British Journal of Mathematical and Statistical Psychology (2018), 71, 499–522

The British Psychological Society

www.wileyonlinelibrary.com

A penalized likelihood method for multi-group structural equation modelling

Po-Hsien Huang*

Department of Psychology, National Cheng Kung University, Taiwan

In the past two decades, statistical modelling with sparsity has become an active research topic in the fields of statistics and machine learning. Recently, Huang, Chen and Weng (2017, *Psychometrika, 82*, 329) and Jacobucci, Grimm, and McArdle (2016, *Structural Equation Modeling: A Multidisciplinary Journal, 23*, 555) both proposed sparse estimation methods for structural equation modelling (SEM). These methods, however, are restricted to performing single-group analysis. The aim of the present work is to establish a penalized likelihood (PL) method for multi-group SEM. Our proposed method decomposes each group model parameter into a common reference component and a group-specific increment component. By penalizing the increment components, the heterogeneity of parameter values across the population can be explored since the null group-specific effects are expected to diminish. We developed an expectation-conditional maximization algorithm to optimize the PL criteria. A numerical experiment and a real data example are presented to demonstrate the potential utility of the proposed method.

1. Introduction

After Tibshirani (1996) introduced L_1-penalized regression, the so-called least absolute shrinkage and selection operator (LASSO), statistical modelling with sparsity became an active research topic in the fields of statistics and machine learning (see Bühlmann & van de Geer, 2011; Fan & Lv, 2010; Wellner & Zhang, 2012; for reviews). By adding a sparsity-inducing penalty (e.g., the L_1, penalty) in the estimation criterion (e.g., a likelihood function), the resulting penalized (or regularized) estimate can have elements that are exactly zero. Through the sparsity pattern of the estimate, the relationships between the variables considered can be easily probed. Hence, the estimation results from sparse modelling have useful interpretations. Penalized estimators can outperform their unpenalized counterparts in terms of mean squared error (e.g., Knight & Fu, 2000) and achieve variable selection consistency under suitable conditions (e.g., Fan & Li, 2001; Zhao & Yu, 2006; Zou, 2006). Sparse modelling has been thought of as an effective approach to learning association patterns among a large number of variables (see Hastie, Tibshirani, & Wainwright, 2015, for a review).

Recently, the idea of sparse modelling was introduced to the field of psychometrics (e.g., Chen, Liu, Xu, & Ying, 2015; Hirose & Yamamoto, 2014, 2015; Huang, Chen, & Weng, 2017; Jacobucci, Grimm, & McArdle, 2016; Tutz & Schauberger, 2015; Zou, Choi,

*Correspondence should be addressed to Po-Hsien Huang, Department of Psychology, National Cheng Kung University, No.1, University Road, Tainan City, 701, Taiwan (email: psyphh@mail.ncku.edu.tw).

DOI:10.1111/bmsp.12130

& Oehlert, 2011). In particular, Huang *et al.* (2017) and Jacobucci *et al.* (2016) both proposed sparse estimation methods for structural equation modelling (SEM). The two methods are conceptually similar, along with their pros and cons. Under the 'all y model' without covariates (Muthén, 1984), Huang *et al.* (2017) established a penalized likelihood (PL) method. They designed an algorithm for optimizing the PL criterion with L_1, smoothly clipped absolute deviation (SCAD; Fan & Li, 2001), and minimax concave penalty (MCP; Zhang, 2010). They also described the asymptotic properties of the PL estimator. The R package *lsl* was written to implement Huang *et al.*'s PL method. On the other hand, Jacobucci *et al.* (2016) considered a methodology based on a general L_1/L_2-regularized fitting function under the reticular action model (RAM) formulation (McArdle & McDonald, 1984). The R package *regsem* implements Jacobucci *et al.*'s regularization method. Under these two sparse estimation methods, users can flexibly specify model parameters to be penalized and then obtain a final sparse estimate by choosing the penalty level. From the viewpoint of modelling capacity, the method of Jacobucci *et al.* is broad since it adopts the RAM formulation and allows other regularized estimation criteria beyond PL (e.g., regularized least squares criterion). However, Jacobucci *et al.* did not propose an appropriate algorithm to optimize the regularized criterion. Their solution relies on general-purpose optimization routines in R. Because any L_1-regularized criterion is non-differentiable, in general, Jacobucci *et al.*' method cannot efficiently find a local maximizer, especially when the model is relatively complex.

Although the existing penalization or regularization methods make a good starting point for sparse estimation in SEM, they are restricted to the case of single-group analysis. In psychological studies, understanding the heterogeneity of relationships among variables when varying the population is a crucial question. Heterogeneity is often studied through multi-group SEM (MGSEM; Jöreskog, 1971; Sörbom, 1974). By comparing the test statistics and the goodness-of-fit indices of models under different parameter constraints, the potential heterogeneity across groups can be evaluated. An important application of MGSEM is to examine the factorial invariance of psychological measurements (Meredith, 1993). In that case, researchers should specify a series of models reflecting different degrees of invariance and then chose an optimal one based on the chi-squared difference test or other fit indices (see Millsap, 2011, for a review). In our experience, MGSEM for examining heterogeneity is often conducted in an exploratory manner. MGSEM users may try a variety of models with different constraints to identify the potential heterogeneity of effects. Because sparse estimation is an efficient way to explore the sparsity pattern, establishing a PL method for MGSEM could be helpful if the heterogeneity across groups can be represented by a sparsity pattern of parameters.

The aim of the present work is to establish a PL method for MGSEM. In particular, the proposed method decomposes each group model parameter into a reference component and an increment component. The reference component is common across groups, while the increment component reflects group-specific effects. The exact meaning of the two components depends on the form of parameter constraints (see Section 2). By penalizing the increment components, the heterogeneity of parameter values across populations can be explored since the null group-specific effects diminish. Therefore, under the proposed method, it is not necessary to specify *a priori* a set of candidate models with different heterogeneity patterns. The PL will 'automatically learn' a heterogeneity pattern based on the given data and the chosen penalty level.

This paper is organized as follows. Section 2 introduces the MGSEM formulation to establish our method. In Section 3 the proposed PL method is presented. Section 4 describes an algorithm for optimizing the PL criterion. In Sections 5 and 6 a numerical

experiment and a real data illustration for exploring partial invariance are discussed. Finally, the merits and limitations of the current work are discussed.

Before describing the proposed method, some notation is introduced (see also Huang *et al.*, 2017). Given a P-dimensional vector x, we use $x[j]$ to denote the jth coordinate of x. Let A be a $P \times M$ matrix with elements a_{pm}. $A[j,]$, $A[,k]$, and $A[j,k]$ are used to denote the jth row, the kth column, and the (j,k)th element of A. Similarly, $A[-j,]$, $A[,-k]$, and $A[-j,-k]$ respectively represent the submatrices of A with $A[j,]$, $A[,k]$, and both $A[j,]$ and $A[,k]$ deleted. When A^{-1} exists, a^{jk} denotes the (j,k)th element of A^{-1}.

2. Multi-group structural equation modelling formulation

The proposed PL method is established using the following RAM formulation for MGSEM. For group g, let η_g denote a $(P+M)$-dimensional random vector with elements η_{gi}. We partition η_g into two parts, v_g and f_g, where v_g is a P-dimensional random vector of the observed variables and f_g is an M-dimensional random vector of the latent factors. The MGSEM model incorporates the following linear equation for η_g:

$$\eta_g = \alpha_g + \mathrm{B}_g \eta_g + \zeta_g, \tag{1}$$

where α_g is the $(P+M)$-dimensional intercept vector, B_g is the $(P+M) \times (P+M)$ regression coefficient matrix, and ζ_g is the $(P+M)$-dimensional random residual vector with mean zero and covariance matrix Φ_g. Since $\eta_g = (v_g, f_g)$, equation (1) can be rewritten as

$$\begin{pmatrix} v_g \\ f_g \end{pmatrix} = \begin{pmatrix} \alpha_g^{(v)} \\ \alpha_g^{(f)} \end{pmatrix} + \begin{pmatrix} \mathrm{B}_g^{(vv)} & \mathrm{B}_g^{(vf)} \\ \mathrm{B}_g^{(fv)} & \mathrm{B}_g^{(ff)} \end{pmatrix} \begin{pmatrix} v_g \\ f_g \end{pmatrix} + \begin{pmatrix} e_g \\ d_g \end{pmatrix}, \tag{2}$$

where $\alpha_g^{(v)}$ and $\alpha_g^{(f)}$ denote the intercepts of v_g and f_g respectively, and $\mathrm{B}_g^{(vv)}$, $\mathrm{B}_g^{(vf)}$, $\mathrm{B}_g^{(fv)}$, and $\mathrm{B}_g^{(ff)}$ are the regression coefficient matrices that describe different relations among those observed variables and latent factors. In particular, $\mathrm{B}_g^{(vf)}$ is a $P \times M$ loading matrix in factor analysis and $\mathrm{B}_g^{(ff)}$ is an $M \times M$ path coefficients matrix among the latent variables. The covariance matrix of ζ_g can be also partitioned into

$$\Phi_g = \begin{pmatrix} \Phi_g^{(ee)} & \Phi_g^{(ed)} \\ \Phi_g^{(de)} & \Phi_g^{(dd)} \end{pmatrix}.$$

In almost all SEM applications, $\Phi_g^{(de)} = \Phi_g^{(ed)^T}$ is set to zero to avoid the identifiability problem.

Let I denote an identity matrix of appropriate size. Under the existence of $(I - \mathrm{B}_g)^{-1}$, the model-implied mean vector and the covariance matrix of η_g are

$$\mu_g^{(\eta)}(\theta_g) = (I - \mathrm{B}_g)^{-1} \alpha_g \tag{3}$$

and

$$\Sigma_g^{(\eta\eta)}(\theta_g) = (I - \mathrm{B}_g)^{-1}\Phi_g(I - \mathrm{B}_g)^{-1^T}, \tag{4}$$

respectively. For group g, θ_g denotes a Q_g-dimensional model parameter vector with elements θ_{gq}. The model parameter vector θ_g contains all unknown and all constrained non-zero parameters from the model parameter matrices, including α_g, B_g, and Φ_g. When collecting these model parameters into θ_g, we utilize a slightly different parameterization for ϕ_{gjj}, the diagonal elements of Φ_g, from the usual SEM. Instead of using ϕ_{gjj}, the conditional variance parameters φ_{gjj} are considered:

$$\varphi_{gjj} = \phi_{gjj} - \Phi_g[j, -j]\Phi_g[-j, -j]^{-1}\Phi_g[-j, j].$$

If ζ_{gj} is uncorrelated with all other ζ_{gi} $(i \neq j)$, we have $\varphi_{gjj} = \phi_{gjj}$. Under this parameterization, penalized covariance coefficients can be easily optimized by a modified iterative conditional fitting method (Chaudhuri, Drton, & Richardson, 2007).

The current framework assumes that the model structures for the G groups are all identical, that is, $\mu_g^{(\eta)}(\cdot) = \mu^{(\eta)}(\cdot)$ and $\Sigma_g^{(\eta\eta)}(\cdot) = \Sigma^{(\eta\eta)}(\cdot)$. As a result, the dimensions of θ_1, $\theta_2, \ldots, \theta_G$ are all equal to Q, while $\theta_{1q}, \theta_{2q}, \ldots, \theta_{Gq}$ represent the same element of a given model parameter matrix. For each g and q, our method reparameterizes θ_{gq} as a sum of a reference component $\underline{\theta}_q$ and an increment component $\underline{\theta}_{gq}$, that is, $\theta_{gq} = \underline{\theta}_q + \underline{\theta}_{gq}$. Thus we can represent the three parameter matrices as $\alpha_g = \underline{\alpha} + \underline{\alpha}_g, \mathrm{B}_g = \underline{\mathrm{B}} + \underline{\mathrm{B}}_g$ and $\Phi_g = \underline{\Phi} + \Phi_g$.[1] By the fact that $\theta_{gq} = \underline{\theta}_q$ if and only if $\underline{\theta}_{gq} = 0$ for all g, any model parameter is homogeneous across the G populations when all corresponding increment components are zero. Therefore, the heterogeneity patterns can be efficiently identified if we develop a sparse estimation method for the increment components.

Let $\underline{\theta} = (\underline{\theta}_1, \underline{\theta}_2, \ldots, \underline{\theta}_Q)$ and $\underline{\theta}_g = (\underline{\theta}_{g1}, \underline{\theta}_{g2}, \ldots, \underline{\theta}_{gQ})$ be the vector of all reference components and the vector of all increment components for group g, respectively. Our parameterization implies that $\theta_g = \underline{\theta} + \underline{\theta}_g$. We use $\theta = \{\underline{\theta}, \underline{\theta}_1, \underline{\theta}_2, \ldots, \underline{\theta}_G\}$ to denote the whole parameter vector of dimension $Q_+ = (G + 1)Q$. To determine the meaning of the reference components and the increment components, we must restrict enough elements in θ to be zero. Let the parameter θ_{gq} denote the qth parameter in group g. If its reference component $\underline{\theta}_q$ equals zero, then each increment component is equal to its corresponding group parameter, $\underline{\theta}_{gq} = \theta_{gq}$. On the other hand, if an increment component $\underline{\theta}_{jq}$ equals zero, then the corresponding reference component is equal to the model parameter of group j, $\underline{\theta}_q = \theta_{jq}$, and the increment component of group g $(g \neq j)$ now represents the difference in the model parameter between group g and j, $\underline{\theta}_{gq} = \theta_{gq} - \theta_{jq}$. By examining the sparsity of $\underline{\theta}_{gq}$, the heterogeneity between groups g and j can be easily identified. In the proposed method, the latter type of parameter constraint is generally adopted.

For example, suppose that $\beta_{1pm}^{(vf)}$ and $\beta_{2pm}^{(vf)}$ represent the factor loadings from the mth factor to the pth observed variable of group 1 and 2, respectively. Our parameterization implies that $\beta_{1pm}^{(vf)} = \underline{\beta}_{pm}^{(vf)} + \underline{\beta}_{1pm}^{(vf)}$ and $\beta_{2pm}^{(vf)} = \underline{\beta}_{pm}^{(vf)} + \underline{\beta}_{2pm}^{(vf)}$. Here, $\underline{\beta}_{pm}^{(vf)}$ is the reference component of the parameter considered, and $\underline{\beta}_{1pm}^{(vf)}$ along with $\underline{\beta}_{2pm}^{(vf)}$ denote the corresponding increment components. If $\underline{\beta}_{1pm}^{(vf)} = 0$, we have $\beta_{1pm}^{(vf)} = \underline{\beta}_{pm}^{(vf)}$ and

[1] In the current framework, $\underline{\phi}_{jj}$, the jth diagonal element of $\underline{\Phi}$, is restricted to be zero and hence ϕ_{gjj}, the jth diagonal element of $\underline{\Phi}_g$, can simply be set to Φ_{gjj}.

$\underline{\beta}_{2pm}^{(vf)} = \beta_{2pm}^{(vf)} - \beta_{1pm}^{(vf)}$. Hence, $\underline{\beta}_{2pm}^{(vf)} = 0$ implies that $\beta_{2pm}^{(vf)} = \beta_{1pm}^{(vf)}$, that is, the loading parameter is invariant across the two groups. The proposed PL can treat $\underline{\beta}_{2pm}^{(vf)}$ as a penalized parameter to explore the group heterogeneity of this loading. Heterogeneity is found if the PL estimate of $\underline{\beta}_{2pm}^{(vf)}$ does not equal zero.

Because v_g can be written as $v_g = G^{(v)}\eta_g$ for an appropriate selection matrix $G^{(v)}$, the model-implied mean and covariance structure of v_g, simply denoted by $\mu(\theta_g)$ and $\Sigma(\theta_g)$, are $G^{(v)}\,\mu^{(\eta)}\,(\theta_g)$ and $G^{(v)}\Sigma^{(\eta\eta)}(\theta_g)G^{(v)^T}$, respectively. In SEM applications, $\mu(\theta_g)$ and $\Sigma(\theta_g)$ are the quantities we wish to evaluate.

3. A penalized likelihood method for multi-group structural equation modelling

Let $\mathcal{V} = \{\{v_{gn}\}_{n=1}^{N_g}\}_{g=1}^{G}$ denote a random sample from the given G populations, where N_g denotes the sample size of group g. In our PL method, we consider maximizing the PL criterion

$$\mathcal{U}(\theta, \lambda) = \mathcal{L}(\theta) - \mathcal{R}(\theta, \lambda), \tag{5}$$

under a simple constraint of $\mathcal{C}(\theta) = 0$, where $\mathcal{L}(\theta)$ is the normal log-likelihood function, $\mathcal{R}(\theta, \lambda)$ is a penalty term, and λ is a regularization parameter. Specifically, the form of the log-likelihood function is

$$\begin{aligned}\mathcal{L}(\theta) = &-\frac{1}{2}\sum_{g=1}^{G} w_g\left[\log\left|\Sigma(\theta_g)\right| + \mathrm{tr}\left(\Sigma(\theta_g)^{-1}S_g\right)\right] \\ &-\frac{1}{2}\sum_{g=1}^{G} w_g(m_g - \mu(\theta_g))^T\Sigma(\theta_g)^{-1}(m_g - \mu(\theta_g)),\end{aligned} \tag{6}$$

where $m_g = \frac{1}{N_g}\sum_{n=1}^{N_g} v_{gn}$, $S_g = \frac{1}{N_g}\sum_{n=1}^{N_g}(v_{gn} - m_g)(v_{gn} - m_g)^T$, and $w_g = \frac{N_g}{N_+}$, with $N_+ = \sum_{g=1}^{G} N_g$. The penalty term is the sum of individual penalty functions with the structure

$$\mathcal{R}(\theta, \lambda) = \sum_{q=1}^{Q} c_{\underline{\theta}_q}\rho\left(\left|\underline{\theta}_q\right|, \lambda\right) + \sum_{g=1}^{G}\sum_{q=1}^{Q} c_{\underline{\theta}_{gq}}\rho\left(\left|\underline{\theta}_{gq}\right|, \lambda\right), \tag{7}$$

where $\rho(|\vartheta|, \lambda)$ is a non-negative penalty function for parameter ϑ and c_ϑ is the penalty indicator for ϑ. To obtain a sparse estimate, three types of penalty function are considered: L_1, SCAD (Fan & Li, 2001), and MCP (Zhang, 2010). The forms of the three functions can be found in Table 1. The value of penalty indicator is either 1 or 0. If its value is 1, the corresponding parameter is a penalized parameter; otherwise the parameter is freely estimated without penalty. In the current work, only a conditional variance parameter falls beyond the above two categories, that is, it can only be set as a freely estimated parameter with the identification constraint $\underline{\varphi}_{jj} = 0$ for each reference component.

The constraint function $\mathcal{C}(\theta) = 0$ restricts some elements of θ to have fixed specified values. For model identification, two types of constraints must be imposed: (1) we should choose an indicator for any given latent factor and fix the value of the corresponding loading for scale setting; (2) constrain enough elements in θ to zero to be able to determine the meaning of the reference components and increment components.

Table 1. Mathematical expressions for L_1, SCAD and MCP

Penalty function	Mathematical expressions
L_1	$\rho_{L_1}(\lvert\vartheta\rvert,\lambda)=\lambda\lvert\vartheta\rvert$
SCAD	$\rho_{\mathrm{SCAD}}(\lvert\vartheta\rvert,\lambda)=\begin{cases}\lambda\lvert\vartheta\rvert & \text{if } \lvert\vartheta\rvert\le\lambda\\ -\frac{\lvert\vartheta\rvert^2+\lambda^2-2\lambda\delta\lvert\vartheta\rvert}{2(\delta-1)} & \text{if } \lambda<\lvert\vartheta\rvert\le\lambda\delta\\ \frac{\lambda^2(\delta^2-1)}{2(\delta-1)} & \text{if } \lambda\delta<\lvert\vartheta\rvert\end{cases}$
MCP	$\rho_{\mathrm{MCP}}(\lvert\vartheta\rvert,\lambda)=\begin{cases}\lambda\lvert\vartheta\rvert-\frac{\vartheta^2}{2\delta} & \text{if } \lvert\vartheta\rvert\le\lambda\delta\\ \frac{1}{2}\lambda^2\delta & \text{if } \lambda\delta<\lvert\vartheta\rvert\end{cases}$

Note. ϑ is a real-valued model parameter. λ is a non-negative regularization parameter. δ is a shape parameter specific to SCAD and MCP. The lower bounds of δ for SCAD and MCP depend on both data and model.

Consider an example of a one-factor model of two groups. If we hope to understand whether a set of loadings $\{\beta_{gp1}^{(vf)}\}_{p=1}^{P}$ is invariant between two groups, then we (1) restrict $\beta_{111}^{(vf)}=\beta_{211}^{(vf)}=1$ and $\underline{\beta}_{111}^{(vf)}=\underline{\beta}_{211}^{(vf)}=0$ for scale setting; (2) constrain $\underline{\beta}_{1p1}^{(vf)}=0$, for $p=2,3,\ldots,P$, to determine the meaning of the reference components; (3) set the penalty term to $\mathcal{R}(\theta,\lambda)=\sum_{p=2}^{P}c_{\underline{\beta}_{2p1}^{(vf)}}\rho\big(\lvert\underline{\beta}_{2p1}^{(vf)}\rvert,\lambda\big)$. The heterogeneity of these loadings can be identified by examining the sparsity pattern of the PL estimates for $\{\underline{\beta}_{2p1}^{(vf)}\}_{p=2}^{P}$. Note that the choice of measurement for scale setting is crucial. When the chosen measurement is not invariant, the proposed method may fail (see also Johnson, Meade, & DuVernet, 2009).

A PL estimate of θ under λ, denoted by $\hat{\theta}\equiv\hat{\theta}(\lambda)$, is defined as a local maximizer for $\mathcal{U}(\theta,\lambda)$. Note that different values of λ may result in different PL estimates since $\hat{\theta}$ is a function of λ. In practice, an optimal value of λ can be chosen from a pre-specified candidate set $\Lambda=\{\lambda_1,\lambda_2,\ldots,\lambda_J\}$ via the Akaike information criterion (AIC; Akaike, 1974),

$$\mathrm{AIC}(\lambda)=\mathcal{D}(\hat{\theta})+\frac{2}{N_+}e(\lambda), \tag{8}$$

or via the Bayesian information criterion (BIC; Schwarz, 1978),

$$\mathrm{BIC}(\lambda)=\mathcal{D}(\hat{\theta})+\frac{\log(N_+)}{N_+}e(\lambda), \tag{9}$$

where $\mathcal{D}(\theta)$ is the maximum likelihood (ML) discrepancy function defined as

$$\begin{aligned}\mathcal{D}(\theta)=&\sum_{g=1}^{G}w_g\big[\mathrm{tr}\big(S_g\Sigma(\theta_g)^{-1}\big)-\log\big\lvert S_g\Sigma(\theta_g)^{-1}\big\rvert-P\big]\\&+\sum_{g=1}^{G}w_g(m_g-\mu(\theta_g))^{T}\Sigma(\theta_g)^{-1}(m_g-\mu(\theta_g))\end{aligned} \tag{10}$$

and $e(\lambda)$ denotes the number of effective parameters. In the current framework, $e(\lambda)$ is just the number of distinct non-zero elements in the PL estimator $\hat{\theta}$ under penalty level λ. It has

been shown that the BIC can yield a consistent selection result for the quasi-true model. In contrast, the AIC can only choose a model that attains minimum ML discrepancy with respect to the given Λ (Huang *et al.*, 2017). If SCAD or MCP is implemented, it is possible to simultaneously choose λ and δ from $\Lambda \times \Delta$ with $\Delta = \delta_1, \delta_2, \ldots, \delta_K$.

Let $\hat{\lambda}$ be the selected value of the regularization parameter. The final PL estimate for the parameter vector is denoted by $\hat{\theta}(\hat{\lambda})$. We may then obtain $\hat{\theta}_g(\hat{\lambda})$, the estimated parameter vector for group g, via $\underline{\hat{\theta}}(\hat{\lambda}) + \underline{\hat{\theta}}_g(\hat{\lambda})$. The final PL estimates for the three parameter matrices can be obtained in a similar manner.

After obtaining the final PL estimate, the model-implied covariance and mean for each group can be derived via $\hat{\Sigma}_g = \Sigma\big(\hat{\theta}_g(\hat{\lambda})\big)$ and $\hat{\mu}_g = \mu\big(\hat{\theta}_g(\hat{\lambda})\big)$. Hence, the appropriateness of the final model can be evaluated by examining the discrepancy between $\{\hat{\Sigma}_g, \hat{\mu}_g\}_{g=1}^G$ and $\{S_g, m_g\}_{g=1}^G$. Goodness-of-fit indices can be calculated in a similar manner. For example, the multi-group version of the root mean square error of approximation (RMSEA; Steiger, 1998) can be calculated as

$$\text{RMSEA} = \sqrt{G \times \max\left\{\frac{\mathcal{D}\left(\hat{\theta}(\hat{\lambda})\right)}{df(\hat{\lambda})} - \frac{1}{N_+}, 0\right\}}, \tag{11}$$

where $df(\hat{\lambda})$ denotes the degrees of freedom under penalty level $\hat{\lambda}$,

$$df(\hat{\lambda}) = \frac{GP(P+3)}{2} - e(\hat{\lambda}).$$

4. An expectation-conditional maximization algorithm

In this section we describe an expectation-conditional maximization (ECM) algorithm (Meng & Rubin, 1993) to optimize the PL criterion. The ECM algorithm is an iterative method composed of one E-step and several CM-steps. It is a variant of the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) and has been used to solve the PL problem (Huang *et al.*, 2017). At each iteration of the ECM algorithm, an updated PL estimate, denoted by $\hat{\theta}^{(t)} \equiv \hat{\theta}^{(t)}(\lambda)$, is calculated. The iteration continues until $\left|\left|\hat{\theta}^{(t+1)} - \hat{\theta}^{(t)}\right|\right| < \epsilon$ for some required $\epsilon > 0$.

4.1. E-step

The E-step calculates $\mathcal{M}(\theta|\hat{\theta}^{(t)}) = \mathbb{E}[\mathcal{U}^C(\theta, \lambda)|\mathcal{V}, \hat{\theta}^{(t)}]$, the conditional expectation of the complete-data PL given the random sample $\mathcal{V}$ and the current parameter estimate $\hat{\theta}^{(t)}$, where $\mathcal{U}^C(\theta, \lambda)$ is the PL criterion that treats latent factors as observable but missing. The complete-data PL can be written as

$$\mathcal{U}^C(\theta, \lambda) = -\frac{1}{2}\sum_{g=1}^{G} w_g \left[\log|\Phi_g| + \frac{1}{N_g}\sum_{n=1}^{N_g} \zeta_{gn}^T \Phi_g^{-1} \zeta_{gn}\right] - \mathcal{R}(\theta, \lambda), \tag{12}$$

where $\zeta_{gn} = \eta_{gn} - \alpha_g - B_g \eta_{gn}$. Given two sets of random samples $\{v_{gn}\}_{n=1}^{N_g}$ and $\{w_{gn}\}_{n=1}^{N_g}$, we define $e_g^{(v)(t)} = \mathbb{E}\big[\frac{1}{N_g}\sum_{n=1}^{N_g} v_{gn}|\mathcal{V}, \hat{\theta}^{(t)}\big]$ and $C_g^{(vw)(t)} = \mathbb{E}\big[\frac{1}{N_g}\sum_{n=1}^{N_g} v_{gn} w_{gn}^T|\mathcal{V}, \hat{\theta}^{(t)}\big]$.

By the functional form of $\mathcal{U}^C(\theta, \lambda)$ and by our parameterization, it suffices to calculate $e_g^{(\eta)} \equiv e_g^{(\eta)(t)}$ and $C_g^{(\eta\eta)} \equiv C_g^{(\eta\eta)(t)}$ to obtain $\mathcal{M}(\theta|\hat{\theta}^{(t)})$ (see Appendix A for the derivation).

4.2. CM-steps

To perform CM at iteration step $t + 1$ for $\underline{\theta}_q$ or $\underline{\theta}_{gq}$, we try to find $\hat{\underline{\theta}}_q^{(t+1)} \equiv \hat{\underline{\theta}}_q^{(t+1)}(\lambda)$ and $\hat{\underline{\theta}}_{gq}^{(t+1)} \equiv \hat{\underline{\theta}}_{gq}^{(t+1)}(\lambda)$, the term that maximizes $\mathcal{M}(\theta|\hat{\theta}^{(t)})$ with all other parameters fixed at their updated values. With no penalty ($\lambda = 0$), the formula for each $\hat{\underline{\theta}}_q^{(t+1)}(0)$ and for each $\hat{\underline{\theta}}_{gq}^{(t+1)}(0)$ can be found in Table 2 (see Appendix B for the derivation). If the penalty level is not zero, a shrinkage step for $\hat{\underline{\theta}}_q^{(t+1)}(0)$ and $\hat{\underline{\theta}}_{gq}^{(t+1)}(0)$ is necessary to obtain $\hat{\underline{\theta}}_q^{(t+1)}(\lambda)$ and $\hat{\underline{\theta}}_{gq}^{(t+1)}(\lambda)$. The shrinkage formulae can be found in Table 3 (see also Huang *et al.*, 2017). If the value of a parameter is constrained, it should remain fixed at each iteration step.

5. Numerical experiment

In this section, a numerical experiment is conducted to evaluate the performance of the proposed PL method for identifying the pattern of partial factorial invariance. Within a two-group factor analysis model, the mean and covariance structures for the measurement v_g ($g = 1, 2$) are

$$\mu(\theta_g) = \alpha_g^{(v)} + B_g^{(vf)}\alpha_g^{(f)} \tag{13}$$

and

$$\Sigma(\theta_g) = B_g^{(vf)}\Phi_g^{(dd)}B_g^{(vf)^T} + \Phi_g^{(ee)}, \tag{14}$$

respectively. A measurement v_g is said to satisfy the so-called weak factorial invariance (or metric invariance) condition if $B_g^{(vf)} = B^{(vf)}$ (Widaman & Reise, 1997). When both $\alpha_g^{(v)} = \alpha^{(v)}$ and $B_g^{(vf)} = B^{(vf)}$ hold, we say that the measurement v_g satisfies the strong factorial invariance (or scalar invariance) condition (Meredith, 1993). Under strong factorial invariance, the observed differences in test scores across groups can be attributed to differences in the latent attributes. If strong factorial invariance is violated, we may try to identify the measurements or items that are not invariant across groups. The condition of partial factorial invariance applies to this case (Byrne, Shavelson, & Muthén, 1989). Because previous simulations mainly focused on the examination of weak factorial invariance (e.g., French & Finch, 2006; Meade & Bauer, 2007; Yoon & Millsap, 2007), our numerical experiment also considers that case only, that is, only the heterogeneity of factor loadings across groups is explored.

In the current simulation, the size of differences across groups (null, small, medium, and large) and the sample sizes (200, 400, 600, 800, and 1,000) are manipulated. The population model for data generation is a single-factor model with 12 measured variables. The parameter values are assumed to be invariant across groups except for loadings and measurement error variances. In all conditions, the loading matrix and the error covariance matrix for group 1 are set up as

Table 2. Updating formulae for the reference and the increment components under no penalty

Parameter and updating formula	Working weight
$\hat{\underline{\alpha}}_j^{(t+1)}(0) = w_{\underline{\alpha}_j}^{(t+1)}\left[\sum_{g=1}^{G} w_g\hat{\phi}_g^{jj}\left(e_g^{(\eta)}[j] - \underline{\hat{\alpha}}_{gj} - \hat{\mathrm{B}}_g[j,]e_g^{(\eta)}\right) + \sum_{g=1}^{G} w_g \sum_{i\neq j}\hat{\phi}_g^{ji}\left(e_g^{(\eta)}[i] - \hat{\alpha}_{gi} - \hat{\mathrm{B}}_g[i,]e_g^{(\eta)}\right)\right]$	$w_{\underline{\alpha}_j}^{(t+1)} = \frac{1}{\sum_{g=1}^{G} w_g\hat{\phi}_g^{jj}}$
$\hat{\underline{\alpha}}_{gj}^{(t+1)}(0) = w_{\underline{\alpha}_{gj}}^{(t+1)}\left[w_g\hat{\phi}_g^{jj}\left(e_g^{(\eta)}[j] - \underline{\hat{\alpha}}_j - \hat{\mathrm{B}}_g[j,]e_g^{(\eta)}\right) + w_g\sum_{i\neq j}\hat{\phi}_g^{ji}\left(e_g^{(\eta)}[i] - \hat{\alpha}_{gi} - \hat{\mathrm{B}}_g[i,]e_g^{(\eta)}\right)\right]$	$w_{\underline{\alpha}_{gj}}^{(t+1)} = \frac{1}{w_g\hat{\phi}_g^{jj}}$
$\hat{\underline{\beta}}_{jk}^{(t+1)}(0) = w_{\underline{\beta}_{jk}}^{(t+1)}\left[\sum_{g=1}^{G} w_g\hat{\phi}_g^{jj}\left(C_g^{(\eta\eta)}[j,k] - e_g^{(\eta)}[k]\hat{\alpha}_{gj} - \underline{\hat{\mathrm{B}}}[j,-k]C_g^{(\eta\eta)}[-k,k] - \underline{\hat{\mathrm{B}}}_g[j,]C_g^{(\eta\eta)}[,k]\right)\right.$ $\left. + \sum_{g=1}^{G} w_g\sum_{i\neq j}\hat{\phi}_g^{ji}\left(C_g^{(\eta\eta)}[i,k] - e_g^{(\eta)}[k]\hat{\alpha}_{gi} - (\underline{\hat{\mathrm{B}}}[i,] + \underline{\hat{\mathrm{B}}}_g[i,])C_g^{(\eta\eta)}[,k]\right)\right]$	$w_{\underline{\beta}_{jk}}^{(t+1)} = \frac{1}{\sum_{g=1}^{G} w_g\hat{\phi}_g^{jj}C_g^{(\eta\eta)}[k,k]}$
$\hat{\underline{\beta}}_{gjk}^{(t+1)}(0) = w_{\underline{\beta}_{gjk}}^{(t+1)}\left[w_g\hat{\phi}_g^{jj}\left(C_g^{(\eta\eta)}[j,k] - e_g^{(\eta)^{(t+1)}}[k]\hat{\alpha}_{gj} - \hat{\mathrm{B}}[j,]C_g^{(\eta\eta)}[,k] - \underline{\hat{\mathrm{B}}}_g[j,-k]C_g^{(\eta\eta)}[-k,k]\right)\right.$ $\left. + w_g\sum_{i\neq j}\hat{\phi}_g^{ji}\left(C_g^{(\eta\eta)}[i,k] - e_g^{(\eta)^{(t+1)}}[k]\hat{\alpha}_{gi} - (\hat{\mathrm{B}}[i,] + \underline{\hat{\mathrm{B}}}_g[i,])C_g^{(\eta\eta)}[,k]\right)\right]$	$w_{\underline{\beta}_{gjk}}^{(t+1)} = \frac{1}{w_g\hat{\phi}_g^{jj}C_g^{(\eta\eta)}[k,k]}$
$\hat{\underline{\phi}}_{jk}^{(t+1)}(0) = w_{\underline{\phi}_{jk}}^{(t+1)}\sum_{g=1}^{G}\frac{w_g}{\phi_{gjj}}C_g^{(\tilde{\zeta}\tilde{\zeta})}[l_k,j] - \underline{\hat{\Phi}}[j,-(j,k)]C_g^{(\tilde{\zeta}^{(j)}\tilde{\zeta}^{(j)})}[-l_k,l_k] - \underline{\hat{\Phi}}_g[j,-j]C_g^{(\tilde{\zeta}^{(j)}\tilde{\zeta}^{(j)})}[,l_k])$	$w_{\underline{\phi}_{jk}}^{(t+1)} = \frac{1}{\sum_{g=1}^{G}\frac{w_g}{\phi_{gjj}}C_g^{(\tilde{\zeta}^{(j)}\tilde{\zeta}^{(j)})}[l_k,l_k]}$
$\hat{\underline{\phi}}_{gjk}^{(t+1)}(0) = w_{\underline{\phi}_{gjk}}^{(t+1)}\frac{w_g}{\phi_{gjj}}(C_g^{(\tilde{\zeta}\tilde{\zeta})}[l_k,j] - \underline{\hat{\Phi}}[j,-j]C_g^{(\tilde{\zeta}^{(j)}\tilde{\zeta}^{(j)})}[,l_k] - \underline{\hat{\Phi}}_g[j,-(j,k)]C_g^{(\tilde{\zeta}^{(j)}\tilde{\zeta}^{(j)})}[-l_k,l_k])$	$w_{\underline{\phi}_{gjk}}^{(t+1)} = \frac{1}{\frac{w_g}{\phi_{gjj}}C_g^{(\tilde{\zeta}^{(j)}\tilde{\zeta}^{(j)})}[l_k,l_k]}$
$\hat{\underline{\phi}}_{gjj}^{(t+1)}(0) = C_g^{(\zeta\zeta)}[j,j] - 2\hat{\Phi}_g[j,-j]C_g^{(\tilde{\zeta}\tilde{\zeta})}[,j] + \hat{\Phi}_g[j,-j]C_g^{(\tilde{\zeta}^{(j)}\tilde{\zeta}^{(j)})}\hat{\Phi}_g[-j,j]$	None

Note. For all types of parameters, $j \neq k$ is assumed. $\hat{\underline{\phi}}_{gjj}^{(t+1)}$ is derived under the constraint $\varphi_{jj} = 0$. l_i denotes the column index of φ_{gji} in $\Phi_g[j,-j]$.

Table 3. Updating formulae for the model parameter under penalty

Penalty	Final updating formula with shrinkage
L_1	$\hat{\vartheta}^{(t+1)}(\lambda) = S\big(\hat{\vartheta}^{(t+1)}, c_\vartheta w_\vartheta^{(t+1)}\lambda\big)$
SCAD	$\hat{\vartheta}^{(t+1)}(\lambda) = \begin{cases} S\big(\hat{\vartheta}^{(t+1)}, c_\vartheta w_\vartheta^{(t+1)}\lambda\big) & \text{if } \big\lvert\hat{\vartheta}^{(t+1)}\big\rvert \le \lambda\big(1 + c_\vartheta w_\vartheta^{(t+1)}\big) \\ S\big(\hat{\vartheta}^{(t+1)}, \frac{c_\vartheta w_\vartheta^{(t+1)}\lambda\delta}{\delta-1}\big)\big(1 - \frac{w_\vartheta^{(t+1)}c_\vartheta}{\delta-1}\big)^{-1} & \text{if } \lambda\big(1 + c_\vartheta w_\vartheta^{(t+1)}\big) < \big\lvert\hat{\vartheta}^{(t+1)}\big\rvert \le \lambda\delta \\ \hat{\vartheta}^{(t+1)} & \text{if } \lambda\delta < \big\lvert\hat{\vartheta}^{(t+1)}\big\rvert \end{cases}$
MCP	$\hat{\vartheta}^{(t+1)}(\lambda) = \begin{cases} S\big(\hat{\vartheta}^{(t+1)}, c_\vartheta w_\vartheta^{(t+1)}\lambda\big)\big(1 - \frac{w_\vartheta^{(t+1)}c_\vartheta}{\delta}\big)^{-1} & \text{if } \big\lvert\hat{\vartheta}^{(t+1)}\big\rvert \le \lambda\delta \\ \hat{\vartheta}^{(t+1)} & \text{if } \lambda\delta < \big\lvert\hat{\vartheta}^{(t+1)}\big\rvert \end{cases}$

Note. $\hat{\vartheta}^{(t+1)}$ is the updated value of parameter ϑ at iteration $t + 1$ with no penalty ($\lambda = 0$ or $c_\vartheta = 0$). c_ϑ is the penalty indicator of parameter ϑ. $w_\vartheta^{(t+1)}$ is the working weight of ϑ at iteration $t + 1$. Detailed formulae for both $\hat{\vartheta}^{(t+1)}$ and $w_\vartheta^{(t+1)}$ can be found in Table 2. $S(\vartheta, \lambda) = \text{sign}(\vartheta)\max\{|\vartheta| - \lambda, 0\}$ is the soft-threshold operator.

$$
\begin{aligned}
\mathrm{B}_1^{(vf)} &= (1\ 1)^T \otimes (.8\,.7\,.6\,.6\,.7\,.8)^T, \\
\Phi_1^{(ee)} &= \mathrm{diag}(1, 1) \otimes \mathrm{diag}(.36,\ .51,\ .64,\ .64,\ .51,\ .36),
\end{aligned}
$$

respectively, where $\otimes$ denotes the Kronecker product. The loading matrix and the error covariance matrix for group 2 in each condition are set up as

$$
\begin{aligned}
\text{Null}\ : \mathrm{B}_2^{(vf)} &= (1\ 1)^T \otimes (.8\,.7\,.6\,.6\,.7\,.8)^T, \\
\Phi_2^{(ee)} &= \mathrm{diag}(1,\ 1) \otimes \mathrm{diag}(.36,\ .51,\ .64,\ .64,\ .51,\ .36); \\
\text{Small}\ : \mathrm{B}_2^{(vf)} &= (1\ 1)^T \otimes (.8\,.7\,.5\,.5\,.6\,.7)^T, \\
\Phi_2^{(ee)} &= \mathrm{diag}(1,\ 1) \otimes \mathrm{diag}(.36,\ .51,\ .75,\ .75,\ .64,\ .51); \\
\text{Medium}\ : \mathrm{B}_2^{(vf)} &= (1\ 1)^T \otimes (.8\,.7\,.4\,.4\,.5\,.6)^T, \\
\Phi_2^{(ee)} &= \mathrm{diag}(1,\ 1) \otimes \mathrm{diag}(.36,\ .51,\ .84,\ .84,\ .75,\ .64); \\
\text{Large}\ : \mathrm{B}_2^{(vf)} &= (1\ 1)^T \otimes (.8\,.7\,.3\,.3\,.4\,.5)^T, \\
\Phi_2^{(ee)} &= \mathrm{diag}(1,\ 1) \otimes \mathrm{diag}(.36,\ .51,\ .91,\ .91,\ .84,\ .75).
\end{aligned}
$$

Across the two groups, the intercept terms are all set to 0 and the factor variances are assumed to be 1. Each data set is generated from the multivariate normal distribution with zero mean and the corresponding covariance structure. The simulation is conducted in the R environment (R Core Team, 2017). The code is available from the author on request. When implementing the proposed PL, the intercept and the variance parameters are all freely estimated without assuming invariance. The loadings are estimated as follows: (1) set the loading for the first measurement at .8 for scaling; (2) specify the remaining reference components as free parameters; (3) restrict the remaining increment components of group 1 to fixed zero parameters; (4) treat the remaining increment components of group 2 as penalized parameters. Hence, the heterogeneity of loadings across groups can be probed by examining the sparsity pattern of the increment

component estimates from group 2. Because L_1 can be seen as a special case of SCAD and MCP with infinite δ, and the theoretical properties and the empirical performances of SCAD and MCP are similar (Huang *et al.*, 2017), only MCP is implemented for the current simulation. For each replication, an optimal pair of λ and δ is chosen based on the value of the AIC or BIC from $\Lambda \times \Delta$, where $\Lambda = \{.01, .02, \ldots, .60\}$ and $\Delta = \{1.5, 2.5, 3.5, \infty\}$. The number of successful replications under each condition is set to 500. A replication is said to be successful if the ECM algorithm converges for each considered pair of λ and δ below $\epsilon = 10^{-6}$ within $t < 1{,}000$ iterations.

To evaluate the performance of the proposed method, five criteria are considered: mean squared error (MSE), squared bias (SB), proportion choosing the true model (PCTM), true positive rate (TPR), and false positive rate (FPR). The MSE is estimated by

$$\widehat{MSE} = \frac{1}{500}\sum_{r=1}^{500}\left(\hat{\theta}^{(r)} - \theta^*\right)^T\left(\hat{\theta}^{(r)} - \theta^*\right), \tag{15}$$

where $\hat{\theta}^{(r)}$ is the PL estimate in replication r and θ^* is the true parameter value. The SB is estimated by

$$\widehat{SB} = \left(\bar{\hat{\theta}} - \theta^*\right)^T\left(\bar{\hat{\theta}} - \theta^*\right), \tag{16}$$

with $\bar{\hat{\theta}} = \frac{1}{500}\sum_{r=1}^{500}\hat{\theta}^{(r)}$. The PCTM is the proportion of occasions where the heterogeneity pattern is correctly identified, that is, all of the true non-zero and zero increment components are correctly identified. The TPR is the chance of correctly identifying the true non-zero increments of loadings. The FPR is the chance of incorrectly identifying the true zero increments of loadings.

Based on the theoretical and empirical results of Huang *et al.* (2017), we expected the following: (1) the MSE and SB of the PL estimator decrease as the sample size increases; (2) the PCTM, TPR, and FPR improve as the sample size increases; (3) the PCTM based on the BIC tends to 1 when the sample size is large; (4) the previous consistency result does not hold for the AIC. The simulation results are presented in Figure 1. Our expectations were well supported except for the consistency of the BIC. Although the BIC could asymptotically identify all zero and non-zero increment components of the loadings under the null effect condition as well as under the medium and large effect conditions, its PCTM was very low when the effect size was small. A previous study has also found that the empirical performance of the BIC in selecting the true model can be quite bad if the non-zero parameter is small (Vrieze, 2012). To compare the performances of the AIC and BIC in terms of model selection, we found that the BIC is mostly better than the AIC, especially in the null-effect condition. However, under non-null cases the AIC could outperform the BIC if either the effect size or the sample size was small. Even if the PCTM based on the AIC does not approach 1, the AIC still yields high TPR and low FPR. In general, our observations about the behaviour of the AIC and BIC were consistent with the existing simulation results (e.g., Haughton, Oud, & Jansen, 1997; Huang, 2017; Vrieze, 2012).

In summary, if the heterogeneity is null or moderate to large, the BIC can perfectly identify the true patterns under large sample sizes. Otherwise, the AIC could still be used to identify most of the zero and non-zero components. Despite this, the AIC seldom identifies the true model in a null condition. Hence, under small sample sizes, neither BIC nor AIC could generally yield correct selection results.

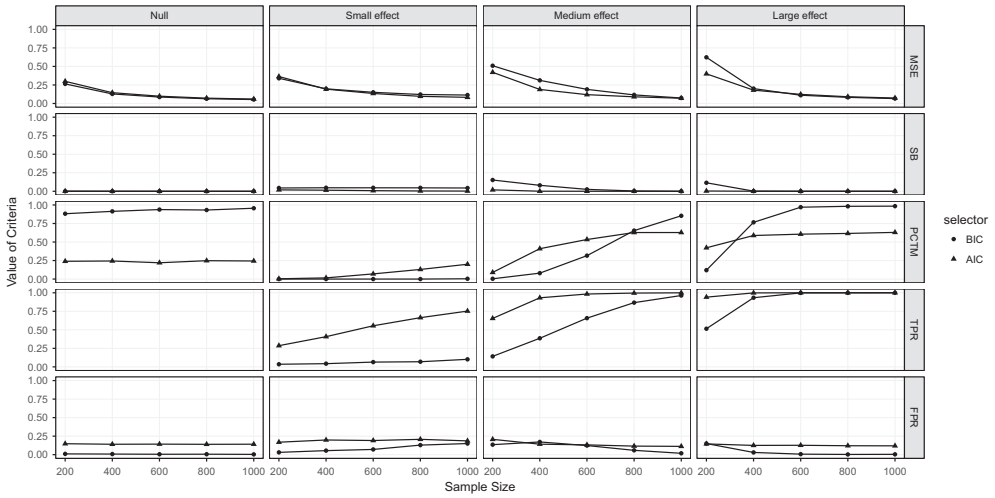

Figure 1. Mean squared error (MSE), squared bias (SB), proportion choosing the true model (PCTM), true positive rate (TPR), and false positive rate (FPR) for multi-group factor analysis via penalized likelihood with minimax concave penalty across four sizes of differences.

6. Real data illustration: Exploring the pattern of partial factorial invariance

In this section the proposed method is applied to explore the pattern of partial factorial invariance for the data of Holzinger and Swineford (1939). The data set includes the responses of 301 junior high school students from the Pasteur High School ($N_1 = 156$) and the Grant-White High School ($N_2 = 145$) on 24 psychological tests. In the following analysis, only the first 19 tests are considered, including visual perception (v_1), cubes (v_2), paper form board (v_3), flags (v_4), general information (v_5), paragraph comprehension (v_6), sentence completion (v_7), word classification (v_8), word meaning (v_9), addition (v_{10}), code (v_{11}), counting groups of dots (v_{12}), straight and curved capitals (v_{13}), word recognition (v_{14}), number recognition (v_{15}), figure recognition (v_{16}), object number (v_{17}), number-figure (v_{18}), and figure-word (v_{19}). These 19 variables are thought to mainly reflect four correlated latent constructs: spatial (f_1; $v_1, \ldots, v_4$), verbal (f_2; $v_5, \ldots, v_9$), speed (f_3; $v_{10}, \ldots, v_{13}$), and memory (f_4; $v_{14}, \ldots, v_{19}$). However, some studies argue that the independent cluster structure (i.e., each variable is only influenced by one factor) may not be appropriate for explaining the data set (e.g., Muthén & Asparouhov, 2012). Hence, when implementing our PL, the whole loading matrix is estimated except for v_1, v_9, v_{12}, v_{14}. They are assumed to be anchor measures with homogeneous fixed loadings – to remove the rotational indeterminacy. Our analysis result also shows that – although the independent cluster structure can fit the data acceptably – some of the indicators are clearly not pure measures.

We first present the results using the traditional approach for examining strong factorial invariance. The traditional method assumes an independent cluster structure for the loading matrix with an uncorrelated error structure. The analysis was conducted using the `measurementInvariance` function of the R package *semTools* (semTools Contributors, 2016). Under no homogeneous constraints on loadings and intercepts did the multi-group factor analysis model fit the data acceptably ($\chi^2 = 475.30$, $df = 292$, RMSEA = .065).

When the loadings across groups were constrained to be equal, the chi-squared difference test indicated no significance ($\chi^2 = 493.24$, $df = 307$, RMSEA = .063, $\Delta\chi^2 = 17.94$, $p = .266$). Hence, weak factorial invariance is satisfied. After further imposing constraints on the intercepts, a significant difference was observed ($\chi^2 = 611.53$, $df = 322$, RMSEA = .077, $\Delta\chi^2 = 118.291$, $p \leq .001$), implying that the strong factorial invariance condition is invalid and some intercepts are not invariant across the two schools.

We now implement the proposed PL to explore the pattern of partial factorial invariance. The Pasteur school is set as the reference group (i.e., $\underline{B}_1^{(vf)} = 0$, $\underline{\alpha}_1^{(v)} = 0$, and $\underline{\alpha}_1^{(f)} = 0$). The latent mean of the Pasteur school is also restricted to zero (i.e., $\underline{\alpha}^{(f)} = 0$). To explore the sparsity pattern of the loading matrix, we estimate all the reference components of loadings that are not in the independent cluster part, with penalization. The increment components corresponding to the loadings and intercepts of non-anchor measurements are all estimated with penalization, including elements in $\underline{B}_2^{(vf)}$ and $\underline{\alpha}_2^{(v)}$. The measurement errors are assumed to be uncorrelated. To dismiss the scaling effect on variables, we standardized each variable by the pooled means and standard deviations. The MCP penalty was utilized and optimal pairs of (λ, δ) were selected from $\Lambda \times \Delta = \{.01, .02, ..., .40\} \times \{2, 3, 4, 5\}$ via both BIC and AIC.

The final PL estimates for loadings and intercepts are presented in Table 4 (BIC solution) and Table 5 (AIC solution). The final model with the BIC fits the data well ($\chi^2 = 372.757$, $df = 300$, RMSEA = .040). We can observe that PL yields a sparse loading

Table 4. The final penalized likelihood (PL) estimate under the BIC selector ($\hat{\lambda} = .13, \hat{\delta} = 3$) for the data taken from Holzinger and Swineford (1939)

	Pasteur school					Grant-white school				
	Intercept	f_1	f_2	f_3	f_4	Intercept	f_1	f_2	f_3	f_4
v_1	−.04	**1**	**0**	**0**	**0**	−.04	**1**	**0**	**0**	**0**
v_2	−.02	.63				−.02	.63			
v_3	−.03	.66				−.03	.66			
v_4	.21[a]	.85				−.30[a]	.85			
v_5	−.29[a]		1.02		−.15	−.20[a]		1.02		−.15
v_6	−.25		.98			−.25		.98		
v_7	−.26	−.10	1.10		−.10	−.26	−.10	1.10		−.10
v_8	−.32[a]	.03	.83	.03		−.09[a]	.03	.83	.03	
v_9	−.25	**0**	**1**	**0**	**0**	−.25	**0**	**1**	**0**	**0**
v_{10}	.19[a]	−.40		1.22		.03[a]	−.40		1.22	
v_{11}	.04		.13	.91	.12	.04		.13	.91	.12
v_{12}	.09	**0**	**0**	**1**	**0**	.09	**0**	**0**	**1**	**0**
v_{13}	.04	.39		.74		.04	.39		.74	
v_{14}	−.07	**0**	**0**	**0**	**1**	−.07	**0**	**0**	**0**	**1**
v_{15}	.02[a]		−.07		.87	−.11[a]		−.07		.87
v_{16}	−.05	.40			.67	−.05	.40			.67
v_{17}	.19[a]	−.29		.44	.87	−.22[a]	−.29		.44	.87
v_{18}	−.02			.34	.64	−.02			.34	.64
v_{19}	−.28[a]			.22	.54	.26[a]			.22	.54

Note. Parameters fixed for identification are bold. Blank table cells indicate a corresponding estimate of zero.

[a]A non-invariant corresponding parameter across the groups.

Table 5. The final penalized likelihood (PL) estimate under the AIC selector ($\hat{\lambda} = .07$, $\hat{\delta} = 3$) for the Data taken from Holzinger and Swineford (1939)

	Pasteur school					Grant-white school				
	Intercept	f_1	f_2	f_3	f_4	Intercept	f_1	f_2	f_3	f_4
v_1	.01	**1**	**0**	**0**	**0**	.01	**1**	**0**	**0**	**0**
v_2	−.09[a]	.69	−.05	−.05	−.03	.13[a]	.69	−.05	−.05	−.03
v_3	−.01	.68			−.38	−.01	.68			.14
v_4	.21[a]	1.19	−.37			−.26[a]	.75	.08		
v_5	−.31[a]	−.08	1.02		−.20	−.07[a]	−.08	1.02	.23[a]	−.20
v_6	−.22		.98			−.22		.98		
v_7	−.25[a]	−.18	1.21[a]	−.08[a]	−.10	−.17[a]	−.18	1.03[a]	.16[a]	−.10
v_8	−.33[a]	.02	.78			.03[a]	.02	.78	.26[a]	
v_9	−.22	**0**	**1**	**0**	**0**	−.22	**0**	**1**	**0**	**0**
v_{10}	.23[a]	−.50	.26	.98	.25	−.31[a]	−.50	.26	.98	.25
v_{11}	.04[a]		.27	.72	.31	−.14[a]		.27	.72	.31
v_{12}	.06	**0**	**0**	**1**	**0**	.06	**0**	**0**	**1**	**0**
v_{13}	.02	.41		.67		.02	.41	.09[a]	.67	
v_{14}	−.11	**0**	**0**	**0**	**1**	−.11	**0**	**0**	**0**	**1**
v_{15}	.07[a]	.13	−.28[a]		.88	−.24[a]	.13	−.08[a]		.88
v_{16}	−.08[a]	.52	−.11	.29[a]	.69	−.01[a]	.52	−.11	−.11[a]	.69
v_{17}	.19[a]	−.23		.55[a]	.85[a]	−.43[a]	−.23		.24[a]	1.14[a]
v_{18}	−.01			.24	.70	−.01	.42[a]	−.21[a]	.24	.70
v_{19}	−.27[a]	.12		.40[a]	.54	.19[a]	.12		.09[a]	.54

Note. Parameters fixed for identification are bold. Blank table cells indicate a corresponding estimate of zero.

[a]A non-invariant corresponding parameter across the groups.

matrix for both groups. Many non-zero 'cross-loadings' show that some of the indicators are not pure measures. The zero increment loading matrix indicates that the weak factorial invariance condition is satisfied. However, the intercepts of v_4, v_5, v_8, v_{10}, v_{15}, v_{17}, and v_{19} are shown to be not invariant across the two groups. The strong factorial invariance condition is invalid. The AIC selector yields an excellent-fitting model (χ^2 = 274.525, *df* =268, RMSEA = .012). Compared to the BIC, the AIC identifies many more non-zero 'cross-loadings' as well as heterogeneous loadings and intercepts, which is consistent with our observation in the simulation. About a quarter of loadings and three-quarters of intercepts are identified as heterogeneous. Although the results using the BIC and AIC are different, the intercepts of v_4, v_5, v_8, v_{10}, v_{15}, v_{17}, and v_{19} are consistently recognized as non-invariant.

In summary, the PL method with the BIC selector yielded the same conclusion as the traditional approach for Holzinger's data, that is, the weak factorial invariance was satisfied but the strong factorial invariance was not. The benefit of using the PL method is that it can further obtain the pattern of partial invariance. Once the non-invariant measurements or items are identified, one may delete or reserve these indicators by evaluating the potential consequence of partial invariance in selection according to the approach of Millsap and Kwok (2004). On the other hand, the result using the AIC showed that most measurements were not invariant. It is difficult to say whether the BIC or AIC

result is more accurate here because the sample sizes for both schools are small. From our personal viewpoint, the conservative BIC solution is preferred due to the exploratory nature of the current examination.

7. Discussion

In this study a penalized likelihood method for multi-group structural equation modelling has been established. The proposed method extends the work of Huang *et al.* (2017) and Jacobucci *et al.* (2016) in the sense that several samples can be simultaneously compared. Through our PL method, the heterogeneity pattern of a moment structure across groups can be identified efficiently. Our numerical experiment and real data example both show that the PL for MGSEM can be used to efficiently explore the pattern of partial factorial invariance as well. This application is in concert with the approaches of Tutz and Schauberger (2015) and Magis, Tuerlinckx, and De Boeck (2015) for differential item functioning, which aims to examine the invariance of intercepts of binary measurements via penalty. On the other hand, the current method only considers continuous measurements. Despite this, it can be further used to explore the invariance of factor loadings.

It is worth contrasting the proposed method for measurement invariance with the alignment method developed by Asparouhov and Muthén (2014). The alignment method is accomplished by simultaneously minimizing a fitting function for multi-group factor analysis and a cost function to evaluate the degree of non-invariance. Hence, the alignment method can be also understood as a PL method with a 'saturated' penalty level and a specific form of penalty term – the sums of component loss functions (Jennrich, 2006) based on pairwise differences among loadings and intercepts. Future studies can incorporate this type of penalty term to develop sparse estimation procedures for MGSEM.

Despite the contributions of the present work, there are still limitations that call for further studies. First, the diagonal elements of Φ_g are not allowed to be penalized and hence the heterogeneity of variance parameters across groups cannot be explored via penalty. This restriction is due to the ECM algorithm, not to PL. One possible way to remove this limitation is to consider Newton-type algorithms for replacing the ECM algorithm (e.g., Friedman, Hastie, & Tibshirani, 2010).

Second, our proposed method can be only applied to deal with continuous data. In psychological research, polytomous data are often encountered because of the widespread use of Likert-type scales. A tentative way to handle polytomous data is to replace the covariance matrix in equation (6) by an estimated polychoric correlation matrix. However, it is worth developing a more direct PL method to incorporate the discrete nature of response variables (e.g., Katsikatsou, Moustaki, Yang-Wallentin, & JÃureskog, 2012; Muthén, 1984).

Third, although our simulation and real data example demonstrated the potential utility of PL for examining measurement invariance, a thorough evaluation is required. Previous studies have shown that many factors may influence the performance of existing approaches for measurement invariance (e.g., French & Finch, 2006; Kaplan, 1989; Meade & Bauer, 2007), such as the number of factors, the ratios of numbers of variables to factors, and the value of factor loading. Future studies can systematically vary these factors to understand the behavior of PL in a broad variety of settings. Also, it would be interesting to compare our PL with specification search regarding partial invariance (e.g., Jung & Yoon, 2016, 2017; Millsap & Kwok, 2004).

Finally, the current work does not propose any formal inference procedure, such as hypothesis testing or confidence interval. It is known that making valid inferences after

model selection (or choosing a penalty level) is a challenging task (Leeb & Pötscher, 2006; Pötscher, 1991). Recently, some encouraging results under lasso regression have been reported (e.g., Lee, Sun, Sun, & Taylor, 2016; Tibshirani, Taylor, Lockhart, & Tibshirani, 2016). Valid post-selection inference procedures in SEM are awaiting further development.

Acknowledgements

The research was supported by Grant MOST 104-2410-H-006-119-MY2 from the Ministry of Science and Technology in Taiwan.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactionson Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508. https://doi.org/10.1080/10705511.2014.919210

Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data*. Berlin: Springer.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Chaudhuri, S., Drton, M., & Richardson, T. S. (2007). Estimation of a covariance matrix with zeros. *Biometrika*, *94*(1), 199–216. https://doi.org/10.1093/biomet/asm007

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of *Q*-matrix based diagnostic classification models. *Journal of the American Statistical Association*, *110*(510), 850–866. https://doi.org/10.1080/01621459.2014.934827

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*, *39*(1), 1–38. https://doi.org/10.2307/2984875

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. https://doi.org/10.1198/016214501753382273

Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space (invited review article). *Statistica Sinica*, *20*(1), 101–148. https://doi.org/10.1007/s11425-015-5062-9

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(3), 378–402. https://doi.org/10.1207/s15328007sem1303_3

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. Boca Raton, FL: CRC Press.

Haughton, D. M., Oud, J. H., & Jansen, R. A. (1997). Information and other criteria in structural equation model selection. *Communications in Statistics – Simulation and Computation*, *26*(4), 1477–1516. https://doi.org/10.1080/03610919708813451

Hirose, K., & Yamamoto, M. (2014). Estimation of an oblique structure via penalized likelihood factor analysis. *Computational Statistics and Data Analysis*, *79*, 120–132. https://doi.org/10.1016/j.csda.2014.05.011

Hirose, K., & Yamamoto, M. (2015). Sparse estimation via nonconcave penalized likelihood in factor analysis model. *Statistics and Computing*, *25*(5), 863–875. https://doi.org/10.1007/s11222-014-9458-0

Holzinger, K. J., & Swineford, F. (1939). *A study in factor analysis: The stability of a bifactor solution* (Supplementary Educational Monograph No. 48). Chicago, IL, USA:University of Chicago.

Huang, P.-H. (2017). Asymptotics of AIC, BIC, and RMSEA for model selection in structural equation modeling. *Psychometrika*, *82*(2), 407–426. https://doi.org/10.1007/s11336-017-9572-y

Huang, P.-H., Chen, H., & Weng, L.-J. (2017). A penalized likelihood method for structural equation modeling. *Psychometrika*, *82*(2), 329–354. https://doi.org/10.1007/s11336-017-9566-9

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 555–566. https://doi.org/10.1080/10705511.2016.1154793

Jennrich, R. I. (2006). Rotation to simple loadings using component loss functions: The oblique case. *Psychometrika*, *71*(1), 173–191. https://doi.org/10.1007/s11336-003-1136-B

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 642–657. https://doi.org/10.1080/10705510903206014

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409–426. https://doi.org/10.1007/BF02291366

Jung, E., & Yoon, M. (2016). Comparisons of three empirical methods for partial factorial invariance: Forward, backward, and factor-ratio tests. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 567–584. https://doi.org/10.1080/10705511.2015.1138092

Jung, E., & Yoon, M. (2017). Two-step approach to partial factorial invariance: Selecting a reference variable and identifying the source of noninvariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(1), 65–79. https://doi.org/10.1080/10705511.2016.1251845

Kaplan, D. (1989). Power of the likelihood ratio test in multiple group confirmatory factor analysis under partial measurement invariance. *Educational and Psychological Measurement*, *49*(3), 579–586. https://doi.org/10.1177/001316448904900308

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., & JÃureskog, K. G. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics and Data Analysis*, *56*(12), 4243–4258. https://doi.org/10.1016/j.csda.2012.04.010

Knight, K., & Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics*, *28*(5), 1356–1378. https://doi.org/10.1017/CBO9781107415324.004

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics*, *44*(3), 907–927. https://doi.org/10.1214/15-AOS1371

Leeb, H., & Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics*, *34*(5), 2554–2591. https://doi.org/10.1214/009053606000000821

Magis, D., Tuerlinckx, F., & De Boeck, P. (2015). Detection of differential item functioning using the lasso approach. *Journal of Educational and Behavioral Statistics*, *40*(2), 111–135. https://doi.org/10.3102/1076998614559747

McArdle, J. J., & McDonald, R. P. (1984). Some algebraic properties of the reticular action model for moment structures. *British Journal of Mathematical and Statistical Psychology*, *37*(2), 234–251. https://doi.org/10.1111/j.2044-8317.1984.tb00802.x

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 611–635. https://doi.org/10.1080/10705510701575461

Meng, X.-L., & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, *80*(2), 267–278. https://doi.org/10.1093/biomet/80.2.267

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/Bf02294825

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

Millsap, R. E., & Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, *9*(1), 93–115. https://doi.org/10.1037/1082-989X.9.1.93

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. https://doi.org/10.1007/BF02294210

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. https://doi.org/10.1037/a0026802

Pötscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory*, 7(2), 163–185. https://doi.org/10.1017/S0266466600004382

R Core Team (2017). *R: A language and environment for statistical computing [Computer software manual]*. Vienna, Austria: R Foundation for Statistical Computing.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, *27*(2), 229–239. https://doi.org/10.1111/j.2044-8317.1974.tb00543.x

Steiger, J. H. (1998). A note on multiple sample extensions of the RMSEA fit index. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(4), 411–419. https://doi.org/10.1080/10705519809540115

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, *58*(1), 267–288. https://doi.org/10.1111/j.1467-9868.2011.00771.x

Tibshirani, R. J., Taylor, J., Lockhart, R., & Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, *111*(514), 600–620. https://doi.org/10.1080/01621459.2015.1108848

semTools Contributors (2016). *semtools: Useful tools for structural equation modeling*. R package version 0.4-14. Retrieved from http://cran.r-project.org/web/packages/semTools/index.html

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika*, *80*(1), 21–43. https://doi.org/10.1007/s11336-013-9377-6

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*(2), 228–243. https://doi.org/10.1037/a0027127

Wellner, J., & Zhang, T. (2012). Introduction to the special issue on sparsity and regularization methods. *Statistical Science*, *27*(4), 447–449. https://doi.org/10.1214/12-STS409

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In J. K. Bryant, M. Windle & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 435–463. https://doi.org/10.1080/10705510701301677

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, *38*(2), 894–942. https://doi.org/10.1214/09-AOS729

Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, *7*, 2541–2563. https://doi.org/10.1109/TIT.2006.883611

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. https://doi.org/10.1198/016214506000000735

Zou, H., Choi, J., & Oehlert, G. (2011). A penalized maximum likelihood approach to sparse factor analysis. *Statistics and its Interface*, *3*, 429–436. https://doi.org/10.4310/SII.2010.v3.n4.a1

Received 2 July 2017; revised version received 12 October 2017

Appendix A: Derivation of E-step

Let $\hat{\alpha}_g$, $\hat{\mathrm{B}}_g$, and $\hat{\Phi}_g$ represent the estimates for the parameter matrices of the group g at iteration step t. The corresponding model-implied mean vector and covariance matrix are denoted by $\hat{\mu}_g^{(\eta)} = (I - \hat{\mathrm{B}}_g)^{-1}\hat{\alpha}_g$ and $\hat{\Sigma}_g^{(\eta\eta)} = (I - \hat{\mathrm{B}}_g)^{-1}\hat{\Phi}_g(I - \hat{\mathrm{B}}_g)^{-1^T}$, respectively. By the working assumption

$$\begin{pmatrix} \boldsymbol{v}_g \\ \eta_g \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mu_g^{(v)} \\ \mu_g^{(\eta)} \end{pmatrix}, \begin{pmatrix} \Sigma_g^{(vv)} & \Sigma_g^{(v\eta)} \\ \Sigma_g^{(\eta v)} & \Sigma_g^{(\eta\eta)} \end{pmatrix} \right],$$

we have

$$\mathbb{E}(\eta_g | \boldsymbol{v}_g) = \mu_g^{(\eta)} + \Sigma_g^{(\eta v)}\Sigma_g^{(vv)^{-1}}(\boldsymbol{v}_g - \mu_g^{(v)})$$

and

$$\mathbb{V}\mathrm{ar}(\eta_g | \boldsymbol{v}_g) = \Sigma_g^{(\eta\eta)} - \Sigma_g^{(\eta v)}\Sigma_g^{(vv)^{-1}}\Sigma_g^{(v\eta)}.$$

Therefore, $e_g^{(\eta)}$ and $C_g^{(\eta\eta)}$ can be derived as

$$e_g^{(\eta)} = J_g + K_g e_g^{(v)}$$

and

$$\begin{aligned} C_g^{(\eta\eta)} =& \hat{\Sigma}_g^{(\eta\eta)} - \hat{\Sigma}_g^{(\eta v)}\hat{\Sigma}_g^{(vv)^{-1}}\hat{\Sigma}_g^{(v\eta)} \\ &+ J_g J_g^T + J_g e_g^{(v)^T} K_g^T + K_g e_g^{(v)} J_g^T + K_g C_g^{(vv)} K_g^T, \end{aligned}$$

where $J_g = \hat{\mu}_g^{(\eta)} - \hat{\Sigma}_g^{(\eta v)}\hat{\Sigma}_g^{(vv)^{-1}}\hat{\mu}_g^{(v)}$ and $K = \hat{\Sigma}_g^{(\eta v)}\hat{\Sigma}_g^{(vv)^{-1}}$. The derivation of $C_g^{(\eta\eta)}$ is based on the formula $\mathbb{V}\mathrm{ar}[\eta_g | \boldsymbol{v}_g] = \mathbb{E}[\eta_g \eta_g^T | \boldsymbol{v}_g] - \mathbb{E}[\eta_g | \boldsymbol{v}_g]\mathbb{E}[\eta_g^T | \boldsymbol{v}_g]$. Note that $e_g^{(v)}$ and $C_g^{(vv)}$ are simply m_g and $S_g + m_g m_g^T$ respectively, because $\boldsymbol{v}_g$ is observable.

When deriving the CM-steps for the covariance parameters, $C_g^{(\zeta\zeta)}$ is also necessary. Its formula is

$$\begin{aligned} C_g^{(\zeta\zeta)} =& C_g^{(\eta\eta)} - e_g^{(\eta)}\hat{\alpha}_g^T - C_g^{(\eta\eta)}\hat{\mathrm{B}}_g^T - \hat{\alpha}_g e_g^{(\eta)^T} + \hat{\alpha}_g\hat{\alpha}_g^T \\ &+ \hat{\alpha}_g e_g^{(\eta)^T}\hat{\mathrm{B}}_g^T - \hat{\mathrm{B}}_g C_g^{(\eta\eta)^T} + \hat{\mathrm{B}}_g e_g^{(\eta)}\hat{\alpha}_g^T + \hat{\mathrm{B}}_g C_g^{(\eta\eta)}\hat{\mathrm{B}}_g^T. \end{aligned}$$

Appendix B: Derivation of CM-step

Regression coefficients and intercepts

For $\left|\underline{\beta}_{jk}\right| > 0$, we have

$$\frac{\partial \mathcal{M}(\theta|\hat{\theta}^{(t)})}{\partial \underline{\beta}_{jk}} = \frac{\partial}{\partial \underline{\beta}_{jk}} \mathbb{E}\left[-\frac{1}{2} \sum_{g=1}^{G} \frac{w_g}{N_g} \sum_{n=1}^{Ng} \sum_{i=1}^{P+M} \sum_{i'=1}^{P+M} \phi_g^{ii'} \zeta_{gni} \zeta_{gni'} \middle| \mathcal{V}, \hat{\theta}^{(t)} \right] + \frac{\partial}{\partial \underline{\beta}_{jk}} c_{\underline{\beta}_{jk}} \rho(\left|\underline{\beta}_{jk}\right|, \lambda).$$

The first term can be simplified to

$$\begin{aligned}
&\mathbb{E}\left[-\frac{1}{2} \sum_{g=1}^{G} \frac{w_g}{N_g} \phi_g^{jj} \sum_{n=1}^{Ng} \frac{\partial}{\partial \underline{\beta}_{jk}} \zeta_{gnj} \zeta_{gnj} \middle| \mathcal{V}, \hat{\theta}^{(t)} \right] - \mathbb{E}\left[\sum_{g=1}^{G} \frac{w_g}{N_g} \sum_{i \neq j} \sum_{n=1}^{Ng} \phi_g^{ji} \frac{\partial}{\partial \underline{\beta}_{jk}} \zeta_{gnj} \zeta_{gni} \middle| \mathcal{V}, \hat{\theta}^{(t)} \right] \\
&= \sum_{g=1}^{G} w_g \phi_g^{jj} \left[C_g^{(\eta\eta)}[j,k] - e_g^{(\eta)}[k] \alpha_{gj} - \sum_{i=1}^{P+M} \left(\underline{\beta}_{ji} + \underline{\beta}_{gji} \right) C_g^{(\eta\eta)}[i,k] \right] \\
&\quad + \sum_{g=1}^{G} w_g \sum_{i \neq j} \phi_g^{ji} \left[C_g^{(\eta\eta)}[i,k] - e_g^{(\eta)}[k] \alpha_{gi} - \sum_{i'=1}^{I} \beta_{gii'} C_g^{(\eta\eta)}[i',k] \right].
\end{aligned}$$

If $\underline{\beta}_{jk}$ is not penalized, $\hat{\underline{\beta}}_{jk}^{(t+1)}$ is the solution of the equation

$$\begin{aligned}
&\sum_{g=1}^{G} w_g \phi_g^{jj} \left[C_g^{(\eta\eta)}[j,k] - e_g^{(\eta)}[k] \hat{\alpha}_{gj} - \hat{\underline{\beta}}_{jk}^{(t+1)} C_g^{(\eta\eta)}[k,k] - \hat{\underline{\mathrm{B}}}[j,-k] C_g^{(\eta\eta)}[-k,k] - \hat{\underline{\mathrm{B}}}_g[j,] C_g^{(\eta\eta)}[,k] \right] \\
&+ \sum_{g=1}^{G} w_g \sum_{i \neq j} \hat{\phi}_g^{ji} \left(C_g^{(\eta\eta)}[i,k] - e_g^{(\eta)}[k] \hat{\alpha}_{gi} - \hat{\mathrm{B}}_g[i,] C_g^{(\eta\eta)}[,k] \right) = 0.
\end{aligned}$$

The solution is

$$\begin{aligned}
\hat{\underline{\beta}}_{jk}^{(t+1)} &= \frac{1}{\sum_{g=1}^{G} w_g \phi_g^{jj} C_g^{(\eta\eta)}[k,k]} \\
&\quad \times \left[\sum_{g=1}^{G} w_g \hat{\phi}_g^{jj} \left(C_g^{(\eta\eta)}[j,k] - e_g^{(\eta)}[k] \hat{\alpha}_{gj} - \hat{\underline{\mathrm{B}}}[j,-k] C_g^{(\eta\eta)}[-k,k] - \hat{\underline{\mathrm{B}}}_g[j,] C_g^{(\eta\eta)}[,k] \right) \right. \\
&\quad \left. + \sum_{g=1}^{G} w_g \sum_{i \neq j} \hat{\phi}_g^{ji} \left(C_g^{(\eta\eta)}[i,k] - e_g^{(\eta)}[k] \hat{\alpha}_{gi} - \hat{\mathrm{B}}_g[i,] C_g^{(\eta\eta)}[,k] \right) \right].
\end{aligned}$$

For the increment component $\left|\underline{\beta}_{gjk}\right| > 0$, we have

$$\frac{\partial \mathcal{M}\left(\theta|\hat{\theta}^{(t)}\right)}{\partial \underline{\beta}_{gjk}} = \frac{\partial}{\partial \underline{\beta}_{gjk}} \mathbb{E}\left[-\frac{1}{2} \sum_{g=1}^{G} \frac{w_g}{N_g} \sum_{n=1}^{Ng} \sum_{i=1}^{P+M} \sum_{i'=1}^{P+M} \phi_g^{ii'} \zeta_{gni} \zeta_{gni'} \middle| \mathcal{V}, \hat{\theta}^{(t)} \right] + \frac{\partial}{\partial \underline{\beta}_{gjk}} c_{\underline{\beta}_{gjk}} \rho\left(\left|\underline{\beta}_{gjk}\right|, \lambda\right).$$

After some calculation, the first term becomes

$$\begin{aligned}
&\mathbb{E}\left[-\frac{1}{2}\frac{w_g}{N_g}\sum_{n=1}^{Ng}\phi_g^{jj}\frac{\partial}{\partial\underline{\beta}_{gjk}}\zeta_{gnj}\zeta_{gnj}\middle|\mathcal{V},\hat{\theta}^{(t)}\right]-\mathbb{E}\left[\frac{w_g}{N_g}\sum_{i\neq j}\sum_{n=1}^{Ng}\phi_g^{ji}\frac{\partial}{\partial\underline{\beta}_{gjk}}\zeta_{gnj}\zeta_{gni}\middle|\mathcal{V},\hat{\theta}^{(t)}\right]\\
&\quad=w_g\phi_g^{jj}\left[C_g^{(\eta\eta)}[j,k]-e_g^{(\eta)}[k]\alpha_{gj}-\sum_{t=1}^{P+M}\left(\underline{\beta}_{ji}+\underline{\beta}_{gji}\right)C_g^{(\eta\eta)}[i,k]\right]\\
&\qquad+w_g\sum_{i\neq j}\phi_g^{ji}\left[C_g^{(\eta\eta)}[i,k]-e_g^{(\eta)}[k]\alpha_{gi}-\sum_{i'=1}^{P+M}\beta_{gii'}C_g^{(\eta\eta)}[i',k]\right].
\end{aligned}$$

When $\underline{\beta}_{gjk}$ is not penalized, $\hat{\underline{\beta}}_{gjk}^{(t+1)}$ must satisfy

$$\begin{aligned}
&w_g\hat{\phi}_g^{jj}\left[C_g^{(\eta\eta)}[j,k]-e_g^{(\eta)}[k]\hat{\alpha}_{gj}-\hat{\underline{\mathrm{B}}}_g[j,]C_g^{(\eta\eta)}[,k]-\hat{\beta}_{gjk}^{(t+1)}C_g^{(\eta\eta)}[k,k]-\hat{\underline{\mathrm{B}}}_g[j,-k]C_g^{(\eta\eta)}[-k,k]\right]\\
&+w_g\sum_{i\neq j}\hat{\phi}_g^{ji}\left[C_g^{(\eta\eta)}[i,k]-e_g^{(\eta)}[k]\hat{\alpha}_{gi}-\hat{\mathrm{B}}_g[i,]C_g^{(\eta\eta)}[,k]\right]=0.
\end{aligned}$$

The solution is

$$\begin{aligned}
\hat{\underline{\beta}}_{gjk}^{(t+1)}=&\frac{1}{w_g\hat{\phi}_g^{jj}C_g^{(\eta\eta)}[k,k]}\\
&\times\left[w_g\hat{\phi}_g^{jj}\left(C_g^{(\eta\eta)}[j,k]-e_g^{(\eta)}[k]\hat{\alpha}_{gj}-\hat{\underline{\mathrm{B}}}[j,]C_g^{(\eta\eta)}[,k]-\hat{\underline{\mathrm{B}}}_g[j,-k]C_g^{(\eta\eta)}[-k,k]\right)\right.\\
&\left.+w_g\sum_{i\neq j}\hat{\phi}_g^{ji}\left(C_g^{(\eta\eta)}[i,k]-e_g^{(\eta)}[k]\hat{\alpha}_{gi}-\hat{\mathrm{B}}_g[i,]C_g^{(\eta\eta)}[,k]\right)\right].
\end{aligned}$$

To estimate $\underline{\alpha}_j$ and $\underline{\alpha}_{gj}$, the first terms of $\dfrac{\partial\mathcal{M}(\theta|\hat{\theta}^{(t)})}{\partial\underline{\alpha}_j}$ and $\dfrac{\partial\mathcal{M}(\theta|\hat{\theta}^{(t)})}{\partial\underline{\alpha}_{gj}}$ can be simplified to

$$\sum_{g=1}^{G}w_g\phi_g^{jj}\left[e_g^{(\eta)}[j]-\underline{\alpha}_j-\underline{\alpha}_{gj}-\mathrm{B}_g[j,]e_g^{(\eta)}\right]+\sum_{g=1}^{G}w_g\sum_{i\neq j}\phi_g^{ji}\left[e_g^{(\eta)}[i]-\underline{\alpha}_i-\underline{\alpha}_{gi}-\mathrm{B}_g[i,]e_g^{(\eta)}\right]$$

and

$$w_g\phi_g^{jj}\left[e_g^{(\eta)}[j]-\underline{\alpha}_j-\underline{\alpha}_{gj}-\mathrm{B}_g[j,]e_g^{(\eta)}\right]+w_g\sum_{i\neq j}\phi_g^{ji}\left[e_g^{(\eta)}[i]-\underline{\alpha}_i-\underline{\alpha}_{gi}-\mathrm{B}_g[i,]e_g^{(\eta)}\right],$$

respectively. The formulae for $\hat{\underline{\alpha}}_j^{(t+1)}$ and $\hat{\underline{\alpha}}_{gj}^{(t+1)}$ are

$$\begin{aligned}
\hat{\underline{\alpha}}_j^{(t+1)}=&\frac{1}{\sum_{g=1}^{G}w_g\hat{\phi}_g^{jj}}\\
&\times\left[\sum_{g=1}^{G}w_g\hat{\phi}_g^{jj}(e_g^{(\eta)}[j]-\hat{\underline{\alpha}}_{gj}-\hat{\mathrm{B}}_g[j,]e_g^{(\eta)})+\sum_{g=1}^{G}w_g\sum_{i\neq j}\hat{\phi}_g^{ji}\left(e_g^{(\eta)}[i]-\hat{\alpha}_{gi}-\hat{\mathrm{B}}_g[i,]e_g^{(\eta)}\right)\right]
\end{aligned}$$

and

$$\hat{\underline{\alpha}}_{gj}^{(t+1)} = \frac{1}{w_g \hat{\phi}_g^{jj}} \times \left[w_g \hat{\phi}_g^{jj} (e_g^{(\eta)}[j] - \hat{\underline{\alpha}}_j - \hat{\mathrm{B}}_g[j,] e_g^{(\eta)}) + w_g \sum_{i \neq j} \hat{\phi}_g^{ji} \left(e_g^{(\eta)}[i] - \hat{\alpha}_{gi} - \hat{\mathrm{B}}_g[i,] e_g^{(\eta)} \right) \right],$$

respectively.

Covariance and variance parameters

According to our previously defined notation, $\zeta_g[-j] = (\zeta_{g1}, \ldots, \zeta_{g,j-1}, \zeta_{g,j+1}, \ldots, \zeta_{g(P+M)})$ is a subvector of ζ_g after dropping ζ_{gj}. By the factorization rule of the density function, we have $\mathbb{P}(\zeta) = \mathbb{P}(\zeta_{gj} | \zeta_g[-j]) \mathbb{P}(\zeta_g[-j])$. Hence, to estimate ϕ_{gjk}, only $\mathbb{P}(\zeta_{gj} | \zeta_g[-j])$ is needed. Given $j \neq k$, the normal working assumption for ζ_g implies that

$$\zeta_{gj} | \zeta_g[-j] \sim \mathcal{N}\left(\Phi_g[j,-j] \zeta_g[-j], \varphi_{gjj} \right),$$

and hence

$$\frac{\partial \mathcal{M}(\theta | \hat{\theta}^{(t)})}{\partial \underline{\phi}_{jk}} = \frac{\partial}{\partial \underline{\phi}_{jk}} \mathbb{E}\left[-\frac{1}{2} \sum_{g=1}^{G} \frac{w_g}{N_g} \sum_{n=1}^{Ng} \frac{1}{\varphi_{g,jj}} \left(\zeta_{gnj} - \sum_{i \neq j} (\underline{\phi}_{ji} + \underline{\phi}_{gji}) \tilde{\zeta}_{gnl_i}^{(j)} \right)^2 | \mathcal{V}, \hat{\theta}^{(t)} \right] + \frac{\partial}{\partial \underline{\phi}_{jk}} c_{\underline{\phi}_{jk}} \rho(|\underline{\phi}_{jk}|, \lambda),$$

where $\tilde{\zeta}_{gnl_i}^{(j)}$ is the l_i th element of $\Phi_g[-j,-j]^{-1} \zeta_{gn}[-j]$, l_i is the column index of ϕ_{gji} in $\Phi_g[j,-j]$, and $\varphi_{gjj} = \phi_{gjj} - \Phi_g[j,-j] \Phi_g[-j,-j]^{-1} \Phi_g[-j,j]$. The first term can be rewritten as

$$\begin{aligned} &\mathbb{E}\left[-\frac{1}{2} \sum_{g=1}^{G} \frac{w_g}{N_g} \frac{1}{\varphi_{g,jj}} \sum_{n=1}^{Ng} \frac{\partial}{\partial \underline{\phi}_{jk}} \left(\zeta_{gnj} - \sum_{i \neq j} (\underline{\phi}_{ji} + \underline{\phi}_{gji}) \tilde{\zeta}_{gnl_i}^{(j)} \right)^2 | \mathcal{V}, \hat{\theta}^{(t)} \right] \\ &= \sum_{g=1}^{G} \frac{w_g}{\varphi_{g,jj}} \mathbb{E}\left[\frac{1}{N_g} \sum_{n=1}^{Ng} \tilde{\zeta}_{gnl_k}^{(j)} \left(\zeta_{gnj} - \sum_{i \neq j} (\underline{\phi}_{ji} + \underline{\phi}_{gji}) \tilde{\zeta}_{gnl_i}^{(j)} \right) | \mathcal{V}, \hat{\theta}^{(t)} \right] \\ &= \sum_{g=1}^{G} \frac{w_g}{\varphi_{g,jj}} \left[C_g^{(\tilde{\zeta}^{(j)} \zeta)}[l_k, j] - \sum_{i \neq j} (\underline{\phi}_{ji} + \underline{\phi}_{gji}) C_g^{(\tilde{\zeta}^{(j)} \tilde{\zeta}^{(j)})}[l_k, l_i] \right], \end{aligned}$$

where

$$C_g^{(\tilde{\zeta}^{(j)} \zeta)} = \Phi_g[-j,-j]^{-1} C_g^{(\zeta\zeta)}[-j,]$$

and

$$C_g^{(\tilde{\zeta}^{(j)} \tilde{\zeta}^{(j)})} = \Phi_g[-j,-j]^{-1} C_g^{(\zeta\zeta)}[-j,-j] \Phi_g[-j,-j]^{-1}.$$

With no penalty for $\underline{\phi}_{jk}$, $\hat{\underline{\phi}}_{jk}^{(t+1)}$ should satisfy

$$\sum_{g=1}^{G} \frac{w_g}{\varphi_{gjj}} \Bigg[C_g^{(\tilde{\zeta}^{(j)}\zeta)}[l_k, j] - \hat{\underline{\phi}}_{jk}^{(t+1)} C_g^{(\tilde{\zeta}^{(j)}\tilde{\zeta}^{(j)})}[l_k, l_k] - \hat{\underline{\Phi}}[j, -(j,k)] C_g^{(\tilde{\zeta}^{(j)}\tilde{\zeta}^{(j)})}[-l_k, l_k] - \hat{\underline{\Phi}}_g[j, -j] C_g^{(\tilde{\zeta}^{(j)}\tilde{\zeta}^{(j)})}[, l_k] \Bigg] = 0.$$

The solution is

$$\hat{\underline{\phi}}_{jk}^{(t+1)} = \frac{1}{\sum_{g=1}^{G} \frac{w_g}{\varphi_{gjj}} C_g^{(\tilde{\zeta}^{(j)}\tilde{\zeta}^{(j)})}[l_k, l_k]} \times \left[\sum_{g=1}^{G} \frac{w_g}{\varphi_{gjj}} \left(C_g^{(\tilde{\zeta}^{(j)}\zeta)}[l_k, j] - \hat{\underline{\Phi}}[j, -(j,k)] C_g^{(\tilde{\zeta}^{(j)}\tilde{\zeta}^{(j)})}[-l_k, l_k] - \hat{\underline{\Phi}}_g[j, -j] C_g^{(\tilde{\zeta}^{(j)}\tilde{\zeta}^{(j)})}[, l_k] \right) \right].$$

When processing the CM-step for $\underline{\phi}_{gjk}$ ($j \neq k$), the first term of $\frac{\partial \mathcal{M}(\theta | \hat{\theta}^{(t)})}{\partial \underline{\phi}_{gjk}}$ can be simplified to

$$\begin{aligned} &\mathbb{E}\left[-\frac{1}{2} \frac{w_g}{N_g} \frac{1}{\varphi_{gjj}} \sum_{n=1}^{Ng} \frac{\partial}{\partial \underline{\phi}_{gjk}} \left(\zeta_{gnj} - \sum_{i \neq j} (\underline{\phi}_{ji} + \underline{\phi}_{gji}) \tilde{\zeta}_{gnl_i}^{(j)} \right)^2 \Big| \mathcal{V}, \hat{\theta}^{(t)} \right] \\ &= \frac{w_g}{\varphi_{gjj}} \mathbb{E}\left[\frac{1}{N_g} \sum_{n=1}^{Ng} \tilde{\zeta}_{gnl_k}^{(j)} \left(\zeta_{gnj} - \sum_{i \neq j} (\underline{\phi}_{ji} + \underline{\phi}_{gji}) \tilde{\zeta}_{gnl_i}^{(j)} \right) \Big| \mathcal{V}, \hat{\theta}^{(t)} \right] \\ &= \frac{w_g}{\varphi_{gjj}} \left[C_g^{(\tilde{\zeta}^{(j)}\zeta)}[l_k, j] - \sum_{i \neq j} (\underline{\phi}_{ji} + \underline{\phi}_{gji}) C_g^{(\tilde{\zeta}^{(j)}\tilde{\zeta}^{(j)})}[l_k, l_i] \right]. \end{aligned}$$

When $\underline{\phi}_{gjk}$ is not penalized, $\hat{\underline{\phi}}_{gjk}^{(t+1)}$ is the solution of the equation

$$\frac{w_g}{\hat{\varphi}_{gjj}} \Bigg[C_g^{(\tilde{\zeta}^{(j)}\zeta)}[l_k, j] - \hat{\underline{\Phi}}[j, -j] C_g^{\tilde{\zeta}^{(j)}\tilde{\zeta}^{(j)}}[, l_k] - \hat{\underline{\phi}}_{gjk}^{(t+1)} C_g^{(\tilde{\zeta}^{(j)}\tilde{\zeta}^{(j)})}[l_k, l_k] - \hat{\underline{\Phi}}_g[j, -(j,k)] C_g^{(\tilde{\zeta}^{(j)}\tilde{\zeta}^{(j)})}[-l_k, l_k] \Bigg] = 0.$$

The solution is

$$\hat{\underline{\phi}}_{gjk}^{(t+1)} = \frac{1}{C_g^{(\tilde{\zeta}^{(j)}\tilde{\zeta}^{(j)})}[l_k, l_k]} \times \left[C_g^{(\tilde{\zeta}^{(j)}\zeta)}[l_k, j] - \hat{\underline{\Phi}}[j, -j] C_g^{(\tilde{\zeta}^{(j)}\tilde{\zeta}^{(j)})}[, l_k] - \hat{\underline{\Phi}}_g[j, -(j,k)] C_g^{(\tilde{\zeta}^{(j)}\tilde{\zeta}^{(j)})}[-l_k, l_k] \right].$$

Because the conditional variance is not allowed to be penalized in the current framework, we can only derive $\hat{\underline{\varphi}}_{gjj}^{(t+1)}$ under the constraint $\underline{\varphi}_{jj} = 0$. The first derivative of $\mathcal{M}(\theta | \hat{\theta}^{(t)})$ with respect to $\underline{\varphi}_{gjj}$ equals

$$\begin{aligned}\frac{\partial \mathcal{M}\left(\theta|\hat{\theta}^{(t)}\right)}{\partial \underline{\varphi}_{gjj}} &= \frac{\partial}{\partial \underline{\varphi}_{gjj}} \mathbb{E}\left[-\frac{w_g}{2}\log\underline{\varphi}_{gjj} - \frac{1}{2}\frac{w_g}{N_g}\sum_{n=1}^{Ng}\frac{1}{\underline{\varphi}_{gjj}}\left(\zeta_{gnj} - \sum_{i\neq j}\phi_{gji}\tilde{\zeta}^{(j)}_{gnl_i}\right)^2 \middle| \mathcal{V}, \hat{\theta}^{(t)}\right] \\ &= -\frac{w_g}{2\underline{\varphi}_{gjj}} + \frac{w_g}{2\underline{\varphi}^2_{gjj}}\mathbb{E}\left[\frac{1}{N_g}\sum_{n=1}^{Ng}\left(\zeta_{gnj} - \Phi_g[j,-j]\tilde{\zeta}^{(j)}_{gn}\right)^2 \middle| \mathcal{V}, \hat{\theta}^{(t)}\right] \\ &= -\frac{w_g}{2\underline{\varphi}_{gjj}} + \frac{w_g}{2\underline{\varphi}^2_{gjj}}\left[C_g^{(\zeta\zeta)}[j,j] - 2\Phi_g[j,-j]C_g^{(\tilde{\zeta}^{(j)}\zeta)}[,j]\right. \\ &\quad \left. + \Phi_g[j,-j]C_g^{(\tilde{\zeta}^{(j)}\tilde{\zeta}^{(j)})}\Phi_g[-j,j]\right],\end{aligned}$$

which implies that

$$\hat{\underline{\varphi}}^{(t+1)}_{gjj} = C_g^{(\zeta\zeta)}[j,j] - 2\hat{\Phi}_g[j,-j]C_g^{(\tilde{\zeta}^{(j)}\zeta)}[,j] + \hat{\Phi}_g[j,-j]C_g^{(\tilde{\zeta}^{(j)}\tilde{\zeta}^{(j)})}\hat{\Phi}_g[-j,j].$$