Published in final edited form as:

Stat Med. 2013 October 30; 32(24): 4229–4239. doi:10.1002/sim.5825.

A Latent Factor Linear Mixed Model for High Dimensional Longitudinal Data Analysis

Xinming An^a, Qing Yang^c, and Peter M Bentler^b

^aSAS Institute Inc

^bDepartments of Psychology and Statistics, UCLA

^cDepartment of Biostatistics, UCLA

Abstract

High dimensional longitudinal data involving latent variables such as depression and anxiety that cannot be quantified directly are often encountered in biomedical and social sciences. Multiple responses are used to characterize these latent quantities, and repeated measures are collected to capture their trends over time. Furthermore, substantive research questions may concern issues such as interrelated trends among latent variables that can only be addressed by modeling them jointly. While statistical analysis of univariate longitudinal data has been well developed, methods for modeling multivariate high dimensional longitudinal data are still under development. In this paper we propose a latent factor linear mixed model (LFLMM) for analyzing this type of data. This model is a combination of the factor analysis and multivariate linear mixed models. Under this modeling framework, the high dimensional responses are reduced to low dimensional latent factors by the factor analysis model, while the multivariate linear mixed model is used to study the longitudinal trends of these latent factors. An EM algorithm is developed to estimate the model. Simulation studies are used to investigate the computational properties of the EM algorithm and compare the LFLMM model with other approaches for high dimensional longitudinal data analysis. A real data example is used to illustrate the practical usefulness of the model.

Keywords

longitudinal data; factor analysis model; linear mixed model

1. Introduction

Longitudinal data gathering designs play a critical role in various research areas, including medical, social and behavioral sciences, because they allow the exploration of change over time as well as the identification of factors that influence the change patterns [1]. A large number of statistical modeling techniques have been proposed for the analysis of longitudinal data, including mixed models to handle various response types and change patterns [2]. Typically these approaches accept the observed responses as the main variables of interest. However, there are situations where the observed responses are best considered as partially adequate indicators of one or more underlying latent variables that cannot be observed or quantified directly. A trivial but almost universal example occurs when observed variables contain varied and unknown amounts of measurement error variance

whose presence can bias parameter estimates as well as scientific conclusions. In such cases, it is critical to be able to focus on longitudinal trends among multiple latent variables using joint modeling to draw joint inference for different outcomes. For example, research interest may focus on the association of changes for multiple latent quantities over time [3]. Since inference on latent variables furthermore requires understanding their effects on observed variables, a large number of longitudinal responses need to be collected and analyzed jointly.

While statistical analysis of univariate longitudinal data is well developed and widely available in many statistical packages, such as the mixed proc in SAS, how to appropriately and efficiently model multivariate or high dimensional longitudinal data has been a hot research topic during the last decade [4]. The challenges are introduced by the complex correlation structures among the repeatedly measured multiple outcomes. Valid inferences can be preserved only when these complex correlation structures are well addressed. These correlations include (1) the correlation between repeated measurement within each outcome; (2) the correlation between outcomes at a specific time point; (3) the correlation between outcomes at different time points.

A number of approaches have been proposed to characterize these correlations. These approaches can be classified into different categories based on how they model these correlations. The first category includes the multivariate marginal models [e.g., 5, 6] and conditional models [e.g., 7] that try to model the joint density either directly or as a product of marginal and conditional densities. Some drawbacks of these approaches are that they can not handle unbalanced data or be easily extended to dimensions higher than two. Approaches in the second category use random effects to account for the correlations between repeated measurements within each outcome. The correlation between outcomes at a specific time point or at different time point are modeled by introducing different random effects for different outcomes. Notable examples in this category include Reinsel [8], Shah et al. [9], Gueorguieva and Agresti [10]. While these approaches can be extended to more than two outcomes, the total number of random effects grows rapidly as the number of outcomes increases. Since a large number of random effects often lead to computational difficulties, especially when some of the outcomes are best described by generalized linear or non-linear models, these approaches are not feasible for the analysis of high dimensional longitudinal data.

Other approaches that are more suitable for high dimensional longitudinal data often incorporate dimension reduction techniques into the modeling process. One technique that has been widely used by applied researchers includes two separate steps. In the first step, high dimensional responses are reduced to low dimensional factor scores by either factor analysis, principal components analysis or simply the average of the raw scores. Then in the second step, these factor scores are analyzed using any of the classical longitudinal models. While these stepwise approaches are conceptually and computationally easy and straightforward, they do have some drawbacks: (1) estimators produced by these approaches are biased [11]; (2) approaches based on factor analysis and principal components can not be used for highly unbalanced data.

Simultaneous modeling can avoid all or some of the above disadvantages associated with the stepwise procedures. In the fields of social sciences, such as psychometrics, structural equation models (SEM) are often used to model longitudinal data [12]. This approach has several appealing modeling abilities and can be used for high dimensional longitudinal data [13, chapter 7]. Under the SEM framework, missing data can be handled using full information maximum likelihood estimation under the cost of computational efficiency. In the case of highly unbalanced longitudinal data, this approach becomes cumbersome. Covariates other than the time variable can also be added into the model, but it is not as flexible as the linear mixed model. For example, covariates can not serve as random but only fixed covariates and they can not have interactions with the time variable. Voelkle et al. [14] proposed the continuous time SEM model to avoid some issues associated with the autoregressive and cross-lagged models. Since the model is formulated using the traditional SEM model, as stated by the authors, the same assumptions and limitations of SEM apply to the continuous time SEM. Dunson [15] and Cagnone et al. [16] combined factor analysis with transitional models to analyze multivariate longitudinal data with mixed outcome types. Roy and Lin [17] proposed a so called latent-variable linear mixed model for the analysis of multivariate continuous longitudinal data, which is a combination of factor analysis and linear mixed models. As compared with the other approaches, this model enjoys the following advantages: (1) it shares the advantages associated with mixed models as compared with marginal and transitional models; (2) it can handle more complex data structures, for example highly unbalanced data; (3) other covariates can be included into the model very easily. Liu and Hedeker [18] combined the item response theory (IRT) model with the linear mixed model for the analysis of longitudinal multivariate ordinal data. However, the number of latent factors is restricted to be one in these two models. This restrictive assumption undermines their capability for the high dimensional situation, since it is unrealistic to assume one latent factor for a large number of outcomes.

In this paper, we propose a modeling framework similar to Roy and Lin [17] that will release the restriction on the number of latent factors. This model is a combination of factor analysis and multivariate linear mixed models and is called latent factor linear mixed model (LFLMM). The idea of combining different modeling techniques has been successfully used to address various modeling issues [e.g., 19, 20]. The rest of this paper is organized as follows. Model specification is presented in the next section. An EM algorithm is developed in section 3 to estimate the model. In section 4, simulation studies are used to investigate the computational properties of this EM algorithm and the modeling ability of our model as compared with other approaches that are available now. A real data example is used to illustrate the practical usefulness of this modeling technique. The paper concludes with some discussions.

2. Model Specification

Suppose there are n subjects, and each subject has J responses, $y_{it} = (y_{it1}, \dots, y_{itJ})$, at time t , $t = 1, \dots, t_i$, where t_i is the number of occasions for subject i . Let's also assume that there are d latent factors, $\eta_{it} = (\eta_{it1}^1, \dots, \eta_{itd}^d)$, underlying these J observed responses and d is much less

than J . A factor analysis model is used to represent the relationship between the observed and latent variables,

$$y_{it} = \Lambda \eta_{it} + \epsilon_{it}, \quad (1)$$

where Λ is the factor loading matrix, $\eta_{it} = (\eta_{it}^1, \dots, \eta_{it}^d)$ and $\epsilon_{it} = (\epsilon_{it1}, \dots, \epsilon_{itJ})$ are latent factors and unique factors for subject i at time t . We assume that $\epsilon_{itj} \sim N(0, \tau_j^2)$, and $\epsilon_{itj} \perp \epsilon_{ith}, j \neq h$. This assumption is often referred as the conditional independence assumption that $y_{it1}, \dots, y_{itJ}$ are independent conditional on latent factors η_{it} . This assumption also implies that the correlations among responses at any given time point are caused entirely by the latent factors they share. In addition, this model also assumes that the latent structure of these observed responses, represented by Λ , is the same over time. This assumption is adopted here to ensure that latent factors at different times are on the same basis and comparable. Otherwise, latent factor differences across time can be attributed to the fact that latent structure has changed over time rather than longitudinal development. This assumption is also needed in order to identify the model. Furthermore, the scale of the latent factors and the rotation of the factor loading matrix also need to be fixed by introducing more restrictions. The scale of latent factors can be fixed by restricting either the variance of latent factors, or one element in each column of Λ to be a constant, which is usually 1. To fix the rotation of the factor loading matrix, we need to apply enough restrictions to the factor loading matrix. For confirmatory analysis, we assume that the user has some hypothesis about the latent structure of the responses, for example simple structure, and this hypothesis will introduce enough restrictions to fix the rotation. For exploratory analysis, usually we set the upper diagonal part of Λ to be 0, that is $\lambda_{ij} = 0, j > i$. For more information about factor analysis, such as model identification, please refer to Basilevsky [21].

Next we use a multivariate linear mixed model to study the longitudinal development of these latent factors. This can be expressed by the following equation

$$\eta_{it}^m = x_{it}^m \beta^m + z_{it}^m b_i^m + e_{it}^m, m=1, \dots, d \quad (2)$$

where x_{it}^m and z_{it}^m are p and q elements covariates corresponding to fixed and random effects respectively to factor m , $b_i^m = (b_{i1}^m, \dots, b_{iq}^m)$ and $e_i^m = (e_{i1}^m, \dots, e_{it_i}^m)$, $m = 1, \dots, d$, are the random effects and error for subject i and factor m . $b_i = (b_i^1, \dots, b_i^d)$ is a vector of all the random effects for subject i . The random effects are assumed to be normally distributed with mean 0 and variance covariance matrix Σ_b . Since the number of latent factors is much smaller than the number of items, the number of random effects needed here is much less than directly applying the multivariate linear mixed model to these items. In this model, covariates can only be included for the multivariate linear mixed model rather than the factor analysis model. Besides the time variable, covariates, both fixed and random, can be specified and different latent factors can have partially or totally different covariates. Thus the model can be used to study the effects of different covariates for different latent factors simultaneously. In practice, substantive research is often interested in the association of changes of different qualities and often these qualities cannot be observed directly. For

example, in mental health studies of depression and anxiety, research interest may focus on how the longitudinal trends of depression and anxiety are correlated with each other. In child development, an interesting question is how the development of different abilities are associated with each other. Such research questions cannot be addressed if outcomes or latent factors are modeled separately. Under our modeling framework, they are characterized by parameters in Σ_b , the variance and covariance matrix of random effects. For example, a positive $\Sigma_b[1, 2]$ suggests that if latent factor $\tilde{\eta}^1$ increases over time for subject i , it is very reasonable to expect that $\tilde{\eta}^2$ will also increase over time for the same subject.

3. Model Estimation with EM algorithm

In this section we discuss model estimation using the EM algorithm [22]. An advantage of using the EM algorithm for this model is that it is easy to implement, because closed form solutions are available for both E and M steps. To derive the EM algorithms for this model, the model is re-expressed in the following vectorized fashion

$$y_i = (I_{t_i} \otimes \Lambda) \eta_i + \epsilon_i, \quad (3)$$

$$\eta_i = x_i \beta + z_i b_i + e_i. \quad (4)$$

For this model setup, we have the following notation:

$$\begin{aligned} y_i &= (y_{i1}^T, \dots, y_{it_i}^T)^T, \\ \eta_i &= (\eta_{i1}^T, \dots, \eta_{it_i}^T)^T, \\ \Lambda_{[J \times d]} &= \begin{pmatrix} \lambda_1^T \\ \vdots \\ \lambda_J^T \end{pmatrix}, \\ x_i &= \begin{pmatrix} x_{i1} \\ \vdots \\ x_{it_i} \end{pmatrix}_{[d \times ti, p \times d]}, \quad x_{it} = \begin{pmatrix} x_{it1} & 0 & \dots & 0 \\ 0 & x_{it2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & x_{itd} \end{pmatrix}_{[d, p \times d]}, \\ z_i &= \begin{pmatrix} z_{i1} \\ \vdots \\ z_{it_i} \end{pmatrix}_{[d \times ti, q \times d]}, \quad z_{it} = \begin{pmatrix} z_{it1} & 0 & \dots & 0 \\ 0 & z_{it2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & z_{itd} \end{pmatrix}_{[d, q \times d]}, \\ \beta &= (\beta_1^T, \dots, \beta_d^T)^T_{[p \times d, 1]}, \\ e_i &= (e_{i1}^T, \dots, e_{it_i}^T)^T, \quad e_{it_1} \perp e_{it_2}, t_1 \neq t_2, e_{it} \sim N(0, \Sigma_e), \\ b_i &= (b_{i1}^T, \dots, b_{id}^T)^T_{q \times d, 1} \sim N(0, \Sigma_b). \end{aligned}$$

The complete data include $\{x, y, z, v\}$, and the complete data log likelihood is

$$\begin{aligned}
\log L &= \sum_{i=1}^n [\log P(y_i|\eta_i, \theta) + \log P(\eta_i|x_i, z_i, b_i, \theta) + \log P(b_i|\theta)] \\
&= \sum_{i=1}^n \left[\sum_{j=1}^J \log P(y_{ij}|\eta_i, \theta) + \sum_{t=1}^{t_i} \log P(\eta_{it}|x_i, z_i, b_i, \theta) + \log P(b_i|\theta) \right] \\
&= \sum_{i=1}^n \sum_{j=1}^J \left\{ -\frac{t_i}{2} \log \tau_j^2 - \frac{1}{2\tau_j^2} (y_{ij} - \eta_i^T \lambda_j)^T (y_{ij} - \eta_i^T \lambda_j) \right\} \\
&\quad + \sum_{i=1}^n \sum_{t=1}^{t_i} \left(-\frac{1}{2} \log |\Sigma_e| - \frac{1}{2} (\eta_{it} - x_{it}\beta - z_{it}b_i)^T \Sigma_e^{-1} (\eta_{it} - x_{it}\beta - z_{it}b_i) \right) \\
&\quad + -\frac{1}{2} \sum_{i=1}^n \left[\log |\Sigma_b| - \frac{1}{2} b_i^T \Sigma_b^{-1} b_i \right].
\end{aligned} \tag{5}$$

After maximizing the above complete data log likelihood with respect to the model parameters, we get the following estimating equations

$$\hat{\Lambda}^T = \left[\sum_{i=1}^n \sum_{t=1}^{t_i} \eta_{it} \eta_{it}^T \right]^{-1} \left[\sum_{i=1}^n \sum_{t=1}^{t_i} \eta_{it} y_{it}^T \right], \tag{6}$$

$$\begin{aligned}
\tau_j^2 &= \frac{1}{N} \text{diag} \left[\sum_{i=1}^n \sum_{t=1}^{t_i} \left[(y_{it} - \hat{\Lambda} \eta_{it}) (y_{it} - \hat{\Lambda} \eta_{it})^T \right] \right] \\
&= \frac{1}{N} \text{diag} \left[\sum_{i=1}^n \sum_{t=1}^{t_i} \left[y_{it} y_{it}^T - 2 y_{it} \eta_{it}^T \hat{\Lambda}^T + \hat{\Lambda} \eta_{it} \eta_{it}^T \hat{\Lambda}^T \right] \right] \\
&= \frac{1}{N} \text{diag} \left[\sum_{i=1}^n \sum_{t=1}^{t_i} y_{it} y_{it}^T - 2 \sum_{i=1}^n \sum_{t=1}^{t_i} y_{it} \eta_{it}^T \hat{\Lambda}^T + \hat{\Lambda} \sum_{i=1}^n \sum_{t=1}^{t_i} \eta_{it} \eta_{it}^T \hat{\Lambda}^T \right],
\end{aligned} \tag{7}$$

$$\hat{\Sigma}_b = \frac{1}{n} \sum_{i=1}^n b_i b_i^T. \tag{8}$$

Estimation of β and Σ_e can be obtained iteratively. First for a fixed Σ_e , we have

$$\hat{\beta} = \left[\sum_{i=1}^n \sum_{t=1}^{t_i} x_{it}^T \Sigma_e^{-1} x_{it} \right]^{-1} \left[\sum_{i=1}^n \sum_{t=1}^{t_i} x_{it}^T \Sigma_e^{-1} (\eta_{it} - z_{it} b_i) \right], \tag{9}$$

then for fixed β , we have

$$\begin{aligned}
\hat{\Sigma}_e &= \frac{1}{N} \sum_{i=1}^n \sum_{t=1}^{t_i} \left[\eta_{it} - x_{it}\hat{\beta} - z_{it}b_i \right] \left[\eta_{it} - x_{it}\hat{\beta} - z_{it}b_i \right]^T \\
&= \frac{1}{N} \sum_{i=1}^n \sum_{t=1}^{t_i} \left[\eta_{it} \eta_{it}^T - \eta_{it} \hat{\beta}^T x_{it}^T - \hat{\beta} x_{it} \eta_{it}^T - \eta_{it} b_i^T z_{it}^T - b_i z_{it} \eta_{it}^T \right. \\
&\quad \left. + x_{it} \hat{\beta} b_i^T z_{it}^T + \hat{\beta} b_i z_{it} x_{it}^T + x_{it} \hat{\beta} \hat{\beta}^T x_{it}^T + z_{it} b_i b_i^T z_{it}^T \right].
\end{aligned} \tag{10}$$

The above equations indicate that in the E step we only need to calculate the conditional expectation of the following statistics: η_{it} , $\eta_{it} \eta_{it}^T$, b_i , $b_i b_i^T$, $\eta_{it} b_i^T$. These conditional expectations and corresponding derivations are provided in the appendix. To identify the factor analysis model, we need to fix some values in Λ . The approach proposed by Shi and

Lee [23] is adopted here to estimate free parameters in Λ conditional on these fixed value constraints. To summarize, the EM algorithm includes the following two steps.

E step: calculate the conditional expectations of v_{it} , $\eta_{it}\eta_{it}^T$, b_i , $b_ib_i^T$, and $\eta_{it}b_i^T$ using equations (24) to (28) in the Appendix.

M step: update parameter estimates using equations (6) to (10).

These two steps run iteratively until the algorithm is converged. The convergence of the EM algorithm is determined by monitoring the largest parameter change after each iteration. In the following simulation studies and real data example, the EM algorithm is considered as converged if the largest parameter change is smaller than a prespecified value.

4. Simulation studies

In this section, simulation studies are used to investigate the computational properties of the EM algorithm and compare the model proposed here with two existing approaches under different settings. The first approach include two steps. It first summarizes observations at each time point with the averages of the items corresponding to each factor and then analyzes these average scores at each time point with the multivariate linear mixed model using PROC MIXED in SAS. In the following studies, a bi-variate linear mixed model is used to illustrate the modeling process. For more information about how to fit a bi-variate linear mixed model with PROC MIXED, please refer to Thiébaud et al. [24]. The second approach is based on SEM and the model can be fitted by popular SEM packages, such as PROC CALIS in SAS and EQS. This approach is also called the latent curve model and more information about this approach can be found at Bollen et al. [13]. To evaluate the accuracy of parameter estimates, we repeat the simulation 30 times and obtained 30 sets of estimates. The bias of parameter estimates, which is the absolute difference between true value and the mean of 30 estimates, are calculated as a measure of performance.

In the first study, we want to investigate the case when the data is well balanced and factor loadings are the same. This study includes 1000 subjects with 6 waves. Responses are simulated based on a mixed model with random intercept and slope. We assume that the current research focuses on two attributes that cannot be observed directly but instead are measured by 12 items. A simple structure is adopted to model the relationship between these items and the latent factors, which suggests that each item loads on only one of the latent factors. Here we assume that the first 6 items load on the first factor and the remaining 6 items load on the second factor. The simple structure suggests that the last 6 elements of the first column and the first 6 elements of the second column of the factor loading matrix will be fixed at 0, thus not allowing any factor rotation. To fix the scale of latent factors, $\lambda_{1,1}$ and $\lambda_{12,2}$ are fixed to be 1. The true values of the free parameters in the model are presented in Table 1.

The LFLMM model is estimated by the EM algorithm with the following initial values: (1) 0.5 for all the free parameters in the factor loading matrix; (2) 0 for all the fixed effects; (3) identity matrix for the variance and covariance matrix of random effects and random errors

in the mixed model; (4) 1 for the variance of specific factors in factor model. Initial values for the other two approaches are set by the program.

Results are presented in Table 1. The estimation bias for the LFLMM model ranges from 0 to 0.0717, and the average is 0.0111. Overall, the parameters of the measurement model (factor loadings, error variances) are estimated more precisely (less bias, smaller variances) than those of the latent mixed regression model. These results suggest that the EM algorithm can recover the model parameters very well. Because responses at each time point are summarized by their means, factor loadings and residual variances corresponding to the factor model are not applicable to the two-step approach. Estimates for these parameters are replaced by * in Table 1. Estimates for these parameters that are applicable to the two-step approach are very close to the true value. The reason is mainly because factor loadings are the same so a simple average can correctly represent the contribution of each item. Also because the data are well balanced, SEM is also efficient in modeling these data.

In practice, however, it is not uncommon that we may have very limited information about the latent structure. For example, the number of latent factors may be known but the structure of the factor loading matrix may be unknown. In this situation, the latent factor structure needs to be investigated along with the modeling process. Here we would like to explore the possibility of solving the problem with the LFLMM model. First we start with some exploratory analysis that tries to identify the structure of the factor loading matrix. Then it is followed by a confirmatory analysis to finalize the results.

During the exploratory analysis, the factor rotation is fixed by letting the upper diagonal elements of the factor loading matrix to be 0, and the factor scale is fixed by restricting one element in each column of the loading matrix be 1. While these restrictions will identify the model, latent factors defined by this solution may be difficult to interpret. To overcome this ambiguity, rotations can be applied to the factor loading matrix such that each item is affected by a small number of factors, preferably only one. It is recommended to try different rotation methods and choose the one producing the simplest and most interpretable structure if their results are not consistent. To illustrate the idea, the initial factor loading matrix, the one after exploratory analysis, and the final one after promax rotation are shown in Table 2. If we compare these three matrices, we can observe that after rotation, the structure becomes much simpler and the latent factors are much easier to interpret. The factor loading matrix after rotation suggests a simple structure in which the first factor is defined by items 1 to 6 and the second factor is defined by items 7 to 12. This simple structure can then be tested by a confirmatory analysis afterwards. Results from this confirmatory analysis are very close to those presented in Table 1. These results suggest that the LFLMM model also should work very well even when the latent structure is unknown.

In the first part of the previous study, the latent factor structure is known, the responses are well balanced and the factor loadings are the same. As a result, the SEM approach and the two-step approach can fit the model equally well as the LFLMM model. In the following, we will introduce another study where responses are unbalanced and factor loadings are different from each other. In this example, there are 1000 subjects and the number of measurement occasions range from 4 to 10 for different subjects. There are 12 items and the

same latent factor structure as in the first study is used. Factor loadings range from 0.9 to 2.0.

Because of the unbalance nature of the data, SEM needs to be fitted with the full information maximum likelihood method since a large fraction of missing data is introduced to represent the unbalanced structure of the data. On the other hand, we can anticipate that results generated by the two-step approach will become biased because simple averages will not be able to reflect the fact that different items now contribute differently to the latent factors. For this study, because the full information maximum likelihood estimation method is very slow, we only used one simulated data. Results for this study are summarized in Table 3.

The results in Table 3 suggest that (1) with unbalanced data, SEM can recover the true parameter equally well as compared with LFLMM under the cost of much longer computation time; (2) two-step approach will almost always produce biased results since factor loadings are rarely the same in practice.

Results from this study show that the full information based SEM can model unbalanced longitudinal data well. However, under the extreme case where all subjects are measured at different time points, SEM will be very difficult or even impossible to use. In practice, it is often useful to include other random covariate besides time into the model. While this can be easily accomplished under the LFLMM framework, it will be a very difficult task for SEM. Simulation studies are not used here to investigate these cases because SEM is not applicable and as a result meaningful comparison cannot be made.

5. Real Data Example

The real data we used to illustrate the model come from the MacArthur Study of Successful Aging [25], a longitudinal study of relatively high functioning men and women aged 70 to 79 years. Subjects were assessed three times in 1988, 1991 and 1995 for physical and cognitive status by performance as well as by self-report measures. One of the interesting questions raised by geriatrics researchers is how physical and cognitive functioning develop over time, and, more importantly, whether there is any association between the development of these two. Three timing variables, seconds for foot taps, turning a circle and bending over, are selected to measure physical functioning. Cognitive functioning are measured by three tests: spatial memory test, item recall test and copying test.

The design of these items suggests a simple structure, with the first 3 items loading on the first factor and the last 3 items loading on the second factor. Thus a confirmatory latent factor linear mixed model with random intercept and slope is adopted. The EM algorithm is approximately converged after 500 iterations. The estimated factor loadings range from 0.46 to 2.60. The variances of unique factors range from 0.89 to 4.21. These results suggest that the two latent factors, physical and cognitive functioning, are well defined by the corresponding items.

The two main research questions are (1) how physical and cognitive functioning develop over time; and (2) whether there is any association between these two developments. They can be answered by the fixed slope and the correlation matrix of the random effects.

The slopes for physical and cognitive functioning are -0.39 and -1.16 respectively. These results suggest that both physical and cognitive functioning decrease over time with cognitive functioning decreasing faster than physical functioning. The correlation matrix of random effects is shown in Table 4. F1 and F2 correspond to physical and cognitive functioning respectively. b_{11} and b_{12} are the random intercept and random slope for F1. b_{21} and b_{22} are the random intercept and random slope for F2. These correlations suggest several interesting findings. First, the high positive correlation between b_{12} and b_{22} , 0.92 , suggest that the development of physical and cognitive functioning are highly related. That is, if one subject's physical functioning decreases over time, then it is reasonable to expect that his/her cognitive functioning would also decrease over time. Second, the positive correlation between b_{11} and b_{21} , 0.40 , suggests that those who start with better physical functioning tend to start with better cognitive functioning.

6. Discussion

In this paper we proposed a latent factor linear mixed model for the analysis of high dimensional longitudinal data. An EM algorithm is developed to estimate the model parameters. Results from simulation studies suggest that the EM algorithm can recover the model parameters very well. The real data example also demonstrates that the model proposed here can be very useful in applied research. Compared with existing approaches, the LFLMM model can handle high dimensional unbalanced longitudinal data naturally and can be extended easily to incorporate other covariates and handle other responses types. While results from the simulation section suggest the SEM approach with full information maximum likelihood estimation can also handle certain unbalanced data, there are several cases where SEM will be inefficient: (1) subjects are all measured at different time points, and (2) variables other than time need to added as random covariates.

We have to note that the model has some restrictive assumptions and the corresponding EM algorithm may encounter some disadvantages. We will briefly discuss these issues and propose some possible solutions here. First, it is assumed that all the responses are continuous. In practice, categorical or mixed type items are also used to define latent factors. One of the widely used solutions to handle categorical responses is the latent response approach [e.g., 26, 27] which assumes a latent response underlying these categorical responses. This implies extending our approach to adding another level of response transformation. However, the corresponding EM algorithm could be computationally intensive, especially when the number of latent factors is large, because the E step then cannot be solved analytically, Instead it needs to be approximated with either numerical or Monte Carlo quadrature. Another important assumption is that the number of latent factors is known. One approach to release this assumption could be to build several models with different numbers of latent factors and to compare them with regard to fit indexes and interpretability of results.

The EM algorithm also has two drawbacks that are worthy of further research. First, the convergence of the EM algorithm could be very slow under certain conditions. This problem has been observed previously and acceleration techniques have been proposed to deal with this issue [28]. Examples include the ECM [29] and parameter expansion EM [30] methods.

Another problem associated with the EM algorithm is that standard errors cannot be obtained as an automatic by-product. Several approaches, such as supplemented EM [SEM, 31], have been proposed to get standard errors with EM algorithms. These issues are beyond the scope of the current paper. For more information about EM algorithms and corresponding techniques please refer to McLachlan and Krishnan [32].

A. Conditional expectations

To derive the conditional expectations for these statistics, we get the joint distribution of y_i , v_i , b_i

$$\begin{pmatrix} y_i \\ \eta_i \\ b_i \end{pmatrix} \sim N \left(\begin{pmatrix} I_{t_i} \otimes \Lambda x_i \beta \\ x_i \beta \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{y_i} & \Sigma_{y_i \eta_i} & \Sigma_{y_i b_i} \\ \Sigma_{\eta_i y_i} & \Sigma_{\eta_i} & \Sigma_{\eta_i b_i} \\ \Sigma_{b_i y_i} & \Sigma_{b_i \eta_i} & \Sigma_{b_i} \end{pmatrix} \right), \quad (11)$$

where

$$\Sigma_{y_i} = (I_{t_i} \otimes \Lambda) \Sigma_{\eta_i} (I_{t_i} \otimes \Lambda)^T + I_{t_i} \otimes \text{diag}(\tau_1^2, \dots, \tau_J^2), \quad (12)$$

$$\Sigma_{\eta_i} = z_i \Sigma_b (z_i)^T + I_{t_i} \otimes \Sigma_e, \quad (13)$$

$$\Sigma_{b_i} = \Sigma_b, \quad (14)$$

$$\Sigma_{y_i \eta_i} = (I_{t_i} \otimes \Lambda) \Sigma_{\eta_i}, \quad (15)$$

$$\Sigma_{y_i b_i} = (I_{t_i} \otimes \Lambda) (z_i \Sigma_b), \quad (16)$$

and

$$\Sigma_{\eta_i b_i} = z_i \Sigma_b. \quad (17)$$

Thus the conditional distributions of v_i and b_i are normal distributions with the following mean and variance covariance matrices.

$$\mu_{\eta_i | y_i} = x_i \beta + \Sigma_{\eta_i y_i} \Sigma_{y_i}^{-1} (y_i - I_{t_i} \otimes \Lambda x_i \beta), \quad (18)$$

$$\Sigma_{\eta_i | y_i} = \Sigma_{\eta_i} - \Sigma_{\eta_i y_i} \Sigma_{y_i}^{-1} \Sigma_{y_i \eta_i}, \quad (19)$$

$$\mu_{b_i | y_i} = \Sigma_{b_i y_i} \Sigma_{y_i}^{-1} (y_i - I_{t_i} \otimes \Lambda x_i \beta), \quad (20)$$

and

$$\Sigma_{b_i|y_i} = \Sigma_{b_i} - \Sigma_{b_i y_i} \Sigma_{y_i}^{-1} \Sigma_{y_i b_i}. \quad (21)$$

The joint conditional distribution of v_i and b_i also follows a normal distribution with the following variance and covariance matrix.

$$\Sigma_{\eta_i, b_i|y_i} = \begin{pmatrix} \Sigma_{\eta_i|y_i} & \Sigma_{\eta_i b_i|y_i} \\ \Sigma_{b_i \eta_i|y_i} & \Sigma_{b_i|y_i} \end{pmatrix} = \begin{pmatrix} \Sigma_{\eta_i} & \Sigma_{\eta_i b_i} \\ \Sigma_{b_i \eta_i} & \Sigma_{b_i} \end{pmatrix} - \begin{pmatrix} \Sigma_{\eta_i y_i} \\ \Sigma_{b_i y_i} \end{pmatrix} \Sigma_{y_i}^{-1} (\Sigma_{y_i \eta_i}, \Sigma_{y_i b_i}), \quad (22)$$

and the conditional covariance matrix of v_i and b_i is

$$\Sigma_{\eta_i b_i|y_i} = \Sigma_{\eta_i b_i} - \Sigma_{\eta_i y_i} \Sigma_{y_i}^{-1} \Sigma_{y_i b_i}. \quad (23)$$

Thus we have

$$E(\eta_{it}|y_i) = \mu_{\eta_i|y_i} [1 + (t-1) * d : t * d], \quad (24)$$

$$E(\eta_{it} \eta_{it}^T | y_i) = E(\eta_{it}|y_i) E(\eta_{it}|y_i)^T + \Sigma_{\eta_i|y_i} [1 + (t-1) * d : t * d, 1 + (t-1) * d : t * d], \quad (25)$$

$$E(b_i|y_i) = \mu_{b_i|y_i}, \quad (26)$$

$$E(b_i b_i^T | y_i) = E(b_i|y_i) E(b_i|y_i)^T + \Sigma_{b_i|y_i}, \quad (27)$$

and

$$E(\eta_{it} b_i^T) = E(\eta_{it}|y_i) E(b_i|y_i)^T + \Sigma_{\eta_i b_i|y_i} [1 + (t-1) * d : t * d, :]. \quad (28)$$

References

1. Hedeker, D.; Gibbons, RD. Longitudinal Data Analysis. John Wiley & Sons; 2006.
2. Wu, L. Mixed Effects Models for Complex Data. CRC Press; Boca Raton: 2010.
3. Fieuws, S.; Verbeke, G. Joint models for high-dimensional longitudinal data. Chapman & Hall/CRC; 2009. chap. 16; p. 367-394.
4. Verbeke, G.; Davidian, M. Joint models for longitudinal data: Introduction and overview. In: Fitzmaurice, G.; Davidian, M.; Verbeke, G.; Molenberghs, G., editors. Longitudinal Data Analysis. CRC Press; Boca Raton: 2009.
5. Galecki A. General class of covariance structures for two or more repeated factors in longitudinal data analysis. Communications in Statistics-Theory and Methods. 1994; 23(11):3105–3119.
6. Carey V, Rosner B. Analysis of longitudinally observed irregularly timed multivariate outcomes: regression with focus on cross-component correlation. Statistics in medicine. 2001; 20(1):21–31. [PubMed: 11135345]
7. Geys H, Molenberghs G, Ryan L. Pseudolikelihood modeling of multivariate outcomes in developmental toxicology. Journal of the American Statistical Association. 1999:734–745.

8. Reinsel G. Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure. *Journal of the American Statistical Association*. 1982;190–195.
9. Shah A, Laird N, Schoenfeld D. A random-effects model for multiple characteristics with possibly missing data. *Journal of the American Statistical Association*. 1997;775–779.
10. Gueorguieva RV, Agresti A. A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of The American Statistical Association*. 2001; 96(455):1102–1112.
11. Skrondal A, Laake P. Regression among factor scores. *Psychometrika*. 2001; 66(4):563–575.
12. Muthén B, Curran P. General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods*. 1997; 2(4):371.
13. Bollen, K.; Curran, P.; Wiley, J. *Latent curve models: A structural equation perspective*. Wiley-Interscience; 2006.
14. Voelkle MC, Oud JHL, Davidov E, Schmidt P. An SEM Approach to Continuous Time Modeling of Panel Data: Relating Authoritarianism and Anomia. *Psychological Methods*. 2012; 17:176–192. [PubMed: 22486576]
15. Dunson D. Dynamic latent trait models for multidimensional longitudinal data. *Journal of the American Statistical Association*. 2003; 98(463):555–563.
16. Cagnone S, Moustaki I, Vasdekis V. Latent variable models for multivariate longitudinal ordinal responses. *British Journal of Mathematical and Statistical Psychology*. 2009; 62(2):401–415. [PubMed: 18625083]
17. Roy J, Lin XH. Latent variable models for longitudinal data with multiple continuous outcomes. *Biometrics*. 2000; 56(4):1047–1054. [PubMed: 11129460]
18. Liu L, Hedeker D. A mixed-effects regression model for longitudinal multivariate ordinal data. *Biometrics*. 2006; 62(1):261–268. [PubMed: 16542254]
19. An X, Bentler P. Extended mixture factor analysis model with covariates for mixed binary and continuous responses. *Statistics in Medicine*. 2011; 30:2634–2647. [PubMed: 21786282]
20. Hand D, Crowder M. Measuring customer quality in retail banking. *Statistical Modelling*. 2005; 5(2):145–158.
21. Basilevsky, AT. *Statistical factor analysis and related methods*. Wiley; New York: 1994.
22. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B Statistical Methodology*. 1977; 39(1): 1–38.
23. Shi JQ, Lee SY. Latent variable models with mixed continuous and polytomous data. *Journal of the Royal Statistical Society, Series B*. 2000; 62:77–87.
24. Thiébaud R, Jacqmin-Gadda H, Chêne G, Leport C, Commenges D. Bivariate linear mixed models using sas proc mixed. *Computer methods and programs in biomedicine*. 2002; 69(3):249–256. [PubMed: 12204452]
25. Bruce M, Seeman T, Merrill S, Blazer D. The impact of depressive symptomatology on physical disability: Macarthur studies of successful aging. *American Journal of Public Health*. 1994; 84(11):1796. [PubMed: 7977920]
26. Hedeker D, Gibbons R. A random-effects ordinal regression model for multilevel analysis. *Biometrics*. 1994:933–944. [PubMed: 7787006]
27. Tutz G, Hennevogel W. Random effects in ordinal regression models. *Computational Statistics & Data Analysis*. 1996; 22(5):537–557.
28. van Dyk DA. Fitting mixed-effects models using efficient em-type algorithms. *Journal of Computational and Graphical Statistics*. 2000; 9(1):78–98.
29. Meng XL, Rubin DB. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*. 1993; 80(2):267–278.
30. Liu CH, Rubin DB, Wu YN. Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika*. 1998; 85(4):755–770.
31. Meng XL, Rubin DB. Using EM to Obtain asymptotic variance - covariance matrices - the SEM algorithm. *Journal of the American Statistical Association*. 1991; 86(416):899–909.
32. McLachlan, G.; Krishnan, T. *The EM Algorithm and Extensions*. John Wiley & Sons; 2007.

Table 1

Results for the first simulation study

Para	True	Bias			Para	True	Bias		
		LFLMM	SEM	TWOSTEP			LFLMM	SEM	TWOSTEP
$\lambda_{2,1}$	1.00	0.0000	0.0003	*	β_{11}	1.00	0.0227	0.0218	0.0210
$\lambda_{3,1}$	1.00	0.0004	0.0002	*	β_{21}	-1.00	0.0034	0.0029	0.0040
$\lambda_{4,1}$	1.00	0.0000	0.0002	*	β_{12}	0.00	0.0572	0.0496	0.0590
$\lambda_{5,1}$	1.00	0.0008	0.0009	*	β_{22}	0.00	0.0552	0.0510	0.0521
$\lambda_{6,1}$	1.00	0.0005	0.0006	*	β_{13}	1.00	0.0717	0.0875	0.0721
$\lambda_{7,2}$	1.00	0.0003	0.0002	*	β_{23}	1.00	0.0563	0.0445	0.0317
$\lambda_{8,2}$	1.00	0.0002	0.0002	*	β_{14}	1.00	0.0129	0.0279	0.0245
$\lambda_{9,2}$	1.00	0.0002	0.0005	*	β_{24}	-1.00	0.0055	0.0087	0.0079
$\lambda_{10,2}$	1.00	0.0004	0.0002	*	$\sigma_{b,11}$	3.00	0.0318	0.0238	0.0261
$\lambda_{11,2}$	1.00	0.0007	0.0003	*	$\sigma_{b,12}$	1.00	0.0008	0.0007	0.0008
τ_1^2	0.50	0.0019	0.0020	*	$\sigma_{b,13}$	1.50	0.0097	0.0082	0.0105
τ_2^2	0.50	0.0041	0.0039	*	$\sigma_{b,14}$	1.00	0.0006	0.0005	0.0007
τ_3^2	0.50	0.0015	0.0015	*	$\sigma_{b,22}$	3.00	0.0169	0.0140	0.0210
τ_4^2	0.50	0.0000	0.0008	*	$\sigma_{b,23}$	1.00	0.0165	0.0139	0.0153
τ_5^2	0.50	0.0019	0.0013	*	$\sigma_{b,24}$	2.00	0.0260	0.0199	0.0309
τ_6^2	0.50	0.0021	0.0023	*	$\sigma_{b,33}$	3.00	0.0214	0.0150	0.0283
τ_7^2	0.50	0.0004	0.0003	*	$\sigma_{b,34}$	1.00	0.0091	0.0112	0.0109
τ_8^2	0.50	0.0032	0.0030	*	$\sigma_{b,44}$	3.00	0.0414	0.0443	0.0509
τ_9^2	0.50	0.0004	0.0004	*	$\sigma_{e,11}$	0.50	0.0036	0.0042	0.0029
τ_{10}^2	0.50	0.0019	0.0021	*	$\sigma_{e,12}$	0.20	0.0022	0.0018	0.0026
τ_{11}^2	0.50	0.0029	0.0018	*	$\sigma_{e,22}$	0.50	0.0009	0.0011	0.0015
τ_{12}^2	0.50	0.0006	0.0008	*					

Table 2

Three different factor loading matrix structures. The first one in column 2 and 3 is the initial loading matrix. The second one in column 4 and 5 is the factor loading matrix after the exploratory analysis. The last one in column 6 and 7 is obtained by applying the promax rotation on the second one.

Items	<u>Initial</u>		<u>Before Rotation</u>		<u>After Rotation</u>	
	F1	F2	F1	F2	F1	F2
1	0.50	0.00	1.000	0.000	1.002	0.000
2	0.50	0.50	0.995	0.006	0.995	0.000
3	0.50	0.50	0.997	0.014	0.995	0.000
4	0.50	0.50	0.996	0.006	0.996	0.000
5	0.50	0.50	0.995	0.008	0.994	0.000
6	0.50	0.50	0.997	0.005	0.998	0.000
7	0.50	0.50	0.243	0.995	0.000	1.024
8	0.50	0.50	0.243	0.995	0.000	1.023
9	0.50	0.50	0.243	0.995	0.000	1.023
10	0.50	0.50	0.246	0.993	0.000	1.022
11	0.50	0.50	0.245	0.993	0.000	1.022
12	0.50	0.50	0.240	1.000	0.000	1.029

Table 3

Results for the second simulation study.

Para	True	Estimates			Para	True	Estimates		
		LFLMM	SEM	TWOSTEP			LFLMM	SEM	TWOSTEP
$\lambda_{2,1}$	1.50	1.5015	1.5007	*	$\sigma_{b,11}$	3.00	3.0219	2.9685	2.9811
$\lambda_{3,1}$	2.00	2.0020	1.9975	*	$\sigma_{b,12}$	1.00	0.8905	1.1899	2.1660
$\lambda_{4,2}$	1.00	0.9977	0.9985	*	$\sigma_{b,13}$	1.50	1.4526	1.4486	0.9939
$\lambda_{5,2}$	1.00	1.0000	0.9979	*	$\sigma_{b,14}$	1.00	0.9160	1.2082	1.4996
τ_1^2	0.50	0.4959	0.4892	*	$\sigma_{b,22}$	3.00	2.8547	2.9574	6.6740
τ_2^2	0.50	0.5073	0.5333	*	$\sigma_{b,23}$	1.00	0.8708	1.0008	1.8146
τ_3^2	0.50	0.5015	0.4871	*	$\sigma_{b,24}$	2.00	1.9473	1.9848	2.6771
τ_4^2	0.50	0.4807	0.4974	*	$\sigma_{b,33}$	3.00	2.7851	2.9850	3.0003
τ_5^2	0.50	0.5228	0.5041	*	$\sigma_{b,34}$	1.00	0.8826	0.9954	2.9749
τ_6^2	0.50	0.5128	0.5024	*	$\sigma_{b,44}$	3.00	2.9414	3.0046	6.6565
β_{11}	0.00	-0.0481	0.0485	0.0713	$\sigma_{e,11}$	0.50	0.5091	0.5072	0.6627
β_{21}	0.00	-0.0093	0.0141	0.0141	$\sigma_{e,12}$	0.00	0.0051	0.0053	0*
β_{12}	1.00	1.0546	1.0378	1.5573	$\sigma_{e,22}$	0.50	0.5041	0.4969	1.3008
β_{22}	-1.00	-1.0017	-0.8555	-0.8545					

* represent parameters that are not applicable to the two-step approach. $\sigma_{e,12}$ is also fixed at 0, indicated by 0*, for the two-step approach.

Table 4

Correlation matrix of random effects

Random effects	F1		F2	
	b_{11}	b_{12}	b_{21}	b_{22}
b_{11}	1.000	-0.513	0.399	-0.261
b_{12}	-0.513	1.000	-0.060	0.921
b_{21}	0.399	-0.060	1.000	-0.008
b_{22}	-0.261	0.921	-0.008	1.000