Table 2, The parameter estimates ***H***, ***F*** and μ

Separate factor loading matrices ***H*** (parameters $h_{i\alpha}$) for brand A and B:

	no purchase brand A	purchase brand A
state category 0	0.28	0.72
state category 1	0.97	0.03

	no purchase brand B	purchase brand B
state category 0	1.00	0
state category 1	0.47	0.53

Separate transition matrices ***F*** (parameters $f_{\beta\alpha}$) for district 1 and district 2:

	district 1	
	state category 0	state category 1
state category 0	0.91	0.09
state category 1	0.03	0.97

	district 2	
	state category 0	state category 1
state category 0	1.00	0
state category 1	0	1.00

Conditional probability $\mu_{\alpha,a}$ in timepoint $t=1$
for state category α given district a:

	state category 0	state category 1
district 1	0.39	0.61
district 2	0.57	0.43

ontvangen: 25-9-95
geaccepteerd: 16-9-96

The Mixing Approach as a Unifying Framework For Dynamic Multivariate Analysis

Jan de Leeuw,
UCLA Statistics Program

Catrien Bijleveld*,
University of Leiden

Kees van Montfort,
Free University Amsterdam

Frits Bijleveld,
University of Leiden

Abstract

We argue that many models for multivariate longitudinal and cross-sectional data analysis have a common ancestry. They all are based on the qualitative idea that if we knew the actual state of the world, the relations between the observed quantities would be truly simple. This is shown to lead directly to factor analysis, IRT, state space models, mixture densities, latent Markov chains, MIMIC, LISREL, and various other common models and techniques. We show how our approach provides a convenient framework for looking at these models. The EM algorithm can be used to estimate the unknown parameters.

An additional advantage of our approach is that it can incorporate continuous as well as interval, ordinal and categorical variables.

*Catrien Bijleveld, Department of Research Methodology and Psychometrics, University of Leiden, P.O. BOX 9555, 2300 RB Leiden, the Netherlands. E-mail address: byleveld@rulfsw.leidenuniv.nl

1 Introduction

Our starting point in this paper is that we want to describe the relationships between (possibly many) variables, and we want to describe this relationship in simple terms. We look for simplicity, not necessarily because we believe the world is simple, but because simple relationships are easier to manipulate and communicate.

For Quetelet, Galton, and the early Karl Pearson, the normal distribution was simple. When Pearson (1894) first came across non-normal variation in his biometric work, he tried to maintain this notion of simplicity by assuming that the sample came from a *mixture* of normal distributions. Thus normality was still the norm, but unfortunately the sample was impure, because it consisted of a mixture of types. If we could have separated the types by observation, we would have seen the normality, but because we couldn't the statistical analysis has to do the job instead. In the same way, the Pearson polychoric model is based on the notion that multivariate normal is simple. Unfortunately we can only observe discreticized versions of the variables, which means we observe multinormals "mixed" over cell contents.

In the same way for Spearman (1904), intelligence was simple. It was a construct much like the weight of an object, and the test was the spring balance. If w_i is the weight of object i, and a_j is the resistance of spring j, then Hooke's Law tells us that the extension of the spring is $y_{ij} = w_i a_j$. This is basically Spearman's model. Score on the test was proportional to the 'weight' of the subject and to the "resistance" of the test. All other relationships between the tests, if they were indeed proper tests of intelligence, were measurement errors, i.e. they were dictated by chance. If we select a population of persons with a fixed intelligence, then the tests will be perfectly uncorrelated. Correlation between tests is merely a consequence of the fact that we cannot select such "pure" populations, i.e. it is a consequence of the fact that our populations are of mixed intelligence. If we knew the "state" of the system, i.e. the person's intelligence, then the correlation would disappear.

This very same idea comes back in Lazarsfeld's (1968) latent class analysis, in a very simple discrete form. It also has dominated item analysis, or item response theory, ever since the work of Lawley (1944). In item response theory the basic assumption is called *local independence*, and the relationship between the variables is "explained" by mixing populations with local independence.

Factor analysis, latent class analysis, and item response theory are all special cases of the *analysis of inter-dependence*. All variables play the same symmetric role in the model, we do not measure any input to the system, only output. In the *analysis of dependence*, it is precisely the relationship between input and output variables that we are interested in, and the model is inherently asymmetric because of this.

In classical regression analysis there is a very simple model for each cell in the design. All cells corresponds with normal distributions with the same shape, with a cell-specific shift. The joint distribution of the predictors and the outcome is a mixture of the cell distributions, although in this case the mixing proportions are known. By using the design matrix, we automatically unmix the distribution. The mixing variable is completely known, and thus we can analyze conditionally on its values (we consider the design matrix as fixed and known).

The notion of local independence is applied most naturally in the analysis of dependence by using MIMIC models. MIMIC models, introduced by Jöreskog and Goldberger (1975), again revolve the notion of a *state*, similar to intelligence or ability. Within a given state, input and output are independent. Or, to put it differently, the state *splits* input and output, and all influence of the input on the output goes through the state. States are unobserved, as usual, and dependence of input on output comes about by mixing states.

MIMIC notions are easy to generalize to the longitudinal situation, in which we observe the same input-output system at various points in time. This defines a sequence of MIMIC models. Of course replicating a MIMIC model on independent individuals also leads to a sequence of MIMIC models, but in that case the models are unconnected, because of independence. In the case of temporal variation, we need to connect the models because of the time-dependence. The basic idea in dynamic multivariate analysis is to link the models through the state variables. Not only do the state variables split input and output, they also split points in time. Thus all information about the past is collected in the present state of the system, and if we knew the present state, our predictions would not be improved by knowing about the past. Given the present state, we agree with Henry Ford that "history is bunk".

In the next section we will make these notions more precise, but for the time being it suffices to observe that Jöreskog and Goldberger's marriage of factor analysis and regression analysis can be extended in the time dimension to include state space analysis. In time, we have linked MIMIC models, and these linked MIMIC models may be stacked on top of each other if we have independent replications. We are interested in the time evolution of the state, because that summarizes all the relevant information for prediction, and thus all the relevant dynamics in the system. If state in cross-sectional factor analysis is intelligence, then state in state space models in the same context is development of intelligence, with similar interpretations for ability.

It is of importance to emphasize that in cross-sectional latent variable theory, a great deal has been made out of the fact that input, output, and state can all be either discrete or continuous. Regression of output on state can have many different possible forms because of this reason. The basic notion of state, or of latent variables, or of conditional independence, is not related to the nature of the various regressions, which should be tailored to the problem at hand.

2 Dynamic Multivariate Models

The basic model we are interested in is drawn in Figure 1. Actually there are N such models, one for each individual. The individuals are independent. We write

$$\text{prob}[(\wedge_{i=1}^{N}\wedge_{t=1}^{T} y_{it})(\wedge_{i=1}^{N}\wedge_{t=0}^{T} z_{it})(\wedge_{i=1}^{N}\wedge_{t=1}^{T} x_{it})] \tag{2.1}$$

for the probability of observing the data X, Y and Z. Our basic task in this section is to derive a general expression for this probability, taking the properties of the model in Figure 1 into account. The key result used to translate directed acyclic graphs into statements about joint distributions is a simple one. We suppose that, given z_{it}, y_{it} is independent of all other variables. Also, given $z_{i,t-1}$ and x_{it}, z_{it} is independent of all other variables.

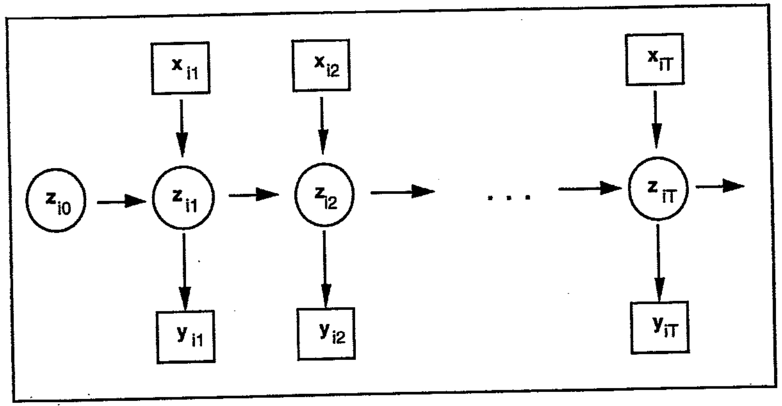

Figure 1: Linear dynamic model for individual i.

Under these conditions Theorem 2.1 is valid.

Theorem 2.1.

$$prob[(\wedge_{t=1}^{T}y_{it})(\wedge_{t=0}^{T}z_{it})(\wedge_{t=1}^{T}x_{it})] =$$
$$prob[\wedge_{t=1}^{T}x_{it} \mid z_{i0}]prob[z_{i0}]\prod_{t=1}^{T}prob[y_{it} \mid z_{it}]prob[z_{it} \mid z_{i,t-1} \wedge x_{i,t}]$$

Proof. The proof is by induction over T. The result is trivially true for $T = 1$. Assume it is true for $T - 1$. Start with a simple application of the definition of conditional probability.

$$\begin{aligned}\text{prob}[(\wedge_{t=1}^{T}y_{it})(\wedge_{t=0}^{T}z_{it})(\wedge_{t=1}^{T}x_{it})] = \\ \text{prob}[y_{iT} \mid (\wedge_{t=1}^{T-1}y_{it})(\wedge_{t=0}^{T}z_{it})(\wedge_{t=1}^{T}x_{it})] \times \\ \text{prob}[z_{iT} \mid (\wedge_{t=1}^{T-1}y_{it})(\wedge_{t=0}^{T-1}z_{it})(\wedge_{t=1}^{T}x_{it})] \times \\ \text{prob}[x_{iT} \mid (\wedge_{t=1}^{T-1}y_{it})(\wedge_{t=0}^{T-1}z_{it})(\wedge_{t=1}^{T-1}x_{it})] \times \\ \text{prob}[(\wedge_{t=1}^{T-1}y_{it})(\wedge_{t=0}^{T-1}z_{it})(\wedge_{t=1}^{T-1}x_{it})]\end{aligned}$$

The assumption about y_{it}, z_{it} and x_{it} tells us

$$\text{prob}[y_{iT} \mid (\wedge_{t=1}^{T-1}y_{it})(\wedge_{t=0}^{T}z_{it})(\wedge_{t=1}^{T}x_{it})] = \text{prob}[y_{iT} \mid z_{iT}],$$

and

$$\text{prob}[z_{iT} \mid (\wedge_{t=1}^{T-1}y_{it})(\wedge_{t=0}^{T-1}z_{it})(\wedge_{t=1}^{T}x_{it})] = \text{prob}[z_{iT} \mid z_{i,T-1} \wedge x_{i,T}],$$

and

$$\text{prob}[x_{iT} \mid (\wedge_{t=1}^{T-1} y_{it})(\wedge_{t=0}^{T-1} z_{it})(\wedge_{t=1}^{T-1} x_{it})] = \text{prob}[x_{iT} \mid \wedge_{t=1}^{T-1} x_{it} \wedge z_{i0}].$$

But this means that we have proved the recursion

$$\begin{aligned}&\text{prob}[(\wedge_{t=1}^{T} y_{it})(\wedge_{t=0}^{T} z_{it})(\wedge_{t=1}^{T} x_{it})] = \\ &\qquad \text{prob}[y_{iT} \mid z_{iT}]\text{prob}[z_{iT} \mid z_{i,T-1} \wedge x_{i,T}]\text{prob}[x_{iT} \mid \wedge_{t=1}^{T-1} x_{it} \wedge z_{i0}] \\ &\qquad\qquad \text{prob}[(\wedge_{t=1}^{T-1} y_{it})(\wedge_{t=0}^{T-1} z_{it})(\wedge_{t=1}^{T-1} x_{it})].\end{aligned}$$

By the induction hypothesis this means the result is true for T. □

We now introduce some simplifying assumptions, which just serve to make the final result easier to write down. If necessary, they can be gotten rid of again.

Corollary 2.2. *If*

$$prob[\wedge_{t=1}^{T} x_{it} \mid z_{i0}] = prob[\wedge_{t=1}^{T} x_{it}]$$

and z_{i0} is a.s. equal to a constant, then

$$\begin{aligned}&prob[\wedge_{t=1}^{T} y_{it} \mid \wedge_{t=1}^{T} x_{it}] = \\ &\qquad \int \cdots \int_{z_{i0},\ldots,z_{iT}} \prod_{t=1}^{T} prob[y_{it} \mid z_{it}]prob[z_{it} \mid z_{i,t-1} \wedge x_{i,t}]dz_{iT} \ldots dz_{i0}.\end{aligned}$$

Proof. Start with the result in Theorem 2.1. We remove the marginal distribution of the input variables by conditioning, and then integrate out the state variables. □

We see that the latent variables or state variables serve two purposes. They mediate the effect of input on output, and they channel the effect of the past on the present. Actually, the state space process is first-order Markov, although the observed output process can be much more complicated. The first-order Markov property is the basic notion of simplicity used in this context. It is clear that the state variables, with their double function, have to do a lot of work, and consequently the dimensionality of the state space (the "number of factors") may have to be quite big for a satisfactory fit.

3 Specific Submodels

There are a number of useful distinctions that can be drawn in discussing this class of models. In the first place there are models with and without input. There are models in which the state variables are discrete, and models in which they are continuous. In some models the input and/or output variables are discrete, in others continuous. There are models which are cross-sectional, in the sense that $T = 1$, and models which are time-series, in the sense that $N = 1$. Discussing the models in these terms shows that they do indeed cover a lot of the latent variable models discussed in psychometrics and other disciplines.

We shall discuss a number of these special cases in a little bit more detail. What we propose here is a simple and straightforward widening of the framework introduced by Lazarsfeld and Henry (1968) and Guttman (1941) in the forties, and then extended by Anderson (1959), McDonald (1962), and Bartholomew (1987) for cross-sectional models, and of the framework discussed, for example, by Metz (1977) for time series.

As mentioned in the introduction, in the class of cross-sectional models without input we find factor analysis (continuous state, continuous output), latent class analysis (discrete state, discrete output), latent profile analysis (discrete state, continuous output), latent trait analysis (continuous state, discrete output), and of course various combinations of these techniques. MIMIC models are cross-sectional with input, and again we can have discrete/continuous state-space and discrete/continuous input/output to describe various MIMIC variations. Classical state space models are usually for the time-series situation, in which $N = 1$, although this is by no means necessary.

If $N = 1$ the state space model has far too many parameters, and we need to get statistical stability from additional assumptions. The obvious one is *stationarity*, which means that the "structural" parameters of the model are constant over time. In this way observing more time points gives us more information about these parameters, and thus, in the limit, they can be estimated consistently. If $N >> 1$ we can hope to estimate nonstationary models, such as the discrete latent Markov chains discussed by Van der Pol and De Leeuw (1986), or the continuous LISREL-type models discussed by MacCallum and Ashby (1986) and Oud *et al.* (1990).

There is an elegant device, familiar from the psychometric tradition, but actually starting with Pearson, which makes it possible to generate models with discrete (or truncated, or transformed) output from models with continuous output. This can be translated in terms of simplicity. We think the continuous models are simple, and we build models for the observed variables on the basis of these simple models. Before we discuss these, we shall look at continuous models in more detail.

4 Continuous output variables

Models with continuous variables remain important, because they are parsimonious, and because they can be used as stepping stones to construct models with continuous latent indicators. The previous formulations of the model were given in terms of the density or probability mass function. Now we will introduce model parameters and will switch to the equivalent formulation in terms of random variables, which we distinguish from fixed quantities by underlining. The conditional expected value and variance of the output given the input are computed.

The natural assumptions in this case are linearity of regression and homoscedasticity. We look at the equations for a fixed value of i. More precisely, we assume that the conditional expectation of $\underline{y}_t$ given $\wedge_{s=0}^{t}\underline{z}_s$ and $\wedge_{s=1}^{t}\underline{x}_s$ depends linearly on $\underline{z}_t$ only. Moreover, the conditional expectation of $\underline{z}_t$ given $\wedge_{s=0}^{t-1}\underline{z}_s$ and $\wedge_{s=1}^{t}\underline{x}_s$ depends on $\underline{z}_{t-1}$ and $\underline{x}_t$ only. Thus

$$\underline{y}_t = H_t\underline{z}_t + \underline{\delta}_t, \tag{4.1a}$$

$$\underline{z}_t = F_t\underline{z}_{t-1} + G_t\underline{x}_t + \underline{\epsilon}_t, \tag{4.1b}$$

where

$$\begin{aligned} &\underline{\delta}_t \perp \wedge_{s=0}^{t} \underline{z}_s, \\ &\underline{\delta}_t \perp \wedge_{s=1}^{t} \underline{x}_s, \\ &\underline{\epsilon}_t \perp \wedge_{s=0}^{t-1} \underline{z}_s, \\ &\underline{\epsilon}_t \perp \wedge_{s=1}^{t} \underline{x}_s, \end{aligned}$$

and homoscedasticity means

$$\begin{aligned} \mathbf{V}(\underline{\delta}_t) &= \Omega, \\ \mathbf{V}(\underline{\epsilon}_t) &= \Theta. \end{aligned}$$

These equations define the discrete linear system, made famous by Kalman. In these classical state space models it is usually assumed that the matrices F_t, G_t and H_t (and even Ω and Θ) are known from the physics of the problem, and only the state variables have to be estimated. For this the Kalman filter is applied, which is in this sense a method for computing "factor scores".

Lemma 4.1.

$$\begin{aligned} &\underline{\delta}_t \perp \wedge_{s=1}^{t} \underline{\epsilon}_s, \\ &\underline{\epsilon}_t \perp \wedge_{s=1}^{t-1} \underline{\epsilon}_s. \end{aligned}$$

□

Proof. Again we use recursion.

Theorem 4.2. *Suppose, again to simplify notation, that $\underline{z}_{i0}$ is a.s. equal to zero. We define*

$$P_{ts} = \begin{cases} \prod_{k=s+1}^{t} F_k & s < t \\ I & s = t \end{cases}.$$

Then

$$\mathbf{E}\left(\underline{y}_t \mid \underline{z}_0 \wedge_{s=1}^{T} \underline{x}_s\right) = \sum_{s=1}^{t} H_t P_{ts} G_s \underline{x}_s,$$

$$\mathbf{V}\left(\underline{y}_t \mid \underline{z}_0 \wedge_{s=1}^{T} \underline{x}_s\right) = \Theta + \sum_{s=1}^{t} H_t P_{ts} \Omega P_{ts}' H_t',$$

while

$$\mathbf{C}\left(\underline{y}_t, \underline{y}_v\right) = \sum_{s=1}^{\min t,v} H_t P_{ts} \Omega_s P_{vs}' H_v'.$$

Proof. We can solve the stochastic difference equation (4.1), and we obtain

$$\underline{z}_t = \sum_{s=1}^{t} P_{ts} G_s \underline{x}_s + \sum_{s=1}^{t} P_{ts} \underline{\epsilon}_s,$$

and thus

$$\underline{y}_t = \sum_{s=1}^{t} H_t P_{ts} G_s \underline{x}_s + \sum_{s=1}^{t} H_t P_{ts} \underline{\epsilon}_s + \underline{\delta}_t.$$

□

The Theorem follows directly.

5 Latent output variables

If our models have categorical, ordinal or interval output variables, but we have the idea that these outcomes are really an imprecisely observed continuous process, then we can apply the same ideas as used by Pearson in his polychoric models. Actually, there are two versions of this idea, which both fit rather neatly into the general framework.

In the *first* approach we change the formulation somewhat to obtain the model

$$\text{prob}[\wedge_{t=1}^{T}\eta_{it} \mid \wedge_{t=1}^{T}x_{it}] = \int\cdots\int_{z_{i0},\ldots,z_{iT}} \prod_{t=1}^{T} \text{prob}[\eta_{it} \mid z_{it}]\text{prob}[z_{it} \mid z_{i,t-1} \wedge x_{i,t}]dz_{iT}\ldots dz_{i0}.$$

The blocks y_{it} have been replaced by the blocks η_{it}, which are the unobserved or latent output variables. We assume that the η_{it} have some known distribution. The idea here is similar to the latent state as an indicator of the output: the y_{it} are a function of an unobserved or latent output variable (See Figure 2). Such a change of formulation is useful when the output variables have been measured at lower than interval level.

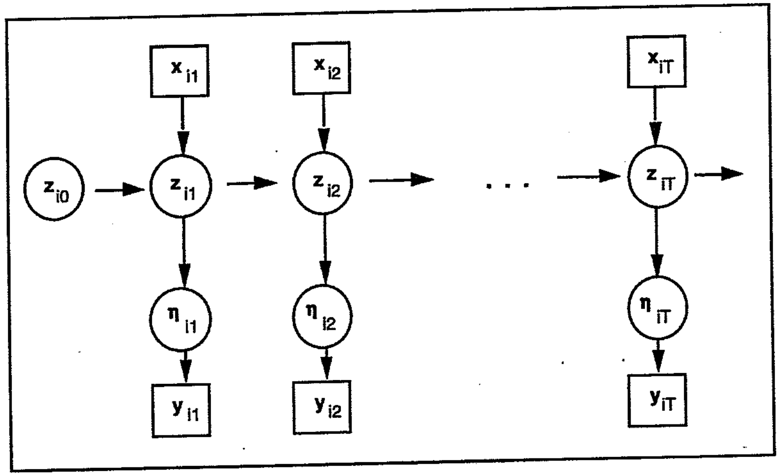

Figure 2: Linear dynamic model for individual i, with y_{it} an observed and η_{it} an unobserved output variable.

Suppose block η_{it} has m_t variables, indexed by j. We allow the number of input and output variables to differ for each t. In that case there is no reason to assume that the number of factors or states would be the same for each t. Of course, we may constrain m_t

to be equal for each time point if that would be appropriate for the particular analysis conducted.

We now have to connect the observed with the unobserved output, and this can be done by assuming some kind of *transformation* for each separate observed output variable:

$$y_{ijt} = F_{jt}(\eta_{ijt}, \lambda),$$

where F_{jt} is a deterministic transformation which depends on a number of parameters, collected in λ. In practice we will often assume F_{jt} to be time-invariant, so that $F_{jt} \equiv F_j$. One obvious case, for instance, for assuming so is when y_{ijt} is the same for every t. This implies that each variable j has a unique transformation. It is not necessary that λ be known beforehand as it can be estimated in a later stage as well.

It follows that:

$$\text{prob}[\wedge_{t=1}^{T}y_{it} \mid \wedge_{t=1}^{T}x_{it}] = \text{prob}[\wedge_{t=1}^{T}y_{it} \mid \wedge_{t=1}^{T}\eta_{it}]\text{prob}[\wedge_{t=1}^{T}\eta_{it} \mid \wedge_{t=1}^{T}x_{it}]. \tag{5.1}$$

Because of the Markov property of z_{it} and because of the fact that y_{it} depends on η_{it} only, it follows that

$$\text{prob}[\wedge_{t=1}^{T}y_{it} \mid \wedge_{t=1}^{T}\eta_{it}] = \prod_{t=1}^{T}\text{prob}[y_{it} \mid \eta_{it}] \tag{5.2}$$

For *nominal* variables, we have a lot of freedom in our choice of transformation, given the categorical properties of the variables: the new values for the categories can in fact be chosen freely as long as they satisfy equality and inequality constraints. For nominal variables, we thus assume that $F_{jt}(\bullet, \lambda)$ is a step-function, mapping the real line into $\{c_{1jt}, \cdots, c_{kjt}\}$, with c_{ijt} the i-th category of variable j at timepoint t. Then, using (5.1) and (5.2) we can write

$$\text{prob}[\wedge_{t=1}^{T}y_{it} \mid \wedge_{t=1}^{T}x_{it}] = \int_{\mathcal{B}_i(\lambda)}\text{prob}[\wedge_{t=1}^{T}\eta_{it} \mid \wedge_{t=1}^{T}x_{it}]\prod_{t=1}^{T}d\eta_{it},$$

where

$$\mathcal{B}_i(\lambda) = \otimes_{t=1}^{T}\otimes_{j=1}^{m_t}F_{jt}^{-1}(y_{ijt}, \lambda).$$

For *ordinal* variables, we have less freedom in our choice of transformation, as the ordering of the categories has to be preserved. This implies that we assume, in the tradition of the Box-Cox approach, that the transformation is a differentiable and strictly monotone function for each parameter λ, in which case

$$\mathrm{prob}[\wedge_{t=1}^{T} y_{it} \mid \wedge_{t=1}^{T} x_{it}] = \mathrm{prob}[\wedge_{t=1}^{T} F^{-1}(y_{it}) \mid \wedge_{t=1}^{T} x_{it}] \prod_{t=1}^{T} \frac{\partial F_t^{-1}(y_{it}, \lambda)}{\partial y_{it}}$$

$F_t(\bullet, \lambda)$ can be, amongst others, a step function or a Box-Cox transformation function. Note how the formulation for ordinal variables is thus a special case of the more general formulation we used above for categorical variables.

A *second* approach is to use mixture distributions. In such a mixing approach the above results simplify to a concise expression:

$$\mathrm{prob}[y_{it} \mid \eta_{it}] = \prod_{j}^{m_t} \mathrm{prob}[y_{ijt} \mid \eta_{ijt}],$$

$$\begin{aligned}\mathrm{prob}[\wedge_{t=1}^{T} y_{it} \mid \wedge_{t=1}^{T} x_{it}] = \\ \int \cdots \int_{\{\eta_{ijt}\}} \prod_{t=1}^{T} \prod_{j=1}^{m_t} \mathrm{prob}[y_{ijt} \mid \eta_{ijt}] \\ \mathrm{prob}[\wedge_{t=1}^{T} \wedge_{j=1}^{m_t} \eta_{ijt} \mid \wedge_{t=1}^{T} x_{it}] \prod_{t=1}^{T} \prod_{j=1}^{m_t} d\eta_{ijt}.\end{aligned}$$

Note how in the mixing approach, transformations for nominal and ordinal variables are combined in a uniform formulation.

The mixing concept, in which the observed output variables are transformations of layers of mixes, can also be applied to distribution-free observed output variables which are continuous. In that case the distributions of the observed variables are transformations or mixes of specified distributions of latent variables.

We have now arrived at a point where relatively simple formulas have been derived for $\mathrm{prob}[\wedge_{t=1}^{T} y_{it} \mid \wedge_{t=1}^{T} x_{it}]$. These simple formulas contain product terms which are easy to work with. Given the structure of the formulas, the EM algorithm is an obvious choice for estimating the unknowns.

We will now show how to fit the model to data, simultaneously finding optimal transformations for output variables.

6 The EM algorithm

Often a sufficiently general framework, which encompasses a lot of different techniques, comes with a "natural" class of algorithms. These algorithms may not always be optimal for any particular special case, but they are usually of a simple structure, and they are guaranteed to be available for any model in the class. For the ALSOS approach to multivariate analysis (Gifi, 1990), these were the Alternating Least Squares algorithms, and for the mixing or latent variable approach in this paper this is the *Expectation/Maximization* or *EM* method, presented first by Dempster *et al.* (1977).

For a discussion of the EM algorithm as a special type of *majorization algorithm*, with majorization provided by Jensen's inequality, we refer to de Leeuw (1994). For our purposes we merely point out that EM algorithms maximize functions of the form

$$g(\theta) \triangleq \log \int f(x, \theta) dx$$

by solving a sequence of maximization problems. The function maximized in step k is

$$h(\theta, \theta^{(k)}) \triangleq \int f(x \mid \theta^{(k)}) \log f(x, \theta) dx, \tag{6.1}$$

where

$$f(x \mid \theta^{(k)}) \triangleq \frac{f(x, \theta^{(k)})}{\int f(u, \theta^{(k)}) du},$$

and $\theta^{(k)}$ contains the last estimates of the parameters in the iterative sequence.

From corollary 2.2 we can now formulate the following theorem.

Theorem 6.1. *In a step of the EM algorithm we maximize, if the current set of parameters is ξ, the following function of θ:*

$$\begin{aligned}\sum_{t=1}^{T} \int_{z_{it}} prob_\xi(z \mid x \wedge y) \log prob_\theta[y_{it} \mid z_{it}] dz_{it} + \\ \sum_{t=1}^{T} \int_{z_{it}} \int_{z_{i,t-1}} prob_\xi(z \mid x \wedge y) \log prob_\theta[z_{it} \mid z_{i,t-1} \wedge x_{it}] dz_{it} dz_{i,t-1}\end{aligned}$$

where

$$prob_\xi(z \mid x \wedge y) \triangleq prob_\xi(\wedge_{t=1}^{T} z_{it} \mid [\wedge_{t=1}^{T} x_{it}] \ \wedge \ [\wedge_{t=1}^{T} y_{it}]),$$

and the ξ are identical to the $\theta^{(k)}$ in (6.1)

Proof. This is just a matter of substituting our multivariate dynamic model in the majorization function, and simplifying. □

It is obvious from the Theorem that the simplifications resulting from majorization will be especially impressive in the case of exponential families (for instance normal distributions), where the logarithm inside the integral sign reduces computing the integral to computing the expected value of a sufficient statistic.

Using the general principles of EM algorithm construction, it is also possible to construct methods to optimize the likelihood functions for the models in the previous section. This is true for the ones using the Box-Cox approach, and for the ones with use additional mixing.

7 Conclusion and discussion

In the above, we presented a framework that encompasses a wide class of models. The class includes amongst others multiple regression, factor analysis, MIMIC-type models and multivariate dynamic models. Furtermore, distribution-free continuous, as well as interval, ordinal and nominal variables can be taken into account. The general approach for estimating the model parameters uses the EM-algorithm.

A first question to be answered is what practical benefits our approach has over existing methodologies for similar data analytic situations. The most obvious competing framework is provided by structural equations modeling. There is much recent research dealing with modeling the dependence structure of variables using causal models, of which structural equations models are a subclass. The modeling should be done in such a way that the structural properties of the graphs that depict the relations between the variables, such as connectivity and separation, are isomorphic with the dependence structure of the variables. This is often called "causal modeling", a description which is unnecessarily controversial. The controversy arises from attributing too much truth to the models, and from not using models as what they are, namely tools. Models are tools in the sense that they are filters, or smoothers, which can be used to bring data in a form which is more interesting, more easy to communicate, or more easy to relate to existing theory. We separate the signal from the noise by using prior theory whenever it is available, and by using inductive techniques otherwise. We can talk about our results in causal language, as long as we realize that this language is simply another (verbal) model with which we

overlay the statistical analysis. Verbal models are however by nature vague, and may thus lead to misconceptions.

A second controversy arises from the fact that the so-called causal models have also been used, mainly in the social and behavioural sciences, to dress up weak data. Using causal terminology suggest a certainty which is, however, mostly absent. In many cases, the outcomes of such models depend largely on accidental properties of the data and arbitrary choices of the researcher. More particularly, there are many reasons to distrust highly specific models in which some of the relations between variables are present, and some of the other comparable relations are absent, and there are no clear reasons for either presence or absence. This has given causal models a bad name, although obviously the problem is not with the model but with the data and the way the model is applied to the data (and perhaps the way the techniques have been sold commercially).

Contrary to common practice with causal models, we propose to use full models, in which the dependence that is modeled depends on global and fairly uncontroversial choices. Exploratory factor analysis is one example, multiple regression is another. All the relations are there between the predictors and the criterion, or between the factors and the indicators. Such full models are much more descriptive than the models whose fine structure suggests much more prior knowledge than we actually have. The former can be used as data reduction techniques, and in fact in most cases they are not far from saturated models. Multiple regression and complete recursive path models are saturated, factor analysis and the state space models we discuss here can be made saturated by introducing sufficiently many latent variables.

The choices we have to make to construct a model are really simple. In multiple regression we only have to choose which variable is the criterion, in factor analysis we only have to choose the number of factors, in MIMIC type models we have to distinguish input and output variables and choose the number of factors. In linear dynamic systems we have to order the blocks of input, state, and output variables in time. These are simple global choices, with which few people will disagree. The filtering done by the model is entirely in the dimensionality of the state space or factor space, and we easily see the effect of this by looking at different dimensionalities.

Finally we have to mention that further research is needed to show the practical relevance of our general framework. For instance, the convergence speed of the EM algorithm in our context has to be investigated. Also, using generated data, comparisons with competing methods for submodels will have to be made, the most notable being the structural equations models (Bentler (1989); Jöreskog (1988)), the nonlinear dynamic systems models proposed by Bijleveld and de Leeuw (1991) and the latent Markov models proposed by Langeheine and others (Langeheine and van de Pol (1990); Van der Pol and De Leeuw (1986)).

References

Anderson, T. W. (1959). Some scaling models and estimation procedures in the latent class model. In U. Grenander, editor, *Probability and Statistics. The Harald Cramér Volume*. Almqvist and Wicksell, Stockholm, Sweden.

Bartholomew, D. J. (1987). *Latent Variable Models and Factor Analysis*. Griffin, London, GB.

Bentler, P. M. (1989). *EQS*. BMDP Statistical Software, Los Angeles.

Bijleveld, C. C. J. H. and de Leeuw, J. (1991). Fitting longitudinal reduced-rank regression models by alternating least squares. *Psychometrika*, **56**, 433–447.

de Leeuw, J. (1994). Block relaxation algorithms in statistics. In H. H. Bock, W. Lenski, and M. M. Richter, editors, *Informations Systems and Data Analysis.*, pages 308–325. Springer, New York.

Dempster, A. P., Liard, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, **B**(34), 183–202.

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. John Wiley & Sons.

Guttman, L. (1941). The quantification of a class of attributes: a theory and method of scale construction. In P. Horst, editor, *The Prediction of Personal Adjustment.*, New York. SSRC.

Jöreskog, K. G. (1988). *Lisrel 7*. SPSS Inc., Chicago.

Jöreskog, K. G. and Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, **70**, 631–639.

Langeheine, R. and van de Pol, F. (1990). A unifying framework for Markov modeling in discrete space and discrete time. *Sociological Methods & Research*, **18**, 416–441.

Lawley, D. N. (1944). The factorial analysis of multiple item tests. In *Proceedings of the Royal Society*, number 62, pages 74–82, Edinburgh.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Ananysis*. Houghton Mifflin, Boston.

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

McDonald, R. P. (1962). A note on the derivation of the general latent class model. *Psychometrika*, **27**, 203–206.

Metz, J. A. J. (1977). Statespace models for animal behaviour. *Annals of System Research*, (6), 65–109.

Oud, J. H. L., Van den Bercken, J. H. L., and Essers, R. J. (1990). Longitudinal factor scores estimation using the Kalman filter. *Applied Psychological Measurement*, (14), 395–418.

Pearson, K. (1894). Contribution to the mathematical theory of evolution. *Philosophical Transactions of the Royal Statistical Society of London.*, **A**(185), 71–110.

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

Van der Pol, F. and De Leeuw, J. (1986). A latent Markov model to correct for measurement error. *Sociological Methods and Research*, (15), 118–141.

ontvangen: 23-12-95
geaccepteerd: 4- 9-96