# Use of Pseudo-Likelihood Approach in Longitudinal Educational Surveys

**Zhi-Hui Fu[1,2],**
**[1]Department of Statistics, School of Mathematics and System Science,**
**Shenyang Normal University, Shenyang, Liaoning, China**
**[2]Corresponding author: Zhi-Hui Fu, e-mail:fuzhihui2001@163.com**

## Abstract

Multidimensional item response theory (MIRT) models can be applied to longitudinal educational surveys where a group of individuals are administered different tests over time. However, computational problems typically arise as the dimension of the latent variables increases. This is especially true when the latent variable distribution cannot be integrated out analytically, as with MIRT models for the data collected from the mixed-type tests, which composed of both dichotomous and polytomous items. Based on the pseudo-likelihood theory, this presentation will describe a pairwise modeling strategy to estimate item and population parameters in longitudinal studies. The pairwise method effectively reduces the dimensionality of the problem and hence is applicable to longitudinal IRT data with high-dimensional latent variables, which are challenging for classical methods. At last, we proposed the Pairwise EM (PEM) algorithm, and proved this algorithm has the ascent property similar to the EM algorithm in that the pairwise likelihood is nondecreasing.

**Keywords**: Expectation-maximization algorithm, mixed-type test, multidimensional item response theory, pairwise modeling,

### 1. Introduction

In educational and psychological research, it is often of interest to investigate changes over time for a group of subjects. For example, the No Child Left Behind Act of 2001 promoted great interest in measuring student progress or growth across time. Many item response theory (IRT)-based methods have been developed to model such data. Roughly, they can be classified into three types, separate calibration linking methods, the fixed common item parameter calibration and equating method, and the concurrent calibration method.

However, all the aforementioned methods do not take into consideration of the dependencies in the item responses over time, which is essential to longitudinal studies. Adam et al. (1997) and Wang & Wu (2004) developed some multidimensional Rasch-type item response models that model the dependency in longitudinal measurements. Adam et al. (1997) used between-item MIRT models, in which a set of items is divided into subsets where each subset is described by a unidimensional IRT model. The latent variables measured by different scales are assumed to be correlated. Fu et al.(2011) proposed a pairwised fitting method to analyze the longitudinal data for the 3PLM. In this paper, with some minor adjustments, we applied this method to analyze longitudinal data in mixed-type tests. Change over time is modeled by assuming that the latent ability $\boldsymbol{\theta}$ at different times points follows a multivariate normal distribution $MVN(\boldsymbol{\mu},\boldsymbol{\Sigma})$, where the covariance matrix $\boldsymbol{\Sigma}$ accounts for the dependence between the item responses over time points. The change is estimated along with the item parameters using our pairwise likelihood method.

There are several alternatives to the pairwise likelihood approach. For instance, Embretson (1991) proposed a two-step procedure for the Rasch model. For general cases, Fischer and Ponocny (1994) modeled the change by imposing linear restrictions on item parameters. More recently, Andrade and Tavares (2005) proposed a marginal maximum likelihood (MML) estimation procedure (referred as the joint modeling method hereafter) to take into account the dependence of the dichotomous responses at different time points. However, the MML approach requires the computation of a multiple integral whose order is the same as the dimension of the latent variables. As such integrals are generally solved by numerical methods, the number of latent variables that can be analyzed simultaneously is hence limited.

Compared to the joint modelling approach, our pairwise likelihood approach replaces the high-dimensional integral over all latent variables by a set of double integrals over each pair of latent variables. In the literature of linear mixed effect models, this pairwise likelihood idea was shown by Fieuws and Verbeke (2006) to effectively solve the dimensionality problem when integrating over latent variables. For a more thorough discussion of pairwise likelihood and its other variants, we refer readers to Lindsay (1988) and Cox and Reid (2004).

## 2. Joint modeling

We assume a group of $N$ individuals, randomly selected from a population, is evaluated at $T$ pre-specified instants. For instance, one group of students evaluated at the end of the $4th$ to $8th$ grades. At time $t$, $t = 1, 2, \cdots, T$, the group of individuals are administered a mixed-type test that consists of $n_t$ items in which $m_t$ are dichotomous and $n_t - m_t$ are polytomous, and we assume that the 3PL and GPCM (Muraki, 1992) models fit the data well. Usually, the total number of distinctive items $n$, is less than $n_c = \sum_{t=1}^{T} n_t$ because of common items among the tests. At each time point $t$, the probability of the correct response on dichotomously scored item $i$ is defined by

$$\begin{aligned} P_{jit} &= P(U_{jit} = 1|\theta_{jt}, a_i, b_i, c_i) \\ &= c_i + (1 - c_i)\{1 + \exp[-Da_i(\theta_{jt} - b_i)]\}^{-1}, \end{aligned}$$

where $a_i$ is the item discrimination parameter, $b_i$ is the item difficulty parameter, and $c_i$ is the parameter that represents the probability of examinees with low ability correctly answering item $i$. As to the polytomous items, the probability of selecting response $k$ (where $k = 0, 1, \cdots, K_i$) of polytomous item $i$ is

$$\begin{aligned} P_{jitk} &= P(U_{jit} = k|\theta_{jt}, \alpha_i, \beta_i) \\ &= \exp[\sum_{v=0}^{k} -D\alpha_i(\theta_{jt} - \beta_{iv})]\{\sum_{v=0}^{K_i} \exp[\sum_{q=0}^{v} -D\alpha_i(\theta_{jt} - \beta_{iq})]\}^{-1}, \end{aligned}$$

where $\alpha_i$ is the discrimination parameter of item $i$ and $\beta_{iv}$ is the location parameter of category $v$. Here $i = 1, 2, \cdots, n_t$, of individual $j$, $j = 1, 2, \cdots, N$, in test $t$, $t = 1, 2, \cdots, T$, where $U_{jit}$ represents the response, $\theta_{jt}$ the latent ability. Assuming conditional independence of the responses to items in test $t$ given $\theta_t$ , and let $\boldsymbol{\xi}_i = (a_i, b_i, c_i, \alpha_i, \boldsymbol{\beta}_i)$ ,we have that

$$\begin{aligned} P(\vec{U}_{jt}|\theta_t, \xi) &= \prod_{i \in I_t} P(U_{jit}|\theta_t, \boldsymbol{\xi}_i) \\ &= [\prod_{i=1}^{m_t} P_{jit}^{U_{jit}}(1 - P_{jit})^{1-U_{jit}}][\prod_{i=m_t+1}^{n_t} \prod_{k=0}^{K_i} P_{jitk} I(U_{jit} = k)], \end{aligned}$$

where $I_t$ represents the set of the indexes of those items presented in test $t$, $\vec{U}_{jt} = (U_{j1t}, \cdots, U_{jn_tt})$ is the $(n_t \times 1)$ vector of responses of individual $j$ in test $t$, and $\boldsymbol{\xi} = (\boldsymbol{\xi}_1^{\mathtt{T}}, \boldsymbol{\xi}_2^{\mathtt{T}}, \cdots, \boldsymbol{\xi}_n^{\mathtt{T}})^{\mathtt{T}}$ (here and hereafter, the Roman superscript $\mathtt{T}$ means the transpose operation) the known vector of the items parameters. Note that, for convenience and without loss of generality, we dropped index $j$ from the ability parameter because we are interested in the distribution of the ability at each instant $t$ and not in any particular $\theta_{jt}$. Furthermore, assuming that item responses in $T$ tests are conditionally independent given the abilities in the $T$ tests, we have

$$P(\vec{U}_j|\boldsymbol{\theta}, \boldsymbol{\xi}) = \prod_{t=1}^{T} P(\vec{U}_{jt}|\theta_t, \boldsymbol{\xi}) = \prod_{t=1}^{T} \prod_{i \in I_t} P(U_{jit}|\theta_t, \boldsymbol{\xi}_i) \tag{1}$$

with $\vec{U}_j = (\vec{U}_{j1}^{\mathtt{T}}, \vec{U}_{j2}^{\mathtt{T}}, \cdots, \vec{U}_{jT}^{\mathtt{T}})$ being the $(n_c \times 1)$ vector of responses of individual $j$ in all tests and $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_T)^{\mathtt{T}}$. Finally, in this article, change over time is modeled by assuming that the latent ability parameters $\boldsymbol{\theta}$ have a multivariate normal distribution with $T$-dimensional vector of means $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_T)^{\mathtt{T}}$ and a $T \times T$ covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{T \times T}$. We denote this latent density function by $g(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. This assumption pertains to the dependence between a subject's latent ability at different time points. In the sequel, the covariance matrix $\boldsymbol{\Sigma}$ can be the covariance over time points for a specific scale. Let $\boldsymbol{\lambda} = (\boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and assuming independence among subjects, the marginal likelihood is

$$L(\boldsymbol{\lambda}; U) = \prod_{j=1}^{N} P(\vec{U}_j|\boldsymbol{\lambda}) = \prod_{j=1}^{N} \int_{R^T} P(\vec{U}_j|\boldsymbol{\theta}, \boldsymbol{\xi}) g(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathrm{d}\boldsymbol{\theta}. \tag{2}$$

Estimates of $\boldsymbol{\lambda}$ can then be obtained from maximizing the likelihood function (2), and inferences immediately follow from the classical maximum likelihood theory. The marginal maximum likelihood (MML) method is widely used in IRT. For a thorough discussion, please refer to Bock and Aitkin (1981), Thissen (1982)and many others. Andrade and Tavares (2005) illustrated in detail the MML method for linear mixed effect modelling of longitudinal data. However, in the longitudinal IRT context, the marginal distribution of $\vec{U}_j$ cannot be derived analytically.

## 3. Pairwise modeling

The computational complexity for MML encountered in longitudinal IRT is due to the dimensionality of the latent variable (i.e. the number of time points). We mitigate this problem by using the pairwise likelihood method(Lindsay, 1988). Instead of considering all $T$ tests simultaneously, the pairwise likelihood is given as the product of bivariate likelihoods over all pairs of two tests

$$PL(\boldsymbol{\lambda}; U) = \prod_{(r,s) \in \ \Gamma} P(\boldsymbol{\lambda}_{r,s}; \vec{U}_r, \vec{U}_s), \tag{3}$$

where $\Gamma$ is a non-empty subset of all possible pairwise neighbors, $\boldsymbol{\lambda}_{r,s}$ represents the vector of all parameters in the bivariate IRT models corresponding to the specific test pair $(r, s)$, $r = 1, \cdots, T-1,\ s = r+1, \cdots, T$. The bivariate likelihood for each pair of tests is given by

$$P(\boldsymbol{\lambda}_{r,s}; \vec{U}_r, \vec{U}_s) = \prod_{j=1}^{N} \int_{R^2} P(\vec{U}_{jr}|\theta_r, \boldsymbol{\xi}) P(\vec{U}_{js}|\theta_s, \boldsymbol{\xi}) g(\theta_r, \theta_s|\boldsymbol{\mu}_{r,s}, \boldsymbol{\Sigma}_{r,s}) \mathrm{d}\theta_r \mathrm{d}\theta_s, \tag{4}$$

where $g_{r,s}$ is the bivariate normal density function for ability vector $(\theta_r, \theta_s)^{\intercal}$ and with mean $\boldsymbol{\mu}_{r,s} = \begin{pmatrix} \mu_r \\ \mu_s \end{pmatrix}$ and covariance matrix $\boldsymbol{\Sigma}_{r,s} = \begin{pmatrix} \sigma_r^2 & \sigma_{r,s} \\ \sigma_{r,s} & \sigma_s^2 \end{pmatrix}$, and $\boldsymbol{\lambda}_{r,s} = (\boldsymbol{\xi}, \boldsymbol{\mu}_{r,s}, \boldsymbol{\Sigma}_{r,s})$. We propose to estimate all unknown parameters by maximizing the pairwise likelihood (3), and the dimensionality of the problem is reduced because only double integrals are involved. Following the pseudo-likelihood theory Geys et al.(1999), parameter estimates can be obtained using a two-step procedure. First, each bivariate likelihood $P\left(\boldsymbol{\lambda}_{r,s}; \vec{U}_r, \vec{U}_s\right)$ is maximized separately and an estimate of $\boldsymbol{\lambda}_{r,s}$ is obtained. Next, the final estimate of any parameter is obtained by taking averages over all pairs $P\left(\boldsymbol{\lambda}_{r,s}; \vec{U}_r, \vec{U}_s\right)$ that involves the parameter. Further, let $\boldsymbol{\Theta}$ be the stacked vector combining all pair-specific parameter vectors $\boldsymbol{\lambda}_{r,s}$. Then results from pseudo-likelihood theory can be used for inference for $\boldsymbol{\Theta}$. The asymptotic multivariate normal distribution for $\hat{\boldsymbol{\Theta}}$ is given by

$$\sqrt{N}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}) \sim MVN(\mathbf{0}, J^{-1}KJ^{-1})$$

where $J^{-1}KJ^{-1}$ is a 'sandwich-type' robust variance estimator.

## 4. An EM algorithm for maximizing pairwise likelihood

The EM algorithm is a method for function maximization which alternates an expectation step and a maximization step. It is widely used for likelihood inference in IRT (Bock & Aitkin, 1981; Thissen, 1982). In this section, we propose the following EM algorithm to maximize the pairwise likelihood $PL(\boldsymbol{\lambda}; U)$ in (3).

**PEM Algorithm.** Choose a starting value $\boldsymbol{\lambda}^{(0)}$ such that $PL(\boldsymbol{\lambda}^{(0)}; U) > 0$ and set $d = 0$. The pairwise EM (PEM) algorithm iterates the following steps until convergence.

- Expectation step: evaluate the sum of the conditional expectations

$$Q(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(d)}) = \sum_{(r,s)\in\Gamma} \int\int \log\{P(\vec{U}_r, \vec{U}_s, \theta_r, \theta_s; \boldsymbol{\lambda}_{r,s})\} P(\theta_r, \theta_s|\vec{U}_r, \vec{U}_s; \boldsymbol{\lambda}_{r,s}^{(d)}) \mathrm{d}\theta_r \mathrm{d}\theta_s.$$

- Maximization step: Solve $\max_{\boldsymbol{\lambda}} Q(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(d)})$ and let $\boldsymbol{\lambda}^{(d+1)}$ be the maximizer.
- Set $d = d + 1$.

The PEM algorithm has the following ascent property similar to the EM algorithm in that the pairwise likelihood is nondecreasing.

**Proposition 1** (Ascent property). Let $\boldsymbol{\lambda}^{(0)}, \boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}, \ldots$, be the iterative sequence of PEM, then the pairwise likelihood does not decrease at each iteration of the PEM algorithm, i.e., $PL(\boldsymbol{\lambda}^{(d)}; U) \leq PL(\boldsymbol{\lambda}^{(d+1)}; U)$ for all $d$.
**Proof.** Let $pl(\boldsymbol{\lambda}; U) = \log PL(\boldsymbol{\lambda}; U)$. We have

$$
\begin{aligned}
pl(\boldsymbol{\lambda};U) &= \sum_{(r,s)\in\Gamma} \log P(\vec{U}_r,\vec{U}_s;\boldsymbol{\lambda}_{r,s}) \\
&= \sum_{(r,s)\in\Gamma} \log P(\vec{U}_r,\vec{U}_s;\boldsymbol{\lambda}_{r,s}) \int\int g(\theta_r,\theta_s|\vec{U}_r,\vec{U}_s;\boldsymbol{\lambda}_{r,s}^{(d)})\mathrm{d}\theta_r\mathrm{d}\theta_s \\
&= \sum_{(r,s)\in\Gamma} \int\int \log\{P(\vec{U}_r,\vec{U}_s,\theta_r,\theta_s;\boldsymbol{\lambda}_{r,s})\} g(\theta_r,\theta_s|\vec{U}_r,\vec{U}_s;\boldsymbol{\lambda}_{r,s}^{(d)})\mathrm{d}\theta_r\mathrm{d}\theta_s \\
&\quad - \sum_{(r,s)\in\Gamma} \int\int \log\{P(\theta_r,\theta_s|\vec{U}_r,\vec{U}_s;\boldsymbol{\lambda}_{r,s})\} g(\theta_r,\theta_s|\vec{U}_r,\vec{U}_s;\boldsymbol{\lambda}_{r,s}^{(d)})\mathrm{d}\theta_r\mathrm{d}\theta_s \\
&= \sum_{(r,s)\in\Gamma} Q_{(r,s)}(\boldsymbol{\lambda}_{r,s}|\boldsymbol{\lambda}_{r,s}^{(d)}) - \sum_{(r,s)\in\Gamma} H_{(r,s)}(\boldsymbol{\lambda}_{r,s}|\boldsymbol{\lambda}_{r,s}^{(d)}) \\
&= Q(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(d)}) - H(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(d)}).
\end{aligned}
$$

Thus, the difference between log-pairwise likelihoods at the $d$th iteration and the $(d+1)$th iteration is

$$
pl(\boldsymbol{\lambda}^{(d+1)};U) - pl(\boldsymbol{\lambda}^{(d)};U) = Q(\boldsymbol{\lambda}^{(d+1)}|\boldsymbol{\lambda}^{(d)}) - Q(\boldsymbol{\lambda}^{(d)}|\boldsymbol{\lambda}^{(d)}) + \sum_{(r,s)\in\Gamma} D_{(r,s)}(\boldsymbol{\lambda}^{(d+1)}|\boldsymbol{\lambda}^{(d)}), \tag{5}
$$

where $D_{(r,s)}$ represents the Kullback-Leibler distance between the two bivariate densities
$g(\theta_r,\theta_s|\vec{U}_r,\vec{U}_s;\boldsymbol{\lambda}_{r,s}^{(d+1)})$ and $g(\theta_r,\theta_s|\vec{U}_r,\vec{U}_s;\boldsymbol{\lambda}_{r,s}^{(d)})$, i. e.,

$$
D_{(r,s)}(\boldsymbol{\lambda}^{(d+1)}|\boldsymbol{\lambda}^{(d)}) = -\int\int \log\left\{\frac{g(\theta_r,\theta_s|\vec{U}_r,\vec{U}_s;\boldsymbol{\lambda}_{r,s}^{(d+1)})}{g(\theta_r,\theta_s|\vec{U}_{\,r},\vec{U}_s;\boldsymbol{\lambda}_{r,s}^{(d)})}\right\} g(\theta_r,\theta_s|\vec{U}_{\,r},\vec{U}_s;\boldsymbol{\lambda}_{r,s}^{(d)})\mathrm{d}\theta_r\mathrm{d}\theta_s.
$$

Similar to the generalized EM (GEM) algorithm, in the PEM algorithm, $\boldsymbol{\lambda}^{(d+1)}$ is chosen such that

$$
Q(\boldsymbol{\lambda}^{(d+1)}|\boldsymbol{\lambda}^{(d)}) \geq Q(\boldsymbol{\lambda}^{(d)}|\boldsymbol{\lambda}^{(d)}). \tag{6}
$$

So, the difference on the right-hand side of (5) is non-negative. On the other hand, by Jensen's inequality, we have

$$
\begin{aligned}
D_{(r,s)}(\boldsymbol{\lambda}^{(d+1)}|\boldsymbol{\lambda}^{(d)}) &\geq -\log\int\int \left\{\frac{g(\theta_r,\theta_s|\vec{U}_r,\vec{U}_s;\boldsymbol{\lambda}_{r,s}^{(d+1)})}{g(\theta_r,\theta_s|\vec{U}_{\,r},\vec{U}_s;\boldsymbol{\lambda}_{r,s}^{(d)})}\right\} g(\theta_r,\theta_s|\vec{U}_{\,r},\vec{U}_s;\boldsymbol{\lambda}_{r,s}^{(d)})\mathrm{d}\theta_r\mathrm{d}\theta_s \\
&= \log\int\int g(\theta_r,\theta_s|\vec{U}_r,\vec{U}_s;\boldsymbol{\lambda}_{r,s}^{(d+1)})\mathrm{d}\theta_r\mathrm{d}\theta_s = 0.
\end{aligned}
$$

This result, together with (5) and (6), leads to the non-negativity of $pl(\boldsymbol{\lambda}^{(d+1)};U) - pl(\boldsymbol{\lambda}^{(d)};U)$, which implies that the map $pl(\cdot;U)$ induced by PEM into the parametric space is non-decreasing. If the expectation $Q(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(d)})$ cannot be expressed in closed form, it needs to be evaluated numerically. In general, the double integrals in $Q$ can be efficiently evaluated using Gauss-Hermite quadrature.

## 5. Conclusions

We developed a pairwise modeling strategy to fit the longitudinal test data within the MIRT framework. Compared to the existing joint modeling approach, the

pairwise method greatly reduces the dimensionality of the problem and hence can be used to model longitudinal test data over many time points.

The pairwise approach presents an elegant solution for fitting MIRT models without suffering from a restriction on the number of dimensions. Thus it provides a valuable alternative to the Bayesian Markov chain Monte Carlo methods proposed for MIRT models. When using Markov chain Monte Carlo methods, the computational burden increases rapidly as the number of examinees or items increases.

**References**

Adams, J. A., Wilson, M., & Wang, W. C. (1997). "The multidimensional random coefficients multinomial logit model," *Applied Psychological Measurement*, 21, 1-23.

Andrade, D. F., & Tavares, H. R. (2005). "Item response theory for longitudinal data: population parameter estimation," *Journal of Multivariate Analysis*, 95, 1-22.

Bock, R. D., & Aitkin, M. (1981). "Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm," *Psychometrika*, 46, 443-459.

Cox, D. R., & Reid, N. (2004)." A note on pseudolikelihood constructed from marginal densities," *Biometrika*, 91, 211-221.

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

Fieuws, S., & Verbeke, G. (2006). "Pairwise fitting of mixed models for the joint modelling of multivariate longitudinal profiles," *Biometrics*, 62, 424-431.

Fischer, G. H., & Ponocny, I. (1994). "An extension of the partial credit model with an application to the measurement of change," *Psychometrika*, 59, 177-192.

Fu, Zhi-Hui et al.(2011). "Analyzing Longitudinal Item Response Data via the Pairwise Fitting Method," *Psychometrika*, 46, 637-651.

Geys, H., Molenberghs, G., & Ryan, L. (1999). "Pseudolikelihood modeling of multivariate outcomes in developmental studies," *Journal of American Statistical Association*, 94, 734-745.

Lindsay, B. (1988). "Composite likelihood methods," In N. U. Prabhu(Ed.), *Statistical Inference from Stochastic Processes* (pp. 221-239). Providence RI: American Mathematical Society.

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

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

Wang, W. C., & Wu, C. Y. (2004). "Gain score in item response theory as an effect size measure," *Educational and Psychological Measurement*, 64, 758-780.