items. No overall goodness of fit measures are calculated here, because χ^2 approximations based on the contingency table of possible response patterns are likely to fail. That is, the total number of observations ($22 \times 90 = 1980$) compared to the number of cells of this table is very small ($7^6 = 117649$). The individuals that show the largest sum of squared standardized residuals are persons 5 and 6.

Table 5.6: Item category parameter estimates

i	k	β_{ik}	SE	i	k	β_{ik}	SE	i	k	β_{ik}	SE
1	1	-2.52	0.21	3	1	-2.13	0.27	5	1	-2.87	0.24
	2	-1.01	0.11		2	-1.78	0.16		2	-1.04	0.10
	3	-0.55	0.08		3	-0.88	0.10		3	-0.40	0.08
	4	0.65	0.06		4	-0.21	0.07		4	0.50	0.07
	5	1.97	0.08		5	1.06	0.07		5	1.67	0.08
	6	3.50	0.14		6	1.86	0.08		6	2.72	0.11
2	1	-1.87	0.22	4	1	-1.82	0.23	6	1	-2.11	0.22
	2	-1.71	0.13		2	-1.91	0.15		2	-1.56	0.13
	3	-0.41	0.08		3	-0.43	0.09		3	-0.53	0.08
	4	0.09	0.07		4	-0.08	0.07		4	0.43	0.07
	5	1.37	0.07		5	1.21	0.07		5	1.73	0.08
	6	2.33	0.09		6	2.31	0.09		6	2.41	0.10

TABLE 5.7: ESTIMATED PERSON MEANS, STANDARD ERROR, AND FIT

Person	$\bar{\theta}$	SE	Fit	Person	$\bar{\theta}$	SE	Fit
1	0.26	0.11	0.42	12	0.68	0.11	0.39
2	0.09	0.11	0.37	13	1.25	0.12	0.96
3	-0.47	0.11	1.00	14	0.80	0.12	0.90
4	1.43	0.12	1.01	15	0.09	0.11	0.58
5	1.50	0.12	1.47	16	0.22	0.11	0.42
6	2.72	0.12	1.38	17	0.47	0.11	0.64
7	0.57	0.11	0.74	18	0.66	0.11	0.76
8	1.28	0.11	1.01	19	0.51	0.11	0.96
9	0.92	0.11	0.69	20	0.85	0.11	0.68
10	-0.07	0.11	0.55	21	-0.38	0.11	0.71
11	1.49	0.12	0.82	22	0.09	0.11	0.74

As a final illustration, first order autoregressive latent processes were fitted to the time series of each person seperately. The penalized likelihood of Equation 5.11 was optimized numerically with the L-BFGS-B method of the R function optim() (see Byrd, Lu, Nocedal, & Zhu, 1995). The item category thresholds were fixed at the values displayed in Table 5.6, the process was restricted to be stationary by restricting $|\phi_1| < 1$, and again each individual's variance was fixed at one. It is investigated if the individual fit improved. The results of this analysis are shown in Table 5.8. In inspecting point estimates and standard errors, it can be said that in about one third of the cases a substantial autoregressive component is found. Remarkably, the change in fit index is negligible in most cases. However, it can be seen that for person 6, a relatively strong autoregressive component is found whereas the fit worsened.

5.5 DISCUSSION

The main purpose of this article was to apply state space methods to the modelling of item responses. Not only can these methods be applied in standard IRT settings, extensions to modelling repeated measurements are easily made within the same framework, and the same estimation methods can be used. An application of Kalman filtering and smoothing techniques to two example data sets illustrated the flexibility of the state space framework. The results of the first analysis indicated that the KFS can result in some bias in the estimation of

Table 5.8: Estimated autoregressive parameter, standard error, and fit

Person	ϕ_1	SE	Fit	Person	ϕ_1	SE	Fit
1	0.17	0.13	0.42	12	0.23	0.11	0.39
2	0.16	0.14	0.37	13	0.26	0.07	0.97
3	0.27	0.20	1.00	14	0.22	0.04	0.92
4	0.30	0.08	1.01	15	0.12	0.12	0.58
5	0.28	0.06	1.48	16	0.17	0.19	0.42
6	0.60	0.06	1.48	17	0.20	0.11	0.65
7	0.16	0.09	0.75	18	0.37	0.21	0.75
8	0.39	0.11	1.00	19	0.21	0.09	0.96
9	0.24	0.09	0.69	20	0.32	0.13	0.68
10	0.15	0.11	0.55	21	0.21	0.21	0.71
11	0.34	0.08	0.82	22	0.15	0.11	0.74

item parameters. This can be related to the fact that a posterior distribution is optimized by the KFS estimation procedure, and this is known to produce bias (see Tsutakawa & Johnson, 1989; Hoijtink & Boomsma, 1995). However, the SEs are consistent with the two standard IRT estimation procedures CML and MML. The discussed framework can be easily equipped to perform typical IRT analyses such as multi-group and DIF analyses.

The second example illustrated some possibilities of the state space approach to the modelling of repeated measurements. The KFS estimation procedure was applied to a data set to obtain item and person parameter estimates. Hereafter, individual time series were analyzed again to investigate the strength of latent autoregressions. It can be stated that this procedure works reasonably well for the discussed situation and might be useful for analyzing various types of longitudinally observed item responses. However, an investigation into the quality of the produced estimates and fit diagnostics, e.g., by means of simulations, remains an important and interesting topic for future research.

Admittedly, many extensions of the discussed models and procedures remain to be explored. For instance, the inclusion of a discrimination parameter as in the two parameter logistic model (Birnbaum, 1968) and generalized partial credit model (Muraki, 1994) is an interesting extension, because in many applications, the Rasch model and its extensions do not fit very well to all item responses. In close connection with this is the extension to differing individual variances of the

latent process in the case of repeated measurements. Fahrmeir and Wagenpfeil (1997) discuss an estimator for this case, but its quality is as yet unknown. Finally, extensions of the model to allow for time varying parameters, e.g., item parameters or person means (trends) might be significant developments for the future, especially when the interest lies in the analysis of change.

In closing, all possible applications of the state space framework are interesting only when the different models and extensions can be compared, and the differences can be tested. That is, reliable diagnostics and fit statistics are necessary to find a fit model, and finding them is perhaps the biggest challenge.

6

Epilogue

This thesis discussed methods for the analysis of psychological measurements in the form of categorical time series. The discourse started out with the analysis of multivariate continuous and categorical time series within the framework of structural equation modelling (SEM). The final part of the thesis concerned the analysis of multi-subject multivariate categorical time series by making reference to the frameworks of state space modelling (SSM) and item response theory (IRT). This final chapter consists of a short review of the general conclusions that can be drawn from the investigations in Chapters 2 to 5, i.e., the core of this thesis. Finally, guidelines for future research in the field of psychological measurement in the form of categorical time series are suggested.

6.1 Conclusions

6.1.1 Structural equation modelling

In Chapter 2, the use of the Toeplitz matrix containing sample auto- and cross-covariances in order to fit multivariate stationary autoregressive (AR) models was investigated. For normally distributed time series, it was found that parameters and standard errors are correctly recovered only for pure vector autoregressions. For multiple indicator autoregressive models, the estimated parameters were correct, but the standard errors were not. This limits the use of the Toeplitz method for this type of measurements. In the second part of the chapter, autoregressive models were fitted to multivariate categorical time series by using the Toeplitz matrix containing polychoric auto- and cross-correlations. The results indicated that the estimates of the parameters are correct for the used models, but the standard errors are not. In view of the simulation results and available asymptotic results on sample auto- and cross-covariances, it can be argued that the Toeplitz method should not be used in investigations with real data in which the type and order of the model is not known. Also, since stationarity is a necessary assumption for the Toeplitz method, but not for filtering methods, the latter are preferable in modelling time series measurements. The Toeplitz method can how-

ever have practical utility when used as a method of moments to obtain parameter estimates.

The above argumentation should not be taken as a case against applying the SEM framework in a time series setting (see also, MacCallum & Ashby, 1986). Rather, it is the methodology of using summary statistics for estimating and fitting models, a regular course of things in standard SEM applications, that is not recommended in the situation of normal and, especially, categorical time series.

6.1.2 State space modelling

Since filtering and smoothing methods to analyse categorical time series are not widely available, the second part of this thesis consisting of Chapters 3 to 5 focused on such methods. Specifically, the SSM framework was referred to for specifying and estimating models for categorical time series. Chapter 3 addressed the case of univariate categorical time series, whereas Chapters 4 and 5 discussed the multivariate case.

In Chapter 3, a Kalman filtering and smoothing method was used to fit a latent autoregressive process to an observed univariate categorical time series through a logistic response function. It was found that autoregressive parameters showed some bias, but could be consistently estimated. The estimates of threshold parameters, associated with particular categories of the time series, were however found to be biased and inconsistent. In general, the presented approach did not seem to work satisfactorily for univariate categorical time series.

The first part of Chapter 4 presented a case in favor of the analysis of intra-individual variation. The second part contained illustrations of a dynamic logistic model for multivariate dichotomous time series with simulated and real data. In this chapter, the presented dynamic logistic model was related to the Rasch model.

Chapter 5 elaborated on Chapter 4, with the focus shifting towards a comparison of the SSM and IRT frameworks. A comparison was made between standard IRT estimation methods and a Kalman filtering and smoothing method for the estimation of item and person parameters in the situation of cross-sectional data. The modelling approach presented in Chapter 4 was extended to polytomous time series. The chapter ended with an application of the methods to multiple subject polytomous time series. The presented approach provided a useful methodology for simultaneously fitting models to polytomous time series of multiple subjects.

The fact that the SSM framework is very useful for modelling time series measurements is not new. That it provides a meaningful framework for psychological time series measurements is therefore to be expected. In the present thesis, we have discussed models and estimation methods for categorical time series within the SSM framework, but also within the frameworks of SEM and IRT. That the SSM framework can be used in a sensible manner in combination with the familiar and comprehensive frameworks of SEM and IRT is one of the main conclusions of the present thesis. Particularly, the frameworks of SSM and IRT form a good combination for the type of categorical time series measurements which were analysed in this thesis. More general, the combination of these two frameworks can provide a unified approach to model cross-sectional, longitudinal, and time series data formed by categorical psychological measurements.

6.2 Guidelines for future research

In closing this thesis, guidelines for future investigations in psychological measurement in the form of categorical time series are suggested, which are distinguished in three interesting topics. First, further comparisons of the Kalman filtering and smoothing estimation method used in this thesis with other filtering methods are important. Many methods for non-normal time series and non-linear time series model are Bayesian (Doucet, de Freitas, & Gordon, 2001), and are known to work well. It would be of interest to compare Bayesian filtering methods with the method used in this thesis. Recently, Chow, Ferrer, and Nesselroade (2007) used an unscented Kalman filter to fit nonlinear dynamic models for dyadic interactions in emotions. Such a filter might also be applied to dynamic logistic models for categorical time series. It is evident that comparative studies are necessary to expose strengths and weaknesses of different methods.

A second important topic is the further development of parameter estimation methods apart from methods for filtering. Particularly, the estimation methods for parameters such as latent autoregressive parameters and variances of innovations are not well developed (Fahrmeir & Wagenpfeil, 1997). The numerical method used in the present thesis can then, for instance, be used for comparison.

Apart from the establishment of the performance of the estimation methods, there is a need to develop and investigate the quality of measures of fit for the particular time series models under scrutiny. This has clearly been a neglected issue in the present thesis, but is largely uncharted and remains one of the most important topics in future investigations. This because a crucial stage in any

statistical analysis is the selection of an appropriate model. Of special interest would be statistical fit procedures to assess if an observed psychological time series is or is not obtained from a stationary stochastic process. If nonstationarity is then found, an immediate need arises for developing statistical techniques for the fit of state space models with time-varying parameters (for the linear Gaussian model, see e.g., Molenaar, 1994). Whereas it is often straightforward to estimate complex statistical models, it is often more difficult to select the best of several straightforward statistical models.

A

A note on classical test theory in heterogeneous populations

A.1 Introduction

A standard definition of the true score in classical test theory is as the mean of the propensity distribution of scores of a fixed person obtained in an infinite series of independent trials with that person (Lord & Novick, 1968). Because it is considered not to be realistic to obtain an infinite series of independent trials with the same person, classical test theory is instead based on the scenario in which an infinite number of persons is measured at a fixed number of independent trials. As is explicitly acknowledged by Lord and Novick (1968, p. 32), this implies that classical test theory is compatible with a situation in which the variance of the propensity distribution of scores for each person in the population differs between persons. We will denote the latter situation by heterogeneity of the population. The only relationship between individual and population variances which exists is that the mean of the variances of individual propensity distributions equals the error variance in the heterogeneous population of persons (Lord & Novick, 1968, p. 35).

The increase in reliability as a consequence of group heterogeneity was mentioned by Gulliksen (1950), Lord and Novick (1968), and studied by, e.g., Zimmerman, Williams, and Burkheimer (1968). The type of heterogeneity studied in this paper differs from group heterogeneity in that it concerns different variances of individual propensity distributions. In a factor analytical context, unobserved heterogeneity produced by varying parameters in a number of subpopulations was studied by means of finite mixture analysis in, e.g., Muthén (1989). More recently, Kelderman and Molenaar (2007) investigate the effects of individual differences in factor loadings. Various types of heterogeneity were investigated using Bayesian factor analysis in Ansari, Jedidi, and Dube (2002). Their definition of heterogeneity in covariance structures partially coincides with our notion of population heterogeneity. In this appendix, however, the situation is studied in the context of classical test theory and different subtypes of heterogeneity, defined by

different measurement forms, i.e., parallel, tau-equivalent, and congeneric, can be explicitly studied.

Since population heterogeneity is explicitly accounted for in classical test theory, an examination in its effects on estimation and its possible detection, is a useful undertaking. The purpose of this study is to investigate whether the application of classical test theory is still justified in the situation of population heterogeneity in the sense that good results in terms of estimation performance are to be expected. It is investigated in a simulation study whether the presence of population heterogeneity in the classical test model has an effect on true score prediction in the situation of parallel, tau-equivalent, and congeneric measurements. Prediction of true scores in classical test theory can be accomplished in various ways, e.g., by means of variants of factor score prediction (Lord & Novick, 1968). The specific aim of this note is to compare the performance of a traditional factor score predictor of true scores with an unconventional predictor of true scores defined by simple pooling of measurements in a heterogeneous population of subjects. Finally, the fit of the factor models defined by different types of measurements is inspected for revealing population heterogeneity.

A.2 POPULATION HETEROGENEITY

There are many ways to construct population heterogeneity in the classical test theory model. In what follows, a description of a possible construction of heterogeneity is given which is in line with the simulation study described in Section A.3. It should be stressed that this construction is a methodological issue in the investigation of population heterogeneity in classical test theory.

Consider a possibly heterogeneous population of subjects in which the propensity density of each individual subject i is Gaussian with mean μ_i and variance σ_i^2: $Y_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$. Let μ_i and σ_i^2 be random variables over subjects. That is, the mean and variance of μ_i are hyperparameters denoted by, respectively: τ_μ and ω_μ^2. The mean and variance of σ_i^2 are hyperparameters denoted by, respectively, τ_{σ^2} and $\omega_{\sigma^2}^2$. Note that if the variance hyperparameter of σ_i^2 is nonzero, $\omega_{\sigma^2}^2 > 0$, then the population is heterogeneous. It is emphasized that the distributions of the hyperparameters do not affect the distribution of measurements, because a drawing from the hyperparameter distributions only determines the values of the parameters of the individual propensity distributions.

Consider now a random sample of N subjects drawn from this population, $i = 1, \ldots, N$, which is measured at T occasions $j = 1, \ldots, T$. If the propensity

density of each subject in the random sample is invariant across measurement occasions, then the T sets of N scores thus obtained constitute parallel measurements according to classical test theory. If the mean of this propensity density is invariant, but not the variance, then the measurements are tau-equivalent. The measurements are congeneric if the mean on one measurement occasion is a linear combination of the mean on another occasion. For all this to hold, it is not required that the propensity density is of the same type for each subject. Hence, our assumption that this propensity density is Gaussian for each subject is not necessary to obtain the different types of measurements, but is only used to ease the presentation.

A.3 Simulation study

Suppose that T parallel measurements $Y_{ij}, j = 1, \ldots, T$, are available for subjects $i = 1, \ldots, N$, and that the correlation between these measurements in the population of subjects is ρ. Suppose also that the variance of scores in the population is σ_Y^2. Then the true score variance ω_μ^2 and the error variance σ_E^2 $(= \tau_{\sigma^2})$ in the population of subjects are, respectively, $\omega_\mu^2 = \rho\sigma_Y^2$ and $\sigma_E^2 = (1-\rho)\sigma_Y^2$.

The T parallel, tau-equivalent, or congeneric measurements define different 1-factor models (see Lord & Novick, 1968, Chapter 24; Jöreskog, 1971; and Jöreskog & Sörbom, 1979). The regression predictor of factor scores as well as the associated prediction variance are given in, e.g., Lawley and Maxwell (1971, p. 109, Eqs. 8.7 and 8.9). For the restricted 1-factor model in the population associated with T parallel measurements the regression predictor of true scores, $RP[\mu_i]$ and its prediction variance var$[RP]$ for the i-th subject are given by

$$RP[\mu_i] = \tau_\mu + \sum_{j=1}^{T} \left(\frac{\omega_\mu^2}{\sigma_E^2 + T\omega_\mu^2} \right) (y_{ij} - \tau_\mu) \quad \text{and} \quad \text{var}[RP] = \frac{\omega_\mu^2 \sigma_E^2}{\sigma_E^2 + T\omega_\mu^2}. \tag{A.1}$$

The alternative unconventional pooling predictor of true score, $PP[\mu_i]$, and its prediction variance var$[PP]$ are defined by pooling across T measurements. For the i-th subject these are given by

$$PP[\mu_i] = \sum_{j=1}^{T} \frac{y_{ij}}{T} \quad \text{and} \quad \text{var}[PP] = \sum_{j=1}^{T} \frac{(y_{ij} - \mu_i)^2}{T^2}. \tag{A.2}$$

Note that var$[RP]$ is invariant across subjects, whereas var$[PP]$ can differ between subjects.

A.3.1 SET UP

To compare both predictors a simulation study was carried out using the following hyperparameter settings. The mean τ_μ and variance ω_μ^2 of the true scores were fixed at $\tau_\mu = 100$ and $\omega_\mu^2 = 10$. The mean τ_{σ^2} of the individual variances was fixed at $\tau_{\sigma^2} = 2.5$. If the variance $\omega_{\sigma^2}^2$ of the individual variances equals zero, then this yields correlations ρ between measurements of $\rho = 0.8$. In the simulation runs for the parallel measurements, the variance hyperparameter $\omega_{\sigma^2}^2$ was increased from $\omega_{\sigma^2}^2 = 0$ to 5 with unit steps. The variance hyperparameter was drawn from a folded normal distribution (see Johnson, Kotz, & Balakrishnan, 1994, p. 170). For the tau-equivalent measurements, within each increase of the variance hyperparameter, the individual error variances on each measurement occassion were drawn from a folded normal distribution with mean equal to the individual error variance and variance 1, 2 or 3. Congeneric measurements were obtained by forming the true scores on one measurement occasion on the basis of a linear combination of the individual true scores on another measurement occassion. The multiplicative parameter in the linear combination was drawn from U(0.95,1.05) or U(0.90,1.10) and the additive parameter was drawn from N(0,1) or N(0,2).

For each simulation, $T = 8$ parallel scores were generated for $N = 10000$ subjects. The restricted 1-factor model associated with parallel, tau-equivalent or congeneric measurements was fitted to the 8×8-dimensional covariance matrix and the 8-dimensional vector of means thus obtained by means of normal theory maximum likelihood estimation (Lawley & Maxwell, 1971). Notice that applicability of normal theory maximum likelihood estimation does not depend upon the distribution of the hyperparameters.

To assess the performance of the both predictors, the mean relative bias and mean absolute bias of true score predictors and associated variance are investigated, defined by, respectively, the mean difference between estimated and true true score and associated variance and mean absolute difference between estimated and true true score and associated variance.

A.3.2 RESULTS

Since there were no effects of heterogeneity on relative bias of both true score predictors and associated variances, these results are not displayed in tables. In short, predictions of true score obtained with parallel measurements showed relatively little bias and prediction with congeneric measurements showed more bias than with tau-equivalent measurements. In general, the relative bias of both true

Table A.1: Absolute biases of regression and pooling predictor for parallel measurements

	RP		PP		
$\omega^2_{\sigma^2}$	Abs-T	Abs-V	Abs-T	Abs-V	P-value
0	0.44	0.12	0.45	0.04	0.31
1	0.43	0.16	0.44	0.04	0.18
2	0.43	0.18	0.43	0.04	0.15
3	0.43	0.20	0.43	0.04	0.57
4	0.43	0.22	0.44	0.04	0.32
5	0.45	0.23	0.45	0.05	0.13

score predictors were comparable. The prediction variance of the pooling predictor, however, showed somewhat more bias than that of the regression predictor.

Table A.1 shows the absolute bias of the regression and pooling predictor (Abs-T) and its associated prediction variances (Abs-V) for parallel measurements. The following observations can be made from Table A.1. Firstly, the fit of the restricted 1-factor model associated with parallel measurements is not affected by population heterogeneity. The p-values of the likelihood-ratio test are excellent in view of the large power ($N = 10000$). Hence, population heterogeneity cannot be detected by inspecting the fit of the restricted 1-factor model. Secondly, the performance of the regression predictor is substantially affected by population heterogeneity in that its estimated prediction variance shows more absolute bias with increasing $\omega^2_{\sigma^2}$. In contrast, the variance of the pooling predictor is much less affected by population heterogeneity. It should be noted that the effective ρ decreases with increasing $\omega^2_{\sigma^2}$, although the decrease was small ($\rho \approx 0.78$ for extreme heterogeneity).

Table A.2 displays the absolute bias of both predictors and associated prediction variances for tau-equivalent measurements. The absolute bias of the variance of the regression predictor increases with increasing heterogeneity whereas that of the pooling predictor remains relatively consant. The 1-factor model associated with tau-equivalent measurements fits in more than 50% of the studied cases, although no clear indication can be given about the situation in which the model does not fit. The results obtained with congeneric measurements are comparable to those obtained with parallel measurements.

In Table A.3, the absolute bias of both predictors and associated prediction variances are given for congeneric measurements with a multiplicative parame-

TABLE A.2: ABSOLUTE BIASES OF REGRESSION AND POOLING PREDICTOR FOR TAU-EQUIVALENT MEASUREMENTS

		RP		PP		
$\omega^2_{\sigma^2}$	τ	Abs-T	Abs-V	Abs-T	Abs-V	P-value
0	1	0.44	0.13	0.45	0.04	0.94
0	2	0.44	0.14	0.45	0.04	0.46
0	3	0.45	0.15	0.45	0.04	0.07[1]
1	1	0.43	0.16	0.43	0.04	0.04[1]
1	2	0.44	0.17	0.45	0.04	0.73
1	3	0.45	0.17	0.46	0.04	0.01[1]
2	1	0.44	0.19	0.45	0.04	0.17
2	2	0.45	0.19	0.45	0.04	0.01[1]
2	3	0.45	0.19	0.46	0.04	0.68
3	1	0.44	0.20	0.45	0.04	0.00[1]
3	2	0.45	0.20	0.45	0.04	0.06[1]
3	3	0.46	0.20	0.46	0.05	0.24
4	1	0.44	0.22	0.45	0.04	0.44
4	2	0.46	0.22	0.47	0.05	0.82
4	3	0.46	0.22	0.47	0.05	0.80
5	1	0.45	0.23	0.46	0.05	0.01[1]
5	2	0.46	0.23	0.47	0.05	0.06[1]
5	3	0.47	0.22	0.48	0.05	0.48

[1] Model does not fit ($\alpha = 0.10$)

ter drawn from U(0.90,1.10) and an additive parameter drawn from N(0,2). It is clearly seen that the absolute biases of the predictors are much larger than in the situation of parallel and tau-equivalent measurements. Also, the 1-factor model for congeneric measurements does not fit for all the situations studied. The absolute bias for the pooling predictor of true score is smaller than that of the regression predictor, although the difference is sometimes small. For the prediction variance, however, the pooling predictor shows substantially less absolute bias than the regression predictor. A clear effect of heterogeneity on the two predictors could not be found in contrast to the situation of parallel and congeneric measurements.

Table A.3: Absolute biases of regression and pooling predictor for congeneric measurements

		RP		PP		
$\omega^2_{\sigma^2}$	τ	Abs-T	Abs-V	Abs-T	Abs-V	P-value
0	1	0.75	0.68	0.46	0.04	0.00[1]
0	2	1.46	1.04	1.40	0.28	0.06[1]
0	3	1.31	1.70	1.02	0.17	0.06[1]
1	1	0.79	0.79	0.47	0.05	0.03[1]
1	2	2.98	2.05	2.99	1.15	0.00[1]
1	3	1.00	1.46	0.54	0.06	0.01[1]
2	1	2.26	1.20	2.25	0.67	0.00[1]
2	2	0.58	0.42	0.50	0.05	0.00[1]
2	3	0.84	0.82	0.54	0.06	0.00[1]
3	1	1.66	1.38	1.59	0.36	0.00[1]
3	2	2.16	1.85	2.13	0.61	0.00[1]
3	3	1.15	1.80	0.73	0.10	0.00[1]
4	1	1.36	1.64	1.19	0.21	0.00[1]
4	2	1.31	1.15	1.18	0.21	0.00[1]
4	3	1.66	1.60	1.59	0.36	0.08[1]
5	1	1.77	1.98	1.67	0.40	0.00[1]
5	2	1.04	1.26	0.71	0.09	0.00[1]
5	3	1.78	1.06	1.76	0.44	0.00[1]

[1] Model does not fit ($\alpha = 0.10$)

A.4 Discussion

First, it is noteworthy that even under extreme heterogeneity, the 1-factor model shows an excellent fit for the parallel measurements. This can be seen as an indication that classical test theory is in this case compatible with the situation that the variance of the propensity distribution of scores for each person in the population differs between persons (Lord & Novick, 1968). Whereas the fit of the 1-factor model is poor for congeneric measurements, the fit for tau-equivalent measurements is moderate, but not affected by population heterogeneity which supports the above indication of applicability of classical test theory.

Second, the results of the simulation study show that population heterogeneity has a clear effect on the variance of the regression predictor in the situation

of parallel and tau-equivalent measurements, but not congeneric measurements. It can be argued that having available N individual estimates of var$[RP]$ and var$[PP]$ obtained with parallel or tau-equivalent measurements, makes it possible to detect population heterogeneity. Under the hypothesis of population homogeneity and normality of the data, these estimates should be considered as N random samples of the chi-squared distribution.

Carefulness has to be taken with respect to the conclusion of population heterogeneity. The reliability of the test, the sample size, the number of repeated measurements, and the type of measure (parallel, tau-equivalent, congeneric) have to be taken into account.

Ansari et al., (2002) show that a Bayesian approach to confirmatory factor analysis can be very useful in dealing with heterogeneity. In closing, applying such a Bayesian factor model to the three types of measurements is a recommendable option in the further investigation and possible detection of population heterogeneity in the context of classical test theory.

References

Agresti, A. (1997). A model for repeated measurements of a multivariate binary response. *Journal of the American Statistical Association, 92*, 315-321.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.

Andersen, E.B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B, 34*, 42-54.

Andersen, E.B. (1995). Polytomous Rasch models and their estimation. In G. Fischer & I.W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 271-292). New York: Springer-Verlag.

Andersen, E.B. & Madsen, M. (1977). Estimating the parameters of the latent population distribution. *Psychometrika, 42*, 357-374.

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

Anderson, T.W. (1963). The use of factor analysis in the statistical analysis of multiple time series. *Psychometrika, 28*, 1-25.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Ansari, A., Jedidi, K., & Dube, L. (2002). Heterogeneous factor analysis models: A Bayesian approach. *Psychometrika, 67*, 49-78.

Arnold, B.C. & Robertson C.A. (1989). Autoregressive logistic processes. *Journal of Applied Probability, 26*, 524-531.

Baker, F.B. (1991). *Item response theory: Parameter estimation techniques.* New York: Marcel Dekker.

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

Bartholomew, D.J. & Knott, M. (1999). *Latent variable models and factor analysis* (2nd ed.). London: Edward Arnold.

Basilevsky, A. (1994). *Statistical factor analysis and related methods: Theory and applications.* New York: Wiley.

Bertoli-Barsotti, L. (2005). On the existence and the uniqueness of JML estimates for the partial credit model. *Psychometrika, 70*, 517-531.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinees ability. In F.M. Lord & M.R. Novick, *Statistical theories of mental test scores* (pp. 395-480). Reading, MA: Addison-Wesley.

Bock, R.D. & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35*, 179-197.

de Boeck, P. & Wilson, M. (Eds., 2004). *Explanatory item response models: A generalized linear and non-linear approach.* New York: Springer-Verlag.

Borkenau, P. & Ostendorf, F. (1998). The Big Five as states: How useful is the five-factor model to describe intraindividual changes over time? *Journal of Research in Personality, 32*, 202-221.

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*, 203-219.

Box, G.E.P. & Jenkins, G.M. (1976). *Time series analysis: Forecasting and control* (Rev. ed.). San Francisco: Holden Day.

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

Browne, M.W. & Nesselroade, J.R. (2005). Representing psychological processes with dynamic factor models: Some promising uses and extensions of ARMA time series models. In A. Maydeu-Olivares & J.J. McArdle (Eds.), *Psychometrics: A festschrift to Roderick P. McDonald* (pp. 415-452). Mahwah, NJ: Lawrence Erlbaum.

Browne, M.W. & Zhang, G. (2005). DyFA: Dynamic Factor Analysis of Lagged Correlation Matrices, Version 2.00 [Computer software and manual]. Retrieved from `http://quantrm2.psy.ohio-state.edu/browne`.

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

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995) A limited memory algorithm for bound constrained optimization. *SIAM Journal of Scientific Computing, 16*, 1190-1208.

Chow, S.-M., Ferrer, E., & Nesselroade, J.R. (2007). An unscented Kalman filter approach to the estimation of nonlinear dynamical systems models. *Multivariate Behavioral Research, 42*, 283-321.

Collins, L.M. & Sayer, A.G. (Eds., 2001). *New methods for the analysis of change.* Washington, DC: American Psychological Association.

Cudeck, R. & McCallum, R.C. (Eds., 2007). *Factor analysis at 100: Historical developments and future directions*. Mahwah, NJ: Lawrence Erlbaum.

De Hoon, M.J.L., van der Hagen, T.H.J.J., Schoonewelle, H., & van Dam, H. (1996). Why Yule-Walker should not be used for autoregressive modeling. *Annals of Nuclear Energy, 23*, 1219-1228.

Dolan, C.V. (2005). MKFM6 [Computer program]. Retrieved from `http://users.fmg.uva.nl/cdolan`.

Doucet, A., de Freitas, N., & Gordon, N. (Eds., 2001). *Sequential Monte Carlo methods in practice*. New York: Springer-Verlag.

Durbin, J. & Koopman, S.J. (1997). Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika, 84*, 669-684.

Durbin, J. & Koopman, S.J. (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. *Journal of the Royal Statistical Society, Series B, 62*, 3-56.

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

Ebbinghaus, H. (1885). *Memory: A contribution to experimental psychology*. New York: Dover.

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

Eggen, T.J.H.M. (2000). On the loss of information in conditional maximum likelihood estimation of item parameters. *Psychometrika, 65*, 337-362.

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

Ellis, J.L. & van den Wollenberg, A.L. (1993). Local homogeneity in latent trait models: A characterization of the homogeneous monotone IRT model. *Psychometrika, 58*, 419-429.

Embretson, S.E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika, 56*, 495-515.

Embretson, S.E. & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Fahrmeir, L. (1992). Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models. *Journal of the American Statistical Association, 87*, 501-509.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models* (2nd ed.). New York: Springer-Verlag.

Fahrmeir, L. & Wagenpfeil, S. (1997). Penalized likelihood estimation and iterative Kalman smoothing for non-Gaussian dynamic regression models. *Computational Statistics & Data Analysis, 24*, 295-320.

Fischer, G. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48*, 3-26.

Fischer, G. (1989). An IRT-based model for dichotomous longitudinal data. *Psychometrika, 54*, 599-624.

Fischer, G. & Molenaar, I.W. (Eds., 1995). *Rasch models: Foundations, recent developments, and applications*. New York: Springer-Verlag.

Fisher, R.A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.

Gill, P.E., Murray, W., Saunders, M.A., & Wright, M.H. (January, 1986). *User's guide for NPSOL (Version 4.0): A Fortran package for nonlinear programming*. Technical report SOL 86-2.

Glas, C.A.W. & Verhelst, N.D. (1995). Testing the Rasch model. In G. Fischer & I.W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 69-96). New York: Springer-Verlag.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

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

Hamaker, E.L., Dolan, C.V. & Molenaar, P.C.M. (2003). ARMA-based SEM when the number of time points T exceeds the number of cases N: Raw data maximum likelihood. *Structural Equation Modeling, 10*, 352-379.

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

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer.

Hamilton, J.D. (1994). *Time series analysis*. Princeton, NJ: Princeton University Press.

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

Harris, C.W. (Ed., 1963). *Problems in measuring change*. Menasha, WI: The University of Wisconsin Press.

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

Hoijtink, H. & Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch model. In G. Fischer & I.W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 53-68). New York: Springer-Verlag.

Holland, P.W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55*, 577-601.

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

Inhelder, B. & J. Piaget (1958). *The growth of logical thinking from childhood to adolescence*. New York: Basic Books.

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

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

Johnson, N., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions, Volume 1*. New York: Wiley.

de Jong, P. (1991). The diffuse Kalman filter. *Annals of Statistics 2*, 1073-1083.

de Jong, P. & Penzer, J. (2004). The ARMA model in state space form. *Statistics & Probability Letters, 70*, 119-125.

Jöreskog, K.G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109-133.

Jöreskog, K.G. (1994). On the estimation of polychoric correlations and their asymptotic correlation matrix. *Psychometrika, 59*, 381-389.

Jöreskog, K.G. & Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Books.

Jöreskog, K.G. & Sörbom, D. (1996). *LISREL 8: Reference guide.* Chicago: Scientific Software International.

Jöreskog, K.G. & Sörbom, D. (1999). *LISREL 8* [*Computer program*]. Chicago: Scientific Software International.

Julier, S.J. & Uhlmann, J.K. (1997). A new extension of the Kalman filter to nonlinear systems. *Proceedings SPIE, 3068*, 182-193.

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

Kakizawa, Y. (1999). Note on the asymptotic efficiency of sample covariances in Gaussian vector stationary processes. *Journal of Time Series Analysis, 20*, 551-558.

Kedem, B. & Fokianos, K. (2002). *Regression models for time series analysis.* Hoboken, NJ: Wiley.

Kelderman, H. & Molenaar, P.C.M. (2007). The effect of individual differences in factor loadings on the standard factor model. *Multivariate Behavioral Research, 42*, 435-456.

Kemeny, J.G., Snell, J.L., & Knapp, A.W. (1966). *Denumerable Markov chains.* Princeton, NJ: Van Nostrand.

Kempf, W.F. (1977). Dynamic models for measurement of "traits" in social behavior. In W.F. Kempf & B.H. Repp (Eds.), *Mathematical models for social pscyhology* (pp. 14-58). Bern, Switzerland: Hans Huber Publishers.

Kitagawa, G. (1987). Non-Gaussian state space modeling of nonstationary time series. *Journal of the American Statistical Association, 82*, 1032-1041.

Kitagawa, G. (1998). A self-organizing state space model. *Journal of the American Statistical Association, 93*, 1203-1215.

Klebaner, F.C. (1998). *Introduction to stochastic calculus with applications.* London: Imperial College Press.

Klein, B.M. (2003). *State space models for exponential family data.* Unpublished doctoral dissertation.

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

Lawley, D.N. & Maxwell, A.E. (1971). *Factor analysis as a statistical method.* London: Butterworth.

Lee, S.-Y. (Ed., 2007). *Handbook of latent variable and related models.* Amsterdam, The Netherlands: North-Holland.

van der Linden, W.J. & Hambleton, R.K. (Eds., 1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

Lord, F.M. (1952). A theory of test scores. *Psychometrika Monograph 7.*

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

Lütkepohl, H. (1991). *Introduction to multiple time series.* Berlin, Germany: Springer-Verlag.

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

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

Mackey, M.C. (1992). *Time's arrow: The origins of thermodynamic behavior.* New York: Springer-Verlag.

Magnus, J.R. & Neudecker, H. (1999). *Matrix differential calculus with applications in statistics and econometrics* (Rev. ed.). Chichester, England: Wiley.

Mair, P. & Hatzinger, R. (2007a). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*, 1-20.

Mair, P. & Hatzinger, R. (2007b). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science, 49*, 26-43.

Marcoulides, G.A. & Moustaki, I. (2002). *Latent variable and latent structure models.* Mahwah, NJ: Lawrence Erlbaum.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Masters, G.N. & Wright, B.D. (1997). The partial credit model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101-122). New York: Springer-Verlag.

McCrae, R.R. & John, O.P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*, 175-215.

McKenzie, E. (2003). Discrete variate time series. In D.N. Shanbhag & C.R. Rao (Eds.), *Handbook of Statistics, Volume 21: Stochastic processes: Modeling and simulation* (pp. 573-606). Amsterdam: Elsevier.

Mélard, G. (1984). A fast algorithm for the exact likelihood of autoregressive-moving average models. *Applied Statistics, 33*, 104-114.

Mellenbergh, G.J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115*, 300-307.

Mellenbergh, G.J. & Van den Brink, W.P. (1998). The measurement of individual change. *Psychological Methods, 3*, 470-485.

Millsap, R.E. (2001). When trivial constraints are not trivial: The choice of uniqueness constraints in confirmatory factor analysis. *Structural Equation Modeling, 8*, 1-17.

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

Molenaar, P.C.M. (1994). Dynamic latent variable models in developmental psychology. In A. von Eye & C.C. Clogg (Eds.), *Analysis of latent variables in developmental research* (pp. 155-180). Newbury Park, CA: Sage.

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

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, P.C.M., Huizenga, H.M., & Nesselroade, J.R. (2003). The relationship between the structure of intra-individual and inter-individual variability: A theoretical and empirical vindication of developmental systems theory. In U.M. Staudinger & U. Lindenberger (Eds.), *Understanding human development: Dialogues with lifespan psychology* (pp. 339-360). New York: Kluwer.

Molenaar, P.C.M. & Nesselroade, J.R. (1998). A comparison of pseudo-maximum likelihood and asymptotically distribution-free dynamic factor analysis param-

eter estimation in fitting covariance-structure models to block-Toeplitz matrices representing single-subject multivariate time series. *Multivariate Behavioral Research, 33*, 313-342.

Molenaar, P.C.M. & 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.

Molenaar, P.C.M. & von Eye, A. (1994). On the arbitrary nature of latent variables. In A. von Eye & C.C. Clogg (Eds.), *Latent variable analysis: Applications for developmental research* (pp. 226-242). Thousand Oaks, CA: Sage.

Moskowitz, D.S. & Hershberger, S.L. (Eds.) (2002). *Modeling intraindividual variability with repeated measures data: Methods and applications.* Mahwah, NJ: Erlbaum.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Murray, J.D. (1993). *Mathematical biology* (2nd ed.). New York: Springer-Verlag.

Muthén, B.O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557-585.

Nesselroade, J.R., McArdle, J.J., Aggen, S.H., & Meyers, J.M. (2002). Dynamic factor analysis models for representing process in multivariate time series. In D.M. Moskowitz & S.L. Hershberger (Eds.), *Modeling intraindividual variability with repeated measures data: methods and applications* (pp. 235-265). Mahwah, NJ: Erlbaum.

Nesselroade, J.R. & Schmidt McCollam, K.M. (2000). Putting the process back in developmental processes. *International Journal of Behavioral Development, 24*, 295-300.

Neyman, J. & Scott, E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrika, 16*, 1-32.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44*, 443-460.

Petersen, K. (1983). *Ergodic theory.* Cambridge, England: Cambridge University Press.

Popper, K.R. (1935). *Logik der Forschung* [The logic of scientific discovery]. Vienna: Springer-Verlag.

Porat, B. (1987). Some asymptotic properties of the sample autocovariances of Gaussian autoregressive moving average processes. *Journal of Time Series Analysis, 8*, 205-220.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna: Retrieved from `http://www.R-project.org`.

Rao, C.R. & Sinharay, S. (Eds., 2007). *Handbook of statistics, Volume 26: Psychometrics.* Amsterdam: Elsevier.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: The Danish Institute of Educational Research. (Expanded edition, 1980. Chicago: The University of Chicago Press.)

van Rijn, P.W. & Molenaar, P.C.M. (2005). Logistic models for single-subject time series. In L.A. van der Ark, M.A. Croon, & K. Sijtsma(Eds.), *New developments in categorical data analysis for the social and behavioral sciences* (pp. 125-146). London: Erlbaum.

Rizopoulos, D. (2006). Ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17*, 125.

Sage, A. & Melsa, J. (1971). *Estimation theory, with applications to communications and control.* New York: McGraw-Hill.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph 17.*

Schnekenburger, B.J. (1988). *A modified extended Kalman filter as parameter estimator for linear discrete-time systems.* Unpublished master thesis.

Shumway, R.H. & Stoffer, D.S. (2006). *Time series analysis and its applications with R examples* (2nd ed.). New York: Springer-Verlag.

Skondral, A. & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* Boca Raton, FL: Chapman & Hall.

Song, P.X.-K. (2000). Monte Carlo Kalman filter and smoothing for multivariate discrete state space models. *Canadian Journal of Statistics, 28*, 641-652.

Song, X.-Y. & Lee, S.-Y. (2003). Full maximum likelihood estimation of polychoric and polyserial correlations with missing data. *Multivariate Behavioral Research, 38*, 57-79.

Spearman, C.E. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47*, 175-186.

Tjøstheim, D. & Paulsen, J. (1983). Bias of some commonly-used time series estimates. *Biometrika, 70*, 389-399.

Tsutakawa, R.K. & Johnson, J.C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika, 55*, 371-390.

Verhelst, N.D. & Glas, C.A.W. (1993). A dynamic generalization of the Rasch model. *Psychometrika, 58*, 395-415.

Vidoni, P. (1999). Exponential family state space models based on a conjugate latent process. *Journal of the Royal Statistical Society, Series B, 61*, 213-221.

Visser, I., Raijmakers, M.E.J., & Molenaar, P.C.M. (2000). Confidence intervals for hidden Markov model parameters. *British Journal of Mathematical and Statistical Psychology, 53*, 317-327.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response models. *Psychometrika, 54*, 427-450.

West, M., Harrison, P.J., &, Migon, H.S. (1985). Dynamic generalized linear models and Bayesian forecasting. *Journal of the American Statistical Association, 80*, 73-83.

Wilson, M. & Masters, G.N. (1993). The partial credit model and null categories. *Psychometrika, 58*, 87-99.

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

Zimmerman, D.W., Williams, R.H., & Burkheimer, G.J. (1968). Dependence of test reliability upon heterogeneity of individual and group score distributions. *Educational and Psychological Measurement, 28*, 41-46.

Samenvatting (Summary in Dutch)

Dit proefschrift behandelt statistische modellen voor psychologische metingen uit de factor-analyse (FA) en item-respons-theorie (IRT). Het betreft hier echter niet, zoals gebruikelijk is in de psychologie, psychologische metingen verkregen bij verschillende personen, maar metingen verkregen bij dezelfde persoon op verschillende momenten in de tijd. Dat wil zeggen, psychologische metingen die een tijdreeks vormen. Speciale belangstelling gaat hierin uit naar een situatie die vaak voorkomt in de psychologie, namelijk de situatie waarin de psychologische meting kan worden geklassificeerd in slechts een beperkt aantal categorieën. Met andere woorden, dit proefschrift gaat over categorische tijdreeksen. Hoofdstuk 1 bestaat uit een korte motivatie voor het onderzoeken van categorische tijdreeksen, die gebaseerd is op ontwikkelingen en aandachtsgebieden in de psychologie en psychometrie. Het hoofdstuk eindigt met een overzicht van de hoofdstukken van dit proefschrift.

In Hoofdstuk 2 worden methoden voor het analyseren van multivariaat normale en categorische tijdreeksen onderzocht binnen het kader van structurele vergelijkingsmodellen ("structural equation modelling", SEM). Dit kader wordt gebruikt vanwege de bekendheid bij onderzoekers in de psychologie, en omdat er verscheidene standaard SEM softwarepakketten voorhanden zijn. Voor het geval van normaal verdeelde tijdreeksen, is onderzocht of de zogeheten Toeplitz matrix bestaande uit auto- en kruis-covarianties gebruikt kan worden om de parameters van verschillende autoregressieve modellen te schatten. Het gebruik van de Toeplitz matrix binnen het SEM kader heeft voordelen, omdat deze matrix eenvoudig is uit te rekenen en zodoende als invoer kan dienen voor SEM software waarmee het model kan worden geschat. De prestaties van een maximum likelihood (ML), gewogen kleinste-kwadraten ("weighted least squares", WLS), en, als referentie, Kalman filter (KF) schattingsprocedure zijn in een simulatie-onderzoek met elkaar vergeleken. De resultaten gaven aan dat de parameters en standaardfouten alleen correct werden geschat in het geval van pure vector autoregressies. Voor autoregressieve modellen met meerdere indicatoren waren de geschatte parameters correct, maar de standaardfouten niet.

Dezelfde benadering kan worden gebruikt voor categorische tijdreeksen. De Toeplitz matrix bestaat in dit geval echter uit polychorische auto- en kruis-correlaties. In een tweede simulatie-onderzoek zijn de prestaties van een WLS

schattingsprocedure onderzocht in termen van het terugschatten van parameters van autoregressieve modellen. De schattingen van parameters van de gebruikte modellen waren in de meeste gevallen correct, maar de standaardfouten niet. De resultaten van de simulatie-onderzoeken en de reeds beschikbare asymptotische resultaten van auto- en kruiscovarianties duiden erop dat de Toeplitz methode slechts beperkte bruikbaarheid kent. Dit moet niet worden gezien als argument tegen het gebruik van SEM in de analyse van tijdreeksen. Het is de methodologie van het gebruik van beschrijvende statistieken voor het schatten en passen van modellen, een reguliere gang van zaken in standaard toepassingen van SEM, die wordt afgeraden voor normaal verdeelde tijdreeksen en voor categorische tijdreeksen in het bijzonder.

In Hoofdstuk 3 worden univariate categorische tijdreeksen geanalyseerd binnen het referentiekader van toestand-ruimte modellen ("state space modelling", SSM). In een simulatie-onderzoek worden de prestaties van een Kalman filter en effen("smoothing")-procedure onderzocht voor het schatten van autoregressieve modellen voor categorische tijdreeksen. De procedure wordt tevens geïllustreerd door een toepassing op echte data. Uit de simulatie-resultaten bleek deze aanpak niet geheel naar bevrediging te werken voor univariate categorische tijdreeksen.

Hoofdstuk 4 van dit proefschrift behandelt in het eerste deel een argumentatie ten faveure van het analyseren van intra-individuele variatie. Er wordt geargumenteerd dat zulke analyses onderbelicht zijn in de psychologie en psychometrie. Het tweede deel betreft de presentatie van een logistisch model voor multivariate dichotome tijdreeksen, dat kan worden gezien als een dynamische extensie van het alomtegenwoordige Rasch model in IRT. Het model wordt geïllustreerd door voorbeelden met gesimuleerde en echte data.

Hoofdstuk 5 borduurt voort op Hoofdstuk 4. Uitbreidingen naar polytome en multi-subject tijdreeksen worden besproken binnen het SSM kader. In een voorbeeld wordt gedemonstreerd dat het SSM kader ook kan worden gebruikt voor gebruikelijke toepassingen van IRT. De resultaten van een Kalman filter en effen-procedure voor het schatten van item- en persoonsparameters in het geval van cross-sectionele data worden vergeleken met standaard IRT methoden. Het hoofdstuk eindigt met een toepassing van de SSM methoden op multi-subject polytome tijdreeksen.

Hoofdstuk 6 bevat de conclusies over de bevindingen in dit proefschrift en biedt wat handvaten voor toekomstig onderzoek op het gebied van categorische tijdreeksen in de psychologie. De conclusies kunnen als volgt worden samengevat. Het kader van SSM is zeer bruikbaar voor het analyseren van psychologische

metingen in de vorm van categorische tijdreeksen en sluit goed aan bij de voor veel onderzoekers bekende kaders van SEM en IRT. Met name, de SSM en IRT kaders vormen een goede combinatie voor het type categorisch tijdreeksen waarop de nadruk ligt in dit proefschrift. Echter, op het gebied van het ontwikkelen en vergelijken van schattingsmethoden voor modellen voor categorische tijdreeksen is meer onderzoek gewenst. Een belangrijk aspect in het modelleren van categorische tijdreeksen, dat in dit proefschrift is onderbelicht, is modelpassing. Het ontwikkelen en onderzoeken van geschikte passingsmaten voor modellen voor categorische tijdreeksen moet als een prioriteit worden gezien in toekomstig onderzoek. Tot slot bevat dit proefschrift een appendix, die een korte noot behelst over het gebruik van klassieke test theorie in het geval van heterogene populaties.