



A Marginal Maximum Likelihood Approach for Extended Quadratic Structural Equation Modeling with Ordinal Data

Shaobo Jin, Johan Vegelius & Fan Yang-Wallentin

To cite this article: Shaobo Jin, Johan Vegelius & Fan Yang-Wallentin (2020) A Marginal Maximum Likelihood Approach for Extended Quadratic Structural Equation Modeling with Ordinal Data, Structural Equation Modeling: A Multidisciplinary Journal, 27:6, 864-873, DOI: [10.1080/10705511.2020.1712552](https://doi.org/10.1080/10705511.2020.1712552)

To link to this article: <https://doi.org/10.1080/10705511.2020.1712552>



© 2020 The Author(s). Published with license by Taylor & Francis Group, LLC.



Published online: 12 Feb 2020.



Submit your article to this journal [↗](#)



Article views: 883



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)

A Marginal Maximum Likelihood Approach for Extended Quadratic Structural Equation Modeling with Ordinal Data

Shaobo Jin, Johan Vegelius, and Fan Yang-Wallentin

Uppsala University

ABSTRACT

The literature on non-linear structural equation modeling is plentiful. Despite this fact, few studies consider interactions between exogenous and endogenous latent variables. Further, it is well known that treating ordinal data as continuous produces bias, a problem which is enhanced when non-linear relationships between latent variables are incorporated. A marginal maximum likelihood-based approach is proposed in order to fit a non-linear structural equation model including interactions between exogenous and endogenous latent variables in the presence of ordinal data. In this approach, the exact gradient of the approximated observed log-likelihood is calculated in order to attain the approximated maximum likelihood estimator. A simulation study shows that the proposed method provides estimates with low bias and accurate coverage probabilities.

KEYWORDS

Non-linear SEM; interaction; adaptive Gauss–Hermite approximation; Laplace approximation

Introduction

Non-linear structural equation modeling (SEM) has been of great interest in the past few decades. Fruitful theories in psychology and business studies incorporate non-linear relations among latent variables. In the Theory of Planned Behavior, Ajzen (1991) emphasized the relevance of interaction between Motivation and Ability on Behavior. In a study focusing on the consumers' loyalty, Agustin and Singh (2005) proposed a quadratic relationship between the latent constructs Satisfaction, Trust, Value, and Loyalty intentions in the framework of marketing sciences. In order to fit a non-linear SEM model, various methods have been proposed, which include but are not limited to the product indicator approaches (e.g., Kelava & Brandt, 2009; Marsh, Wen, & Hau, 2004; Wall & Amemiya, 2001; Yang-Jonsson, 1997), the method of moments (e.g., Mooijaart & Bentler, 2010; Wall & Amemiya, 2000), two-stage least squares (e.g., Bollen, 1995, 1996), the distributional analytic approaches (e.g., Klein & Moosbrugger, 2000; Klein & Muthén, 2007; Lee & Zhu, 2002), and the Bayesian methods (e.g., Lee & Song, 2004a; Zhu & Lee, 1999). The reader is directed to Brandt, Kelava, and Klein (2014) and Harring, Weiss, and Hsu (2012) for reviews and simulation studies comparing the aforementioned methods. Despite that the methods of fitting a non-linear SEM model are numerous, they are not flawless and do not always meet the needs of real data analysis.

First, most studies have focused on models with continuous indicators. In such cases, distributional assumptions are violated if indicators are ordinal. This has been investigated and observed extensively in linear models (e.g., Finney & DiStefano, 2006; Flora & Curran, 2004). Rhemtulla, Brosseau-Liard, and Savalei (2012) showed that ordinal data cannot be treated as continuous, unless the number of categories is large and the categories are equidistant. Due to the complexity of the non-linear model, we do not expect

such methods to work well in a non-linear SEM model with ordinal data either. Lee and Zhu (2000) firstly considered a non-linear SEM model with the presence of polytomous indicators and proposed a Bayesian approach to estimate the parameters. Since then, various studies that directly handle the categorical nature emerge. Lee and Song (2003) extended the model in Lee and Zhu (2000) by incorporating covariates in the measurement and the structural model and proposed an MCEM algorithm (Wei & Tanner, 1990) to estimate the parameters. A similar algorithm was used by Song and Lee (2005) when the non-linear model has only dichotomous indicators. Lee and Song (2008) and Song and Lee (2006a) considered Bayesian estimation and comparison of non-linear models with a mixture of continuous, dichotomous, and ordinal indicators. Rizopoulos and Moustaki (2008) introduced non-linear terms of the latent variables to generalized latent variable models with covariates. When the non-linear model contains missing values, Lee and Song (2004a) proposed a Bayesian approach for a mixture of continuous and ordinal indicators, which is extended by Cai, Song, and Lee (2008) to model non-ignorable missing. Lee and Song (2007) proposed a unified maximum likelihood approach for non-linear models with different types of indicators and with or without missing values. Lee, Song, and Cai (2010) exemplified the use of the Bayesian approach in the non-linear models using dichotomous indicators as examples. Studies mentioned above are devoted to single group data analysis. It is further generalized by Song and Lee (2006b) to multigroup analysis and by Kelava and Brandt (2014), Lee and Song (2004b), Lee et al. (2009), and Song and Lee (2004) to multilevel analysis.

Second, in the Theory of Planned Behavior, Ajzen (1991) proposed that the effect of Intention (endogenous latent variable) on Behavior is affected by Perceived Behavioral Control (exogenous latent variable). Throughout the paper, an exogenous variable is a variable that is not caused by any other variables in the model

and an endogenous variable is a variable that is caused by some variables in the model. Even though some methods can incorporate ordinal data into the model, they generally lack the interaction between endogenous latent variables and exogenous latent variables. In contrast, Agustin and Singh (2005) proposed an interaction effect between Consumer Trust (endogenous latent variable) and Satisfaction (exogenous latent variable). They treated ordinal data as continuous and used the two-step approach of Ping (1995). To our best knowledge, there are no contemporary methods that can estimate non-linear SEM models including interactions between exogenous and endogenous latent variables in the presence of ordinal indicators.

The main purpose of the current paper is to fit the non-linear SEM model with ordinal data in the presence of interactions between latent exogenous and latent endogenous variables. We propose a marginal maximum likelihood approach to estimate the fixed unknown parameters in the model. Instead of the commonly implemented expectation-maximization (EM, Dempster, Laird, & Rubin, 1977) algorithm, we propose to use direct maximization. As we shall explain in later sections, the direct maximum likelihood approach is based on the true gradient of the approximated observed log-likelihood, which makes the sandwich estimator of standard errors easier to obtain. In contrast, the sandwich estimator of standard errors is not directly available in the EM algorithm. Further, when the gradients in the M-step are approximated, the EM algorithm can be less accurate than the direct maximum likelihood approach with the same approximation method (Jin & Andersson, 2019a).

The article is organized as follows. The model is presented in the next section, which is followed by a description of the proposed method and an outline of the suggested procedure for directly maximizing the marginal likelihood. Next, a simulation study is conducted to investigate the small-sample performance of the proposed method. An empirical example is presented to illustrate the method, followed by a discussion with conclusions.

Non-linear structural equation model

Let ξ (exogenous) and η (endogenous) be latent random vectors with dimensions K_ξ and K_η , respectively. The measurement model for X_j , the j^{th} ordinal indicator of ξ , is:

$$g_j^{(x)}(P(X_j \leq u_j | \xi)) = \alpha_{x,j}^{(u_j)} - \beta_{x,j}^T \xi, \quad u_j \in \{1, \dots, U_j\}, \quad (1)$$

where U_j is the number of response categories for the ordinal indicator X_j , $g_j^{(x)}$ is the link function, $\alpha_{x,j}^{(u_j)}$ is the category-specific intercept for observing u_j , and $\beta_{x,j}$ is the factor loading vector for X_j . Such a setting is similar to the cumulative link model in the context of generalized linear model (Agresti, 2015, p. 213), but only with latent variables as explanatory variables. Likewise, the measurement model for Y_j , the j^{th} ordinal indicator of η , is:

$$g_j^{(y)}(P(Y_j \leq v_j | \eta)) = \alpha_{y,j}^{(v_j)} - \beta_{y,j}^T \eta, \quad v_j \in \{1, \dots, V_j\}, \quad (2)$$

where V_j is the number of response categories for the ordinal indicator Y_j , $g_j^{(y)}$ is the link function, $\alpha_{y,j}^{(v_j)}$ is the category-specific intercept for observing v_j , and $\beta_{y,j}$ is the factor loading vector for Y_j . ξ is assumed to have a $\mathbf{0}$ mean vector and

a covariance matrix Φ . Further, partition η into a $K_{\eta_1} \times 1$ vector η_1 and a $K_{\eta_2} \times 1$ vector η_2 . The structural model of interest is:

$$\eta_1 = \tau_1 + B_{11}\eta_1 + B_{12}\eta_2 + \Gamma_1\xi + \left(I_{K_{\eta_1}} \otimes \xi\right)^T \Omega_1\xi + \left(I_{K_{\eta_1}} \otimes \xi\right)^T \Pi\eta_2 + \left(I_{K_{\eta_1}} \otimes \eta_2\right)^T \Xi\eta_2 + \zeta_1, \quad (3)$$

$$\eta_2 = \tau_2 + B_{22}\eta_2 + \Gamma_2\xi + \left(I_{K_{\eta_2}} \otimes \xi\right)^T \Omega_2\xi + \zeta_2, \quad (4)$$

where I_a is an $a \times a$ identity matrix and \otimes denotes the Kronecker product. The B matrices represent linear effects among endogenous latent variables, where B_{11} and B_{22} are $K_{\eta_1} \times K_{\eta_1}$ and $K_{\eta_2} \times K_{\eta_2}$ matrices, respectively, with zero diagonal elements, and B_{12} is a $K_{\eta_1} \times K_{\eta_2}$ matrix. The Ω matrices represent non-linear effects of exogenous latent variables on the endogenous latent variables, where Ω_1 and Ω_2 are $K_{\eta_1}K_\xi \times K_\xi$ and $K_{\eta_2}K_\xi \times K_\xi$ matrices, respectively. Π is a $K_{\eta_1}K_\xi \times K_{\eta_2}$ matrix representing interaction effects between endogenous and exogenous latent variables, and Ξ is a $K_{\eta_1}K_{\eta_2} \times K_{\eta_2}$ matrix, representing non-linear effects among η_2 on η_1 . The matrices Ω_1 , Ξ , and Ω_2 are block matrices stacking upper-triangular matrices on top of one another. For example, if $K_{\eta_1} = K_{\eta_2} = K_\xi = 2$, Ω_1 , Ω_2 , and Ξ consist of two upper-triangular matrices each, i.e.

$$\Omega_1 = \begin{pmatrix} \Omega_{11}^{(1)} & \Omega_{12}^{(1)} \\ 0 & \Omega_{22}^{(1)} \\ \Omega_{31}^{(1)} & \Omega_{32}^{(1)} \\ 0 & \Omega_{42}^{(1)} \end{pmatrix}, \quad \Omega_2 = \begin{pmatrix} \Omega_{11}^{(2)} & \Omega_{12}^{(2)} \\ 0 & \Omega_{22}^{(2)} \\ \Omega_{31}^{(2)} & \Omega_{32}^{(2)} \\ 0 & \Omega_{42}^{(2)} \end{pmatrix}, \quad \Xi = \begin{pmatrix} \Xi_{11} & \Xi_{12} \\ 0 & \Xi_{22} \\ \Xi_{31} & \Xi_{32} \\ 0 & \Xi_{42} \end{pmatrix}.$$

The diagonal elements of the upper-triangular matrices correspond to the quadratic effects, whereas the off-diagonal elements correspond to the interaction effects. The error terms ζ_1 and ζ_2 are assumed to have zero mean vectors and are independent of each other with $K_{\eta_1} \times K_{\eta_1}$ and $K_{\eta_2} \times K_{\eta_2}$ covariance matrices Ψ_{11} and Ψ_{22} , respectively. τ_1 and τ_2 are the intercepts. The intercept τ_2 is determined such that $E(\eta_2) = \mathbf{0}$, that is, the j^{th} element of τ_2 is defined by:

$$\tau_{2,j} = -\text{trace}\left(\Omega_j^{(2)}\Phi\right),$$

where $\Omega_j^{(2)}$ is the j^{th} upper-triangular matrix in Ω_2 . For simplicity, we let $\tau_1 = \mathbf{0}$. Thus, $E(\eta_1)$ is not zero.

To our knowledge, the majority of studies on non-linear SEM overlooked Equation (3) and focused on Equation (4). Take the path diagram of an extended quadratic structural model in Figure 1 as an example. If Equation (3) is ignored, the researchers are not able to model the paths in the dashed rectangle. As mentioned previously, the non-linear effects among endogenous and exogenous latent variables are commonly encountered in practice. Hence, it is of importance to include Equation (3) in the structural model.

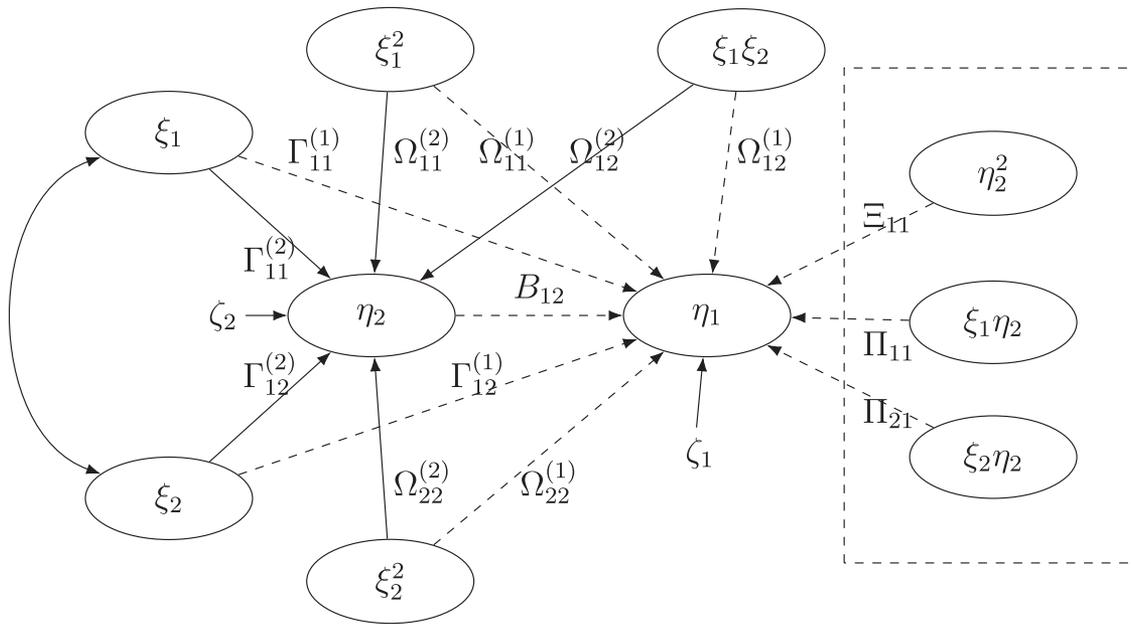


Figure 1. The path diagram of the structural model of an extended quadratic structural equation model. The dashed paths correspond to the effects in Equation (3) and the solid paths correspond to the effects in Equation (4).

Marginal maximum likelihood estimation

Under the assumption of independent observations, the complete likelihood function is:

$$\prod_{i=1}^n f(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\eta}_i, \boldsymbol{\xi}_i),$$

where n is the number of observations, and \mathbf{y}_i and \mathbf{x}_i are the i^{th} observed ordinal vectors of indicators. Under the conditional independence assumption, the complete likelihood becomes:

$$\prod_{i=1}^n f(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\eta}_i, \boldsymbol{\xi}_i) = \prod_{i=1}^n f(\mathbf{y}_i | \boldsymbol{\eta}_i) f(\mathbf{x}_i | \boldsymbol{\xi}_i) f(\boldsymbol{\eta}_{1,i} | \boldsymbol{\eta}_{2,i}, \boldsymbol{\xi}_i) f(\boldsymbol{\eta}_{2,i} | \boldsymbol{\xi}_i) f(\boldsymbol{\xi}_i). \tag{5}$$

We assume that $\boldsymbol{\xi} \sim N(\mathbf{0}, \boldsymbol{\Phi})$, $\zeta_1 \sim N(\mathbf{0}, \boldsymbol{\Psi}_{11})$, and $\zeta_2 \sim N(\mathbf{0}, \boldsymbol{\Psi}_{22})$. Hence, $\boldsymbol{\eta}_{1,i} | (\boldsymbol{\eta}_{2,i}, \boldsymbol{\xi}_i)$, $\boldsymbol{\eta}_{2,i} | \boldsymbol{\xi}_i$, and $\boldsymbol{\xi}_i$ are multivariate normal. The expression of $f(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\eta}_i, \boldsymbol{\xi}_i)$ under these normal assumptions can be found in the appendix.

In practice, $\boldsymbol{\eta}_1$ is often a scalar that has only one binary indicator. For example, $y_1 = 1$ if a behavior is conducted and 0 if not. If that is the case, further restrictions are needed for identification. One way is to let $\boldsymbol{\Psi}_{11} = 0$ and $\beta_{y,1} = 1$. This means that $\boldsymbol{\Psi}_{11}$ is singular and $\boldsymbol{\eta}_{1,i} | (\boldsymbol{\eta}_{2,i}, \boldsymbol{\xi}_i)$ is a one-point distribution. Equivalently, $\boldsymbol{\eta}_1$ is completely determined conditioned on $\boldsymbol{\xi}$ and $\boldsymbol{\eta}_2$.

Approximated observed log-likelihood

Let $\boldsymbol{\theta}$ be the vector of parameters that includes the unconstrained elements in the structural matrices in \mathbf{B}_{11} , \mathbf{B}_{12} , \mathbf{B}_{22} , $\boldsymbol{\Gamma}_1$, $\boldsymbol{\Gamma}_2$, $\boldsymbol{\Omega}_1$, $\boldsymbol{\Omega}_2$, $\boldsymbol{\Pi}$ and $\boldsymbol{\Xi}$, the unconstrained elements in the covariance matrices $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}_{11}$, and $\boldsymbol{\Psi}_{22}$, and the unconstrained intercepts and factor loadings $\alpha_{y,j}^{(v_j)}$, $\alpha_{x,j}^{(u_j)}$, $\beta_{y,j}$, and $\beta_{x,j}$. In order

to obtain the maximum likelihood estimator of $\boldsymbol{\theta}$, the observed likelihood function is calculated by integrating out $\mathbf{Y}_i = (\boldsymbol{\eta}_i^T, \boldsymbol{\xi}_i^T)^T$ with $\boldsymbol{\eta}_i = (\boldsymbol{\eta}_{1,i}^T, \boldsymbol{\eta}_{2,i}^T)^T$ from Equation (5), i.e.

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \log \int f(\mathbf{y}_i, \mathbf{x}_i, \mathbf{Y}_i; \boldsymbol{\theta}) d\mathbf{Y}_i. \tag{6}$$

The integral in Equation (6) is intractable and approximations are needed. The approximated maximum likelihood estimator is obtained by maximizing the approximated observed log-likelihood function. Commonly used numerical approximation methods include the Laplace approximation (e.g., Huber, Ronchetti, & Victoria-Feser, 2004) and the Gauss–Hermite quadrature-based methods (e.g., Moustaki, 1996; Moustaki & Knott, 2000; Rabe-Hesketh & Skrondal, 2004; Rizopoulos & Moustaki, 2008). Alternatively, the observed gradient can be approximated by stochastic methods (e.g., Cai, 2010). In the present study, the second-order Laplace approximation (Shun & McCullagh, 1995) and the adaptive Gauss–Hermite quadrature (AGHQ) approximation of Liu and Pierce (1994) are implemented. The second-order Laplace approximation is second-order accurate and the accuracy of the AGHQ approximation depends on the number of quadrature points. A general introduction of these methods and the expressions of the approximations to Equation (6) can be found in the appendix.

Maximizing the approximated observed log-likelihood

A common approach in latent variable models is to employ the EM algorithm or its variants, where the latent variables are treated as missing values and the gradient of the conditional expectation in the E-step is approximated in the M-step. For example, the EM algorithm is used by Lee and Song (2004b), Lee, Song, and Lee (2003), and Lee and Zhu (2000) among others. The EM algorithm is known for its slow speed in many situations. When the integrals

in the M-step are approximated, various studies (e.g., Bianconcini & Cagnone, 2012; Rizopoulos, Verbeke, & Lesaffre, 2009; Steele, 1996) argued that the fully exponential Laplace approximation (Tierney & Kadane, 1986; Tierney, Kass, & Kadane, 1989) is more accurate than the Laplace approximation. Jin and Andersson (2019a) proved that the estimator based on the EM algorithm with the fully exponential Laplace approximation is the same as the estimator that maximizes the Laplace approximated observed log-likelihood function, if implemented properly. Hence, we propose to directly maximize the approximated $\ell(\theta)$, instead of the EM algorithm.

The gradients of the approximations can be readily obtained by the chain rule, explained in detail in the appendix. A quasi-Newton method (e.g., BFGS algorithm) can be used to find the maximizer of the approximated observed log-likelihood function. Since the approximations are subject to approximation error, the proposed approach can be interpreted as the quasi-maximum likelihood approach (Huber et al., 2004). If the approximated log-likelihood function is directly maximized, the gradient of the approximation is already available, which makes the standard errors easy to estimate (see Equation (10) in the appendix). If the EM algorithm is used, the information matrix needed for standard errors is approximated afterward by, for example, the Louis method (Louis, 1982).

Simulation study

In order to investigate the performance of the suggested approach, a simulation study is conducted. The setup and result are presented in the following subsections.

Simulation design

In the simulation, a four-factor model is considered with two exogenous latent variables $\{\xi_1, \xi_2\}$ and two endogenous latent variables $\{\eta_1, \eta_2\}$. Each latent variable is associated with three ordinal indicators with five categories each. ξ, ζ_1 and ζ_2 are generated from multivariate normal distributions. The population parameters of the measurement model (Equations (1) and (2)) are

$$\alpha_x^T = \begin{pmatrix} -2.32 & -2.13 & -1.96 & -2.32 & -2.13 & -1.96 \\ 0.57 & 0.52 & 0.48 & 0.57 & 0.52 & 0.48 \\ 1.51 & 1.39 & 1.27 & 1.51 & 1.39 & 1.27 \\ 2.87 & 2.64 & 2.42 & 2.87 & 2.64 & 2.42 \end{pmatrix}$$

$$\beta_x^T = \begin{pmatrix} 2.00 & 1.80 & 1.60 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2.00 & 1.80 & 1.60 \end{pmatrix}$$

$$\alpha_y^T = \begin{pmatrix} -1.76 & -1.70 & -1.66 & -2.20 & -2.14 & -2.08 \\ -0.42 & -0.40 & -0.35 & -0.91 & -0.88 & -0.87 \\ 0.60 & 0.56 & 0.52 & 0.00 & 0.00 & 0.00 \\ 2.65 & 2.60 & 2.50 & 1.79 & 1.76 & 1.72 \end{pmatrix}$$

$$\beta_y^T = \begin{pmatrix} 1.00 & 0.95 & 0.90 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.00 & 0.95 & 0.90 \end{pmatrix},$$

where zero elements are restricted to zero. Here the $(j, k)^{th}$ element in α_x is $\alpha_{x,j}^{(k)}$ and the j^{th} row of β_x is $\beta_{x,j}$ in Equation

(1). Likewise, the $(j, k)^{th}$ element in α_y is $\alpha_{y,j}^{(k)}$ and the j^{th} row of β_y is $\beta_{y,j}$ in Equation (2). The factor loadings only load on one latent variable giving this setup a simple structure for simplicity. $\beta_{y,11}$ and $\beta_{y,42}$ are fixed to 1 for identification. The population covariance between the latent exogenous variables is 0.40 and the variances are fixed to 1 for identification. Ordinal observations are generated by the measurement model using the probit link function. Equations (3) and (4) are set to be the same as Figure 1. The population values of the structural parameters are

$$B_{11} = 0, B_{12} = 0.15, \Gamma_1 = (0.15 \quad 0.15),$$

$$\Omega_1 = \begin{pmatrix} 0.15 & 0.20 \\ 0 & 0.15 \end{pmatrix}, \Pi = \begin{pmatrix} 0.10 \\ 0.10 \end{pmatrix}, \Xi = 0.1,$$

$$B_{22} = 0, \Gamma_2 = (0.35 \quad 0.35), \Omega_2 = \begin{pmatrix} 0.20 & 0.25 \\ 0 & 0.20 \end{pmatrix},$$

where the zero elements are restricted to zero. The variances of the error terms are $\Psi_{11} = 2.00$ and $\Psi_{22} = 1.50$. Such population values are chosen to reflect common values in practice. The α_x and β_x are chosen such that the reliabilities of the items range from 0.72 to 0.8 and the probabilities of observing each outcomes are 0.15, 0.45, 0.15, 0.15, and 0.1, respectively, for all items. The structural parameters are chosen such that the R^2 's of the latent regression are approximately 0.43 and 0.34. The α_y and β_y are chosen such that the reliabilities of the items range from 0.65 to 0.78 and the probabilities of observing each outcomes are approximately 0.10, 0.20, 0.20, 0.35, and 0.15 for all items. Four different sample sizes are considered, i.e., $n = \{400, 600, 800, 1000\}$ and 2000 replications are generated for each sample size. Two second-order accurate approximation methods to the observed log-likelihood are considered in the simulation, namely, the second-order Laplace approximation, denoted by Lap(2nd), and the AGHQ with 5 quadrature points per latent variable, denoted by AGHQ(5p). We expect them to perform similarly in terms of bias. The approximated maximum likelihood estimator is obtained by maximizing the approximated observed log-likelihood using the direct maximum likelihood approach by the BFGS algorithm. The algorithm is stopped when the maximum absolute change in the parameter estimates of two consecutive iterations is less than 10^{-4} . The code was written in R (R Core Team, 2018) by the aid of the package Rcpp (Eddelbuettel, 2013).

Simulation results

Replications which do not converge and those with nonpositive definite covariance matrices are excluded from the analysis. The percentage of excluded cases is reported in Table 1. The proportion of inadmissible cases is generally ignorable and decreases as the sample size increases.

In order to investigate the performance of the methods the relative bias $RB_p = 100R^{-1} \sum_{i=1}^R \theta_p^{-1}(\hat{\theta}_{p,i} - \theta_p)$ and root mean

squared error $RMSE_p = \sqrt{R^{-1} \sum_{i=1}^R [\hat{\theta}_{p,i} - \theta_p]^2}$ are

Table 1. Percentage of inadmissible cases excluded from the analysis

Method	<i>n</i>			
	400	600	800	1000
Lap(2nd)	0.00	0.00	0.00	0.00
AGHQ(5p)	0.30	0.01	0.00	0.00

Lap(2nd) refers to the second-order Laplace approximation and AGHQ(5p) refers to the adaptive Gauss–Hermite quadrature approximation with 5 quadrature points per latent dimension.

calculated for each structural parameter θ_p , where R is the number of replications and $\hat{\theta}_{p,i}$ is the corresponding estimate at the i^{th} replication. Table 2 shows the RB and RMSE for the structural parameters B_{11} , B_{12} , B_{22} , Γ_1 , Γ_2 , Ω_1 , Ω_2 , Π and Ξ . Following the common practice in the SEM literature (e.g., Curran, West, & Finch, 1996; Flora & Curran, 2004), relative bias lower than 5 is regarded as a low bias. Generally, the methods provide low biases for the sample sizes investigated. The RMSE decreases with sample size as expected. Lap(2nd) and AGHQ(5p) perform similarly in terms of RB and RMSE across parameters and sample sizes. In order to evaluate the performances of the calculated standard errors, coverage probabilities at the nominal level 95% are presented in Table 3. The coverage probabilities of covering the true values are generally close to the nominal level. Results not shown here indicate that the first-order Laplace approximation that

is equivalent to the adaptive Gauss–Hermite quadrature approximation with 1 quadrature point can perform unacceptably poorly for non-linear terms.

The Laplace approximations are expected to be computationally more efficient than AGHQ approximations for similar error rates if the number of latent variables is large. However, a higher-order Laplace approximation requires calculations of higher-order derivatives which are not straightforward, whereas the AGHQ approximation provides the possibility of decreasing error rates by increasing the number of quadrature points with the drawback of being computationally heavier. In the current simulation, the average estimation time for the sample size $n = 1000$ takes on average 90 s for AGHQ(5p) and 49 s for Lap(2nd). This relative difference is hypothesized to increase with increasing number of latent variables.

Empirical example

In order to demonstrate how the proposed method works in practice, an empirical example is considered in this section. The empirical example contains 707 complete responses of households in two Swedish municipalities Sollentuna and Saltjö-boo regarding their behavior of off-peak hour energy use (Bartusch, Juslin, Stitkvoort, Yang-Wallentin, & Öhrlund,

Table 2. Relative bias and root mean squared error of the estimators of structural parameters in Equations (3) and (4)

<i>n</i>	Method	Parameter													
		B_{12}	$\Gamma_{11}^{(1)}$	$\Gamma_{12}^{(1)}$	$\Omega_{11}^{(1)}$	$\Omega_{12}^{(1)}$	$\Omega_{22}^{(1)}$	Π_{11}	Π_{21}	Ξ_{11}	$\Gamma_{11}^{(2)}$	$\Gamma_{12}^{(2)}$	$\Omega_{11}^{(2)}$	$\Omega_{12}^{(2)}$	$\Omega_{22}^{(2)}$
Relative bias															
400	Lap (2nd)	4.59	0.87	-0.08	-2.22	2.81	-2.16	1.65	-2.88	1.53	1.98	2.25	-0.89	3.12	-0.44
	AGHQ (5p)	2.09	3.03	2.24	0.86	4.69	1.03	2.66	-1.69	-2.23	1.33	1.63	0.29	1.54	0.77
600	Lap (2nd)	2.54	0.53	2.93	-1.27	0.07	0.30	-4.22	-2.80	2.01	1.93	1.81	0.22	2.99	-0.58
	AGHQ (5p)	0.26	2.51	4.79	1.36	2.23	2.83	-3.18	-1.63	-1.62	1.30	1.18	1.07	1.93	0.23
800	Lap (2nd)	2.05	-0.20	0.32	-2.34	1.33	-0.87	-1.45	-0.22	-0.13	0.82	1.44	0.30	0.28	0.48
	AGHQ (5p)	0.04	1.65	2.17	-0.02	3.65	1.57	-0.40	0.88	-3.53	0.22	0.82	0.95	-0.54	1.15
1000	Lap (2nd)	1.54	-0.57	-1.74	-0.40	-1.24	-0.32	2.47	-0.61	0.76	1.27	1.32	-0.31	1.84	-0.24
	AGHQ (5p)	-0.50	1.21	0.07	1.79	1.22	1.90	3.35	0.35	-2.66	0.65	0.71	0.23	1.08	0.30
Root mean squared error multiplied by 10															
400	Lap (2nd)	0.92	1.19	1.22	1.05	1.54	1.01	0.99	0.96	0.56	1.03	0.99	0.86	1.30	0.85
	AGHQ (5p)	0.91	1.19	1.21	1.07	1.53	1.03	0.98	0.95	0.55	1.02	0.99	0.87	1.30	0.87
600	Lap (2nd)	0.74	0.97	0.94	0.82	1.19	0.82	0.76	0.77	0.46	0.85	0.86	0.69	1.01	0.67
	AGHQ (5p)	0.74	0.97	0.94	0.84	1.18	0.85	0.76	0.77	0.45	0.84	0.86	0.71	1.02	0.69
800	Lap (2nd)	0.63	0.82	0.82	0.70	1.02	0.69	0.65	0.65	0.38	0.71	0.71	0.59	0.89	0.58
	AGHQ (5p)	0.63	0.82	0.82	0.72	1.02	0.70	0.65	0.64	0.38	0.71	0.71	0.60	0.90	0.60
1000	Lap (2nd)	0.56	0.72	0.72	0.62	0.89	0.64	0.58	0.58	0.34	0.65	0.63	0.54	0.78	0.53
	AGHQ (5p)	0.56	0.72	0.72	0.64	0.89	0.65	0.59	0.58	0.34	0.65	0.63	0.55	0.79	0.54

Lap(2nd) refers to the second-order Laplace approximation and AGHQ(5p) refers to the adaptive Gauss–Hermite quadrature approximation with 5 quadrature points per latent dimension. The superscript indicates to which η the parameter belongs.

Table 3. Coverage probabilities (%) of the Wald confidence interval for structural parameters in Equations (3) and (4) at the nominal level of 95%

<i>n</i>	Method	Parameter													
		B_{12}	$\Gamma_{11}^{(1)}$	$\Gamma_{12}^{(1)}$	$\Omega_{11}^{(1)}$	$\Omega_{12}^{(1)}$	$\Omega_{22}^{(1)}$	Π_{11}	Π_{21}	Ξ_{11}	$\Gamma_{11}^{(2)}$	$\Gamma_{12}^{(2)}$	$\Omega_{11}^{(2)}$	$\Omega_{12}^{(2)}$	$\Omega_{22}^{(2)}$
400	Lap (2nd)	95.2	94.7	94.3	94.9	94.3	96.0	93.9	95.0	95.0	95.3	96.4	94.8	94.9	95.5
	AGHQ (5p)	95.3	94.4	94.3	94.8	94.4	95.9	94.1	94.8	94.9	95.6	96.4	94.5	95.1	95.3
600	Lap (2nd)	95.2	94.4	95.7	95.1	94.8	95.4	95.3	94.4	94.2	94.5	94.6	95.6	95.7	96.2
	AGHQ (5p)	95.6	94.5	95.9	94.8	95.3	95.1	95.6	94.4	93.6	94.8	94.6	95.0	95.9	95.7
800	Lap (2nd)	95.0	95.3	94.5	95.3	94.8	95.5	95.2	95.3	95.0	95.8	95.3	95.2	94.7	95.4
	AGHQ (5p)	95.1	95.2	95.0	95.0	95.0	95.3	95.3	95.4	94.2	96.0	95.6	95.2	95.0	95.4
1000	Lap (2nd)	95.4	95.6	95.5	95.5	96.1	94.8	95.4	95.2	95.3	95.0	95.9	94.8	95.2	95.0
	AGHQ (5p)	95.7	95.5	95.5	94.8	96.7	94.7	95.0	94.7	94.4	95.1	95.9	94.5	95.4	95.0

Lap(2nd) refers to the second-order Laplace approximation and AGHQ(5p) refers to the adaptive Gauss–Hermite–quadrature approximation with 5 quadrature points per latent dimension. The superscript indicates to which η the parameter belongs.

Table 4. Point estimates and 95% confidence intervals for structural parameters in the empirical example corresponding to non-linear effects for the AGHQ approximation with 5-quadrature points, denoted by AGHQ(5p), and second-order Laplace approximation, denoted by Lap(2nd)

	Parameter							
	B_{12}	$\Gamma_{13}^{(1)}$	$\Omega_{33}^{(1)}$	Π_{31}	Ξ_{11}	$\Gamma_{11}^{(2)}$	$\Gamma_{12}^{(2)}$	$\Gamma_{13}^{(2)}$
AGHQ (5p)								
Point estimate	-0.042	-0.046	-0.074	0.083	-0.008	1.802	0.636	0.069
Lower bound	-0.098	-0.230	-0.252	-0.010	-0.029	1.037	0.302	-0.324
Upper bound	0.014	0.138	0.104	0.177	0.012	2.568	0.971	0.463
Lap (2nd)								
Point estimate	-0.045	-0.034	-0.088	0.083	-0.006	1.919	0.546	0.103
Lower bound	-0.105	-0.190	-0.327	-0.017	-0.033	1.153	0.140	-0.232
Upper bound	0.014	0.123	0.151	0.182	0.021	2.686	0.953	0.438

2019), in order to investigate the roles of the psychological factors, Attitude (ξ_1), Belief (ξ_2), Perceived Behavioral Control (ξ_3) and Intentions (η_2) in relation to the Behavior (η_1) of shifting the energy use to off-peak hours. It is hypothesized that Behavior (η_1) is affected by an interaction between Perceived Behavioral Control (ξ_3) and Intentions (η_2). η_1 has only one continuous indicator and it is dichotomized into a binary indicator. Other latent variables are associated with two indicators each, in the form of 7-point Likert scale items. The 7-point Likert scale items were aggregated into five categories in order to avoid categories with too few observations. For the purpose of modeling the interaction effects, we also include ξ_3^2 and η_2^2 in the model, as suggested by Kelava and Brandt (2009). The structural model of interest is then:

$$\eta_1 = B_{12}\eta_2 + \begin{pmatrix} 0 & 0 & \Gamma_{13}^{(1)} \end{pmatrix} \xi + \xi^T \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \Omega_{33}^{(1)} \end{pmatrix} \xi + \xi^T \begin{pmatrix} 0 \\ 0 \\ \Pi_{31} \end{pmatrix} \eta_2 + \Xi_{11}\eta_2^2,$$

$$\eta_2 = \tau_2 + \begin{pmatrix} \Gamma_{11}^{(2)} & \Gamma_{12}^{(2)} & \Gamma_{13}^{(2)} \end{pmatrix} \xi + \xi^T \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \xi + \zeta_2.$$

Table 4 shows the estimates and the 95% confidence intervals for the structural parameters using the AGHQ approximation with 5 quadrature points per latent variable and the second-order Laplace approximation. None of the non-linear parameters $\Omega_{33}^{(1)}$, Π_{31} and Ξ_{11} is significant at the 5% significance level. The linear effects of Attitude and Belief on Intention ($\Gamma_{11}^{(2)}$ and $\Gamma_{12}^{(2)}$) are significant at the 5% significance level, whereas the linear effects of Perceived Behavioral Control on Intention and Behavior ($\Gamma_{13}^{(2)}$ and $\Gamma_{13}^{(1)}$) and of Intention on Behavior (B_{12}), are not significant at the 5% significance level. Results not presented here, show that similar results are obtained when 7 quadrature points are used, which indicates that the approximation is stabilized.

Discussion and conclusion

In this article, a structural model with quadratic and interaction relations is considered in the presence of ordinal data. Structural models including interactions between exogenous and endogenous latent variables are rarely found in the literature. The current

study provides the possibility to include such features. To this end, a marginal maximum likelihood approach is proposed, where the exact gradient of the approximated observed log-likelihood is computed and used to approximate the Hessian, providing the opportunity to obtain standard errors straightforwardly, which are not easily available using, for example, the EM algorithm.

Simulation results indicate that the second-order Laplace approximation and the AGHQ approximation with 5 quadrature points per latent variable provide accurate parameter estimators at finite sample sizes. On one hand, despite that only simple structures are considered in the current study, the proposed approach works for general loading structures, provided that the model is identified. On the other hand, more research is needed to provide a rule-of-thumb sample size requirement for non-linear models with interactions between endogenous and exogenous variables. We leave it as a future topic.

In this study, the Laplace approximation and the AGHQ approximation are used to approximate the log-likelihood function, where the quadrature nodes are chosen deterministically. Alternative approaches include the Monte Carlo-based maximum likelihood approaches (e.g., Lee & Song, 2003; Song & Lee, 2005) and the Bayesian approaches (e.g., Lee & Song, 2008; Lee & Zhu, 2000). To our knowledge, they are not yet implemented in the non-linear SEM model of our interest. This remains as a topic for future studies.

A drawback of the method is that it relies on normally distributed latent exogenous variables and error terms. It remains for future research to investigate the sensitivity against nonnormality and develop robust approaches. For example, non-normality in the latent exogenous variables could be modeled by mixtures of normals.

Acknowledgment

We thank the reviewers for useful comments that lead to an improved manuscript.

Funding

The work was supported by the Swedish Research Council [contract number 2017-01175].

References

Agresti, A. (2015). *Foundations of linear and generalized linear models*. Hoboken, NJ: John Wiley & Sons.

- Agustin, C., & Singh, J. (2005). Curvilinear effects of consumer loyalty determinants in relational exchanges. *Journal of Marketing Research*, 42, 96–108. doi:10.1509/jmkr.42.1.96.56961
- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50, 179–211. doi:10.1016/0749-5978(91)90020-T
- Barndorff-Nielsen, O., & Cox, D. (1989). *Asymptotic techniques for use in statistics*. London, UK: Chapman and Hall.
- Bartusch, C., Juslin, P., Stitkvoort, B., Yang-Wallentin, F., & Öhrlund, I. (2019). *The black box of demand response* (Unpublished manuscript). Department of Engineering Sciences, Uppsala University, Uppsala, Sweden.
- Bianconcini, S. (2014). Asymptotic properties of adaptive maximum likelihood estimators in latent variable models. *Bernoulli*, 20, 1507–1531. doi:10.3150/13-BEJ531
- Bianconcini, S., & Cagnone, S. (2012). Estimation of generalized linear latent variable models via full exponential Laplace approximation. *Journal of Multivariate Analysis*, 112, 183–193. doi:10.1016/j.jmva.2012.06.005
- Bollen, K. A. (1995). Structural equation models that are nonlinear in latent variables: A least-squares estimator. *Sociological Methodology*, 25, 223–251. doi:10.2307/271068
- Bollen, K. A. (1996). An alternative two stage least squares (2SLS) estimator for latent variable equations. *Psychometrika*, 61, 109–121. doi:10.1007/BF02296961
- Brandt, H., Kelava, A., & Klein, A. (2014). A simulation study comparing recent approaches for the estimation of nonlinear effects in sem under the condition of nonnormality. *Structural Equation Modeling*, 21, 181–195. doi:10.1080/10705511.2014.882660
- Cai, J. H., Song, X. Y., & Lee, S. Y. (2008). Bayesian analysis of nonlinear structural equation models with mixed continuous, ordered and unordered categorical, and nonignorable missing data. *Statistics and Its Interface*, 1, 99–114. doi:10.4310/SII.2008.v1.n1.a9
- Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, 75, 33–57. doi:10.1007/s11336-009-9136-x
- Core Team, R. (2018). *R: A language and environment for statistical computing*. Vienna, AU: R Foundation for Statistical Computing.
- Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, 1, 16–29. doi:10.1037/1082-989X.1.1.16
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39, 1–38. doi:10.1111/rssb.1977.39.issue-1
- Eddelbuettel, D. (2013). *Seamless R and C++ integration with Rcpp*. New York, NY: Springer.
- Finney, S. J., & DiStefano, C. (2006). *Structural equation modeling: A second course*. Charlotte, NC: Information Age Publishing.
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9, 446–491. doi:10.1037/1082-989X.9.4.466
- Harring, J. R., Weiss, B. A., & Hsu, J.-C. (2012). A comparison of methods for estimating quadratic effects in nonlinear structural equation models. *Psychological Methods*, 17, 193–214. doi:10.1037/a0027539
- Huber, P., Ronchetti, E., & Victoria-Feser, M.-P. (2004). Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 66, 893–908. doi:10.1111/j.1467-9868.2004.05627.x
- Jin, S., & Andersson, B. (2019a). *Gains of fully exponential Laplace approximation in latent variable models* (Unpublished manuscript). Department of Statistics, Uppsala University, Uppsala, Sweden.
- Jin, S., & Andersson, B. (2019b). A note on the accuracy of adaptive Gauss-Hermite quadrature. *Accepted by Biometrika*.
- Jin, S., Noh, M., & Lee, Y. (2018). H-likelihood approach to factor analysis for ordinal data. *Structural Equation Modeling*, 25, 530–540. doi:10.1080/10705511.2017.1403287
- Kelava, A., & Brandt, H. (2009). Estimation of nonlinear latent structural equation models using the extended unconstrained approach. *Review of Psychology*, 16, 123–132.
- Kelava, A., & Brandt, H. (2014). A general non-linear multilevel structural equation mixture model. *Frontiers in Psychology*, 5, 748. doi:10.3389/fpsyg.2014.00748
- Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, 65, 457–474. doi:10.1007/BF02296338
- Klein, A., & Muthén, B. (2007). Quasi-maximum likelihood estimation of structural equation models with multiple interaction and quadratic effects. *Multivariate Behavioral Research*, 42, 647–673. doi:10.1080/00273170701710205
- Lee, S. Y., & Song, X. Y. (2003). Maximum likelihood estimation and model comparison of nonlinear structural equation models with continuous and polytomous variables. *Computational Statistics & Data Analysis*, 44, 125–142. doi:10.1016/S0167-9473(02)00305-5
- Lee, S. Y., Song, X. Y., Cai, J. H., So, W. Y., Ma, C. W., & Chan, C. N. J. (2009). Non-linear structural equation models with correlated continuous and discrete data. *British Journal of Mathematical and Statistical Psychology*, 62, 327–347. doi:10.1348/000711008X292343
- Lee, S. Y., & Zhu, H. T. (2002). Maximum likelihood estimation of nonlinear structural equation models. *Psychometrika*, 67, 189–210. doi:10.1007/BF02294842
- Lee, S.-Y., & Song, X.-Y. (2004a). Bayesian model comparison of nonlinear structural equation models with missing continuous and ordinal categorical data. *British Journal of Mathematical and Statistical Psychology*, 57, 131–150. doi:10.1348/000711004849204
- Lee, S.-Y., & Song, X.-Y. (2004b). Maximum likelihood analysis of a general latent variable model with hierarchical mixed data. *Biometrics*, 60, 624–636. doi:10.1111/j.0006-341X.2004.00211.x
- Lee, S.-Y., & Song, X.-Y. (2007). A unified maximum likelihood approach for analyzing structural equation models with missing non-standard data. *Sociological Methods & Research*, 35, 352–381. doi:10.1177/0049124106292357
- Lee, S.-Y., & Song, X.-Y. (2008). On Bayesian estimation and model comparison of an integrated structural equation model. *Computational Statistics and Data Analysis*, 52, 4814–4827. doi:10.1016/j.csda.2008.03.029
- Lee, S.-Y., Song, X.-Y., & Cai, J.-H. (2010). A bayesian approach for nonlinear structural equation models with dichotomous variables using logit and probit links. *Structural Equation Modeling*, 17, 280–302. doi:10.1080/10705511003659425
- Lee, S.-Y., Song, X.-Y., & Lee, J. C. (2003). Maximum likelihood estimation of nonlinear structural equation models with ignorable missing data. *Journal of Educational and Behavioral Statistics*, 28, 111–134. doi:10.3102/10769986028002111
- Lee, S.-Y., & Zhu, H.-T. (2000). Statistical analysis of nonlinear structural equation models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology*, 53, 209–232. doi:10.1348/000711000159303
- Lee, Y., & Nelder, J. A. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, 88, 987–1006. doi:10.1093/biomet/88.4.987
- Liu, Q., & Pierce, D. A. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, 81, 624–629.
- Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 44, 226–233. doi:10.1111/rssb.1982.44.issue-2
- Marsh, H. W., Wen, Z., & Hau, K.-T. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods*, 9, 275–300. doi:10.1037/1082-989X.9.3.275
- Mooijaart, A., & Bentler, P. M. (2010). An alternative approach for nonlinear latent variable models. *Structural Equation Modeling*, 17, 357–373. doi:10.1080/10705511.2010.488997
- Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables. *British Journal of Mathematical and Statistical Psychology*, 49, 313–334. doi:10.1111/j.2044-8317.1996.tb01091.x
- Moustaki, I., & Knott, M. (2000). Generalized latent trait models. *Psychometrika*, 65, 391–411. doi:10.1007/BF02296153

- Ping, R. A. (1995). A parsimonious estimating technique for interaction and quadratic latent variables. *Journal of Marketing Research*, 32, 336–347. doi:10.1177/002224379503200308
- Rabe-Hesketh, S., & Skrondal, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, 69, 167–190. doi:10.1007/BF02295939
- Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? a comparison of robust continuous and categorical sem estimation methods under suboptimal condition. *Psychological Methods*, 17, 354–373. doi:10.1037/a0029315
- Rizopoulos, D., & Moustaki, I. (2008). Generalized latent variable models with non-linear effects. *British Journal of Mathematical and Statistical Psychology*, 61, 415–438. doi:10.1348/000711007X213963
- Rizopoulos, D., Verbeke, G., & Lesaffre, E. (2009). Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 71, 637–654. doi:10.1111/j.1467-9868.2008.00704.x
- Shun, Z., & McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57, 749–760. doi:10.1111/rssb.1995.57.issue-4
- Song, X.-Y., & Lee, S.-Y. (2004). Bayesian analysis of two-level nonlinear structural equation models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology*, 57, 29–52. doi:10.1348/000711004849259
- Song, X.-Y., & Lee, S.-Y. (2005). Maximum likelihood analysis of nonlinear structural equation models with dichotomous variables. *Multivariate Behavioral Research*, 40, 151–177. doi:10.1207/s15327906mbr4002_1
- Song, X.-Y., & Lee, S.-Y. (2006a). Bayesian analysis of structural equation models with nonlinear covariates and latent variables. *Multivariate Behavioral Research*, 41, 337–365. doi:10.1207/s15327906mbr4103_4
- Song, X.-Y., & Lee, S.-Y. (2006b). A maximum likelihood approach for multisample nonlinear structural equation models with missing continuous and dichotomous data. *Structural Equation Modeling*, 13, 325–351. doi:10.1207/s15328007sem1303_1
- Steele, B. M. (1996). A modified EM algorithm for estimation in generalized mixed models. *Biometrics*, 52, 1295–1310. doi:10.2307/2532845
- Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81, 82–86. doi:10.1080/01621459.1986.10478240
- Tierney, L., Kass, R. E., & Kadane, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, 84, 710–716. doi:10.1080/01621459.1989.10478824
- Wall, M. M., & Amemiya, Y. (2000). Estimation for polynomial structural equation models. *Journal of the American Statistical Association*, 95, 929–940. doi:10.1080/01621459.2000.10474283
- Wall, M. M., & Amemiya, Y. (2001). Generalized appended product indicator procedure for nonlinear structural equation analysis. *Journal of Educational and Behavioral Statistics*, 26, 1–29. doi:10.3102/10769986026001001
- Wei, G. C. G., & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistician Association*, 85, 699–704. doi:10.1080/01621459.1990.10474930
- Yang-Jonsson, F. (1997). *Non-linear structural equation models: Simulation studies of the Kenny-Judd model* (Doctoral dissertation), Department of Statistics, Uppsala University, Uppsala, Sweden.
- Zhu, H.-T., & Lee, S.-Y. (1999). Statistical analysis of nonlinear factor analysis models. *British Journal of Mathematical and Statistical Psychology*, 52, 225–242. doi:10.1348/000711099159080

Appendix

Approximated Maximum Likelihood Estimator

In this section, the proposed maximum likelihood estimator is described in detail. For notational convenience, let the complete log-likelihood be:

$$h_i(\mathbf{Y}_i; \boldsymbol{\theta}) = \log f(\mathbf{y}_i, \mathbf{x}_i, \mathbf{Y}_i; \boldsymbol{\theta}).$$

Let p_x and p_y be the number of indicators associated with $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$, respectively. Under the distributional assumptions $\boldsymbol{\xi} \sim N(\mathbf{0}, \boldsymbol{\Phi})$, $\zeta_1 \sim N(\mathbf{0}, \boldsymbol{\Psi}_{11})$ and $\zeta_2 \sim N(\mathbf{0}, \boldsymbol{\Psi}_{22})$,

$$h_i(\mathbf{Y}_i; \boldsymbol{\theta}) = \log f(\mathbf{y}_i | \boldsymbol{\eta}_i) + \log f(\mathbf{x}_i | \boldsymbol{\xi}_i) + \log f(\boldsymbol{\eta}_{1,i} | \boldsymbol{\eta}_{2,i}, \boldsymbol{\xi}_i) + \log f(\boldsymbol{\eta}_{2,i} | \boldsymbol{\xi}_i) + \log f(\boldsymbol{\xi}_i),$$

where

$$\log f(\mathbf{y}_i | \boldsymbol{\eta}_i) = \sum_{j=1}^{p_y} \log P(Y_{ij} = v_{ij} | \boldsymbol{\eta}_i),$$

$$\log f(\mathbf{x}_i | \boldsymbol{\xi}_i) = \sum_{j=1}^{p_x} \log P(X_{ij} = u_{ij} | \boldsymbol{\xi}_i),$$

$$\begin{aligned} \log f(\boldsymbol{\eta}_{1,i} | \boldsymbol{\eta}_{2,i}, \boldsymbol{\xi}_i) &= -\frac{K_{\eta_1}}{2} \log(2\pi) - \frac{1}{2} \log[\det(\boldsymbol{\Psi}_{11})] + \log \left[\det(\mathbf{I}_{K_{\eta_1}} - \mathbf{B}_{11}) \right] \\ &\quad - \frac{1}{2} \left((\mathbf{I}_{K_{\eta_1}} - \mathbf{B}_{11}) \left[\boldsymbol{\eta}_{1,i} - E(\boldsymbol{\eta}_{1,i} | \boldsymbol{\eta}_{2,i}, \boldsymbol{\xi}_i) \right] \right)^T \boldsymbol{\Psi}_{11}^{-1} \left((\mathbf{I}_{K_{\eta_1}} - \mathbf{B}_{11}) \left[\boldsymbol{\eta}_{1,i} - E(\boldsymbol{\eta}_{1,i} | \boldsymbol{\eta}_{2,i}, \boldsymbol{\xi}_i) \right] \right), \end{aligned}$$

$$\begin{aligned} \log f(\boldsymbol{\eta}_{2,i} | \boldsymbol{\xi}_i) &= -\frac{K_{\eta_2}}{2} \log(2\pi) - \frac{1}{2} \log[\det(\boldsymbol{\Psi}_{22})] + \log \left[\det(\mathbf{I}_{K_{\eta_2}} - \mathbf{B}_{22}) \right] \\ &\quad - \frac{1}{2} \left((\mathbf{I}_{K_{\eta_2}} - \mathbf{B}_{22}) \left[\boldsymbol{\eta}_{2,i} - E(\boldsymbol{\eta}_{2,i} | \boldsymbol{\xi}_i) \right] \right)^T \boldsymbol{\Psi}_{22}^{-1} \left((\mathbf{I}_{K_{\eta_2}} - \mathbf{B}_{22}) \left[\boldsymbol{\eta}_{2,i} - E(\boldsymbol{\eta}_{2,i} | \boldsymbol{\xi}_i) \right] \right), \end{aligned}$$

$$\log f(\boldsymbol{\xi}_i) = -\frac{K_{\xi}}{2} \log(2\pi) - \frac{1}{2} \log|\boldsymbol{\Phi}| - \frac{1}{2} \boldsymbol{\xi}_i^T \boldsymbol{\Phi}^{-1} \boldsymbol{\xi}_i,$$

with

$$E(\boldsymbol{\eta}_{1,i}|\boldsymbol{\eta}_{2,i}, \boldsymbol{\xi}_i) = (\mathbf{I}_{K_{\eta_1}} - \mathbf{B}_{11})^{-1} [\mathbf{B}_{12}\boldsymbol{\eta}_{2,i} + \boldsymbol{\Gamma}_1\boldsymbol{\xi}_i + (\mathbf{I}_{K_{\eta_1}} \otimes \boldsymbol{\xi}_i^T)\boldsymbol{\Omega}_1\boldsymbol{\xi}_i + (\mathbf{I}_{K_{\eta_1}} \otimes \boldsymbol{\xi}_i^T)\boldsymbol{\Pi}\boldsymbol{\eta}_{2,i} + (\mathbf{I}_{K_{\eta_1}} \otimes \boldsymbol{\eta}_{2,i}^T)\boldsymbol{\Xi}\boldsymbol{\eta}_{2,i}],$$

and

$$E(\boldsymbol{\eta}_{2,i}|\boldsymbol{\xi}_i) = (\mathbf{I}_{K_{\eta_2}} - \mathbf{B}_{22})^{-1} [\boldsymbol{\tau}_2 + \boldsymbol{\Gamma}_2\boldsymbol{\xi}_i + (\mathbf{I}_{K_{\eta_2}} \otimes \boldsymbol{\xi}_i^T)\boldsymbol{\Omega}_2\boldsymbol{\xi}_i].$$

Further, $P(Y_{ij} = v_{ij}|\boldsymbol{\eta}_i)$ and $P(X_{ij} = u_{ij}|\boldsymbol{\xi}_i)$ depend on the link function. For example, with the probit link,

$$P(Y_{ij} = v_{ij}|\boldsymbol{\eta}_i) = \begin{cases} \Phi(\boldsymbol{\alpha}_{y,j}^{(1)} - \boldsymbol{\beta}_{y,j}^T\boldsymbol{\eta}_i), & v_{ij} = 1, \\ 1 - \Phi(\boldsymbol{\alpha}_{y,j}^{(V_j-1)} - \boldsymbol{\beta}_{y,j}^T\boldsymbol{\eta}_i), & v_{ij} = V_j, \\ \Phi(\boldsymbol{\alpha}_{y,j}^{(v_{ij})} - \boldsymbol{\beta}_{y,j}^T\boldsymbol{\eta}_i) - \Phi(\boldsymbol{\alpha}_{y,j}^{(v_{ij}-1)} - \boldsymbol{\beta}_{y,j}^T\boldsymbol{\eta}_i), & \text{otherwise,} \end{cases}$$

$$P(X_{ij} = u_{ij}|\boldsymbol{\xi}_i) = \begin{cases} \Phi(\boldsymbol{\alpha}_{x,j}^{(1)} - \boldsymbol{\beta}_{x,j}^T\boldsymbol{\xi}_i), & u_{ij} = 1, \\ 1 - \Phi(\boldsymbol{\alpha}_{x,j}^{(U_j-1)} - \boldsymbol{\beta}_{x,j}^T\boldsymbol{\xi}_i), & u_{ij} = U_j, \\ \Phi(\boldsymbol{\alpha}_{x,j}^{(u_{ij})} - \boldsymbol{\beta}_{x,j}^T\boldsymbol{\xi}_i) - \Phi(\boldsymbol{\alpha}_{x,j}^{(u_{ij}-1)} - \boldsymbol{\beta}_{x,j}^T\boldsymbol{\xi}_i), & \text{otherwise,} \end{cases}$$

where V_j is the number of response categories for the ordinal indicator Y_j , U_j is the number of response categories for the ordinal indicator X_j , and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. Then, the observed log-likelihood to be approximated is:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \log \int f(\mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\Upsilon}_i; \boldsymbol{\theta}) d\boldsymbol{\Upsilon}_i = \sum_{i=1}^n \log \int \exp\{h_i(\boldsymbol{\Upsilon}_i; \boldsymbol{\theta})\} d\boldsymbol{\Upsilon}_i.$$

If $\boldsymbol{\eta}_1$ is a scalar that has only one binary indicator, we restrict $\Psi_{11} = 0$ and $\boldsymbol{\beta}_{y,11} = 1$. The modified complete log-likelihood is then:

$$h_i(\boldsymbol{\Upsilon}_i; \boldsymbol{\theta}) = \log f(\mathbf{y}_i|\boldsymbol{\eta}_i) + \log f(\mathbf{x}_i|\boldsymbol{\xi}_i) + \log f(\boldsymbol{\eta}_{2,i}|\boldsymbol{\xi}_i) + \log f(\boldsymbol{\xi}_i),$$

where

$$\boldsymbol{\eta}_{1,i} = (\mathbf{I}_{K_{\eta_1}} - \mathbf{B}_{11})^{-1} [\mathbf{B}_{12}\boldsymbol{\eta}_{2,i} + \boldsymbol{\Gamma}_1\boldsymbol{\xi}_i + (\mathbf{I}_{K_{\eta_1}} \otimes \boldsymbol{\xi}_i^T)\boldsymbol{\Omega}_1\boldsymbol{\xi}_i + (\mathbf{I}_{K_{\eta_1}} \otimes \boldsymbol{\xi}_i^T)\boldsymbol{\Pi}\boldsymbol{\eta}_{2,i} + (\mathbf{I}_{K_{\eta_1}} \otimes \boldsymbol{\eta}_{2,i}^T)\boldsymbol{\Xi}\boldsymbol{\eta}_{2,i}].$$

Accordingly, the vector of latent variables to be integrated out is $\boldsymbol{\Upsilon}_i = (\boldsymbol{\eta}_{2,i}^T, \boldsymbol{\xi}_i^T)^T$.

Laplace Approximation

In general, the first-order Laplace approximation (Barndorff-Nielsen & Cox, 1989) to the integral $\int \exp\{g(\mathbf{t})\} d\mathbf{t}$ is:

$$(2\pi)^{K/2} \left[\det \left(-\frac{\partial^2 g(\hat{\mathbf{t}})}{\partial \mathbf{t} \partial \mathbf{t}^T} \right) \right]^{-1/2} \exp\{g(\hat{\mathbf{t}})\},$$

where \mathbf{t} is a $K \times 1$ vector and $g(\mathbf{t})$ is a unimodal function with the mode $\hat{\mathbf{t}}$. The approximation is essentially achieved by Taylor expanding $g(\mathbf{t})$ about its mode. Shun and McCullagh (1995) further derived the second-order Laplace approximation. The Laplace approximation of the observed log-likelihood is:

$$\ell_{(Lap)}(\boldsymbol{\Upsilon}; \boldsymbol{\theta}) = \sum_{i=1}^n \left[\frac{K}{2} \log(2\pi) + h_i(\boldsymbol{\Upsilon}_i; \boldsymbol{\theta}) - \frac{1}{2} \log \det(-\mathbf{H}_i(\boldsymbol{\Upsilon}_i; \boldsymbol{\theta})) + \log(1 + R_2(\boldsymbol{\Upsilon}_i; \boldsymbol{\theta})) \right], \quad (7)$$

where $K = K_\xi + K_\eta$ or $K = K_\xi + K_{\eta_2}$,

$$\mathbf{H}_i(\boldsymbol{\Upsilon}_i; \boldsymbol{\theta}) = \frac{\partial^2 h_i(\boldsymbol{\Upsilon}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\Upsilon}_i \partial \boldsymbol{\Upsilon}_i^T},$$

is the Hessian matrix, and $\boldsymbol{\Upsilon}_i$ is evaluated at $\hat{\boldsymbol{\Upsilon}}_i$, the solution of

$$\frac{\partial h_i(\boldsymbol{\Upsilon}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\Upsilon}_i} = \mathbf{0},$$

with respect to $\boldsymbol{\Upsilon}_i$ for a fixed $\boldsymbol{\theta}$. In the first-order approximation, $R_2(\boldsymbol{\Upsilon}_i; \boldsymbol{\theta}) = 0$, for any i . In the second-order approximation,

$$R_2(\boldsymbol{\Upsilon}_i; \boldsymbol{\theta}) = -\frac{1}{2} \left[\frac{1}{4} \sum_{j,l,m,r=1}^K \left(-\frac{\partial^4 h_i(\boldsymbol{\Upsilon}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\Upsilon}_j \partial \boldsymbol{\Upsilon}_l \partial \boldsymbol{\Upsilon}_m \partial \boldsymbol{\Upsilon}_r} c_{jm} c_{lr} \right) - \sum_{j,l,m,r,s,t=1}^K \frac{\partial^3 h_i(\boldsymbol{\Upsilon}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\Upsilon}_j \partial \boldsymbol{\Upsilon}_l \partial \boldsymbol{\Upsilon}_m} \frac{\partial^3 h_i(\boldsymbol{\Upsilon}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\Upsilon}_r \partial \boldsymbol{\Upsilon}_s \partial \boldsymbol{\Upsilon}_t} \left(\frac{1}{4} c_{jl} c_{mr} c_{st} + \frac{1}{6} c_{jr} c_{ls} c_{mt} \right) \right],$$

where $c_{j,m}$ is the $(j, m)^{th}$ element of the inverse of $-\mathbf{H}_i(\boldsymbol{\Upsilon}_i; \boldsymbol{\theta})$ and the approximation is second-order accurate (Shun & McCullagh, 1995). The second-order Laplace approximation has been recommended by Lee and Nelder (2001) for generalized linear-mixed models and by Jin, Noh, and Lee (2018) for confirmatory factor analysis models with ordinal data.

Adaptive Gauss–Hermite Quadrature Approximation

The Gauss–Hermite–Hermite quadrature rule approximates the integral $\int \exp\{g(\mathbf{t})\}d\mathbf{t}$ by a weighted sum of function values evaluated at pre-specified quadrature points as:

$$\sum_{k_1=1}^q \cdots \sum_{k_K=1}^q \omega_{k_1 k_2 \dots k_K}^* \exp\{g(z_{k_1}, \dots, z_{k_K})\},$$

where

$$w_{k_1 k_2 \dots k_K}^* = \prod_{j=1}^K w_{k_j} \exp\{z_{k_j}^2\},$$

with z_{k_j} being the Gauss–Hermite quadrature points, w_{k_j} being the corresponding weights, and q being the number of quadrature points per latent variable. Liu and Pierce (1994) proposed to translate and dilate the quadrature points in the weighted sum, known as the AGHQ approximation, to improve the approximation effectiveness when $g(\mathbf{t})$ is an unimodal function. The AGHQ approximated observed log-likelihood is:

$$\begin{aligned} \ell_{(AGHQ)}(\hat{\mathbf{Y}}; \boldsymbol{\theta}) &= \sum_{i=1}^n \left(\frac{K}{2} \log(2) - \frac{1}{2} \log \det(-\mathbf{H}_i(\hat{\mathbf{Y}}_i; \boldsymbol{\theta})) + \log \right. \\ &\quad \left. \sum_{k_1=1}^q \cdots \sum_{k_K=1}^q \left\{ w_{k_1 k_2 \dots k_K}^* \exp \left[h_i \left(\tilde{\mathbf{Y}}_i(\mathbf{z}_{k_1 k_2 \dots k_K}); \boldsymbol{\theta} \right) \right] \right\} \right), \end{aligned} \tag{8}$$

where

$$\tilde{\mathbf{Y}}_i(\mathbf{z}_{k_1 k_2 \dots k_K}) = \sqrt{2} \mathbf{L}_i \mathbf{z}_{k_1 k_2 \dots k_K} + \hat{\mathbf{Y}}_i$$

is the latent variable vector dilated and translated around the mode $\hat{\mathbf{Y}}_i$ with \mathbf{L}_i being the Cholesky decomposition of the inverse of the negative Hessian:

$$\mathbf{L}_i \mathbf{L}_i^T = (-\mathbf{H}_i(\mathbf{Y}_i; \boldsymbol{\theta}))^{-1},$$

and $\mathbf{z}_{k_1 k_2 \dots k_K}$ being the vector containing the k_1^{th} , k_2^{th} , ..., and the k_K^{th} quadrature points. If only one quadrature point per latent variable is used, $\ell_{(AGHQ)}(\hat{\mathbf{Y}}_i; \boldsymbol{\theta})$ is the same as the first-order approximated $\ell_{(Lap)}(\hat{\mathbf{Y}}_i; \boldsymbol{\theta})$. Let $\hat{\boldsymbol{\theta}}_{(AGHQ)}$ be the estimator of $\boldsymbol{\theta}$ that maximizes (8). Following Bianconcini (2014), Jin and Andersson (2019b) showed that the error rate of the approximation is

$$O\left((p_x + p_y)^{-\lfloor (q+2)/3 \rfloor}\right),$$

and that $\hat{\boldsymbol{\theta}}_{(AGHQ)}$ satisfies

$$\hat{\boldsymbol{\theta}}_{(AGHQ)} - \boldsymbol{\theta}_0 = O_p\left[\max\left(n^{-1/2}, (p_x + p_y)^{-\lfloor (q+2)/3 \rfloor}\right)\right],$$

where $\boldsymbol{\theta}_0$ is the true parameter vector and $\lfloor \cdot \rfloor$ takes the largest integer that is less than the enclosed value. Hence, using 5 quadrature points per latent variable yields the same error rate as the second-order Laplace approximation.

Direct Maximum Likelihood

When maximizing functions such as (7) or (8) it is important to realize that the mode $\hat{\mathbf{Y}}_i$ depends on the parameter vector $\boldsymbol{\theta}$. The gradient of the approximated observed log-likelihood (regardless of approximation method) is:

$$\frac{\partial \ell(\hat{\mathbf{Y}}(\boldsymbol{\theta}); \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big|_{\mathbf{Y}=\hat{\mathbf{Y}}} + \left(\frac{\partial \hat{\mathbf{Y}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right)^T \frac{\partial \ell(\mathbf{Y})}{\partial \mathbf{Y}} \Big|_{\mathbf{Y}=\hat{\mathbf{Y}}}, \tag{9}$$

where $\partial \ell(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ is the derivative of the observed log-likelihood function with respect to $\boldsymbol{\theta}$, treating the mode as constant. On the contrary, $\partial \ell(\mathbf{Y})/\partial \mathbf{Y}$ is the derivative of the observed log-likelihood function with respect to \mathbf{Y} , treating $\boldsymbol{\theta}$ as constant. Hence, the second term acknowledges that the mode is affected by $\boldsymbol{\theta}$, by applying the chain rule. By the implicit function theorem,

$$\frac{\partial \hat{\mathbf{Y}}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = - \left(\frac{\partial^2 h_i(\mathbf{Y}_i; \boldsymbol{\theta})}{\partial \mathbf{Y}_i \partial \mathbf{Y}_i^T} \Big|_{\mathbf{Y}_i=\hat{\mathbf{Y}}_i} \right)^{-1} \frac{\partial^2 h_i(\mathbf{Y}_i; \boldsymbol{\theta})}{\partial \mathbf{Y}_i \partial \boldsymbol{\theta}^T} \Big|_{\mathbf{Y}_i=\hat{\mathbf{Y}}_i}.$$

The sandwich estimator of standard errors can be calculated from:

$$\left(\frac{\partial^2 \widehat{\ell}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right)^{-1} \sum_{i=1}^n \left[\frac{\partial \ell_i(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \frac{\partial \ell_i(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}^T} \right] \left(\frac{\partial^2 \widehat{\ell}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right)^{-1}, \tag{10}$$

where ℓ_i is the i^{th} term in Equation (7) or (8). In the BFGS algorithm, the inverse of the Hessian matrix of $\ell(\hat{\mathbf{Y}}(\boldsymbol{\theta}); \boldsymbol{\theta})$ is approximated and updated by low-rank approximations at every iteration. The Hessian matrix can also be approximated from the Louis (1982) method. Since the exact gradient of $\ell(\hat{\mathbf{Y}}(\boldsymbol{\theta}); \boldsymbol{\theta})$ is calculated by Equation (9), the sandwich estimator (10) of standard errors can be readily computed. In contrast, neither the gradient nor the Hessian matrix is the by-product in the EM algorithm and need to be computed afterward.