Nonlinear structural equation modeling with distribution-free method

Yan D. Zhao\textsuperscript{a,}* and Dewi Rahardja\textsuperscript{b}
\textsuperscript{a}Department of Clinical Sciences and Simmons Cancer Center, UT Southwestern Medical Center, Dallas, TX, USA
\textsuperscript{b}Department of Clinical Sciences, UT Southwestern Medical Center, Dallas, TX, USA

Abstract. Structural equation models (SEM) are widely used in many fields including economics and social science. Typical nonlinear SEMs consist of two parts: a linear measurement model relating observed measurements to underlying latent variables, and a nonlinear structural model describing relationships among the latent variables. For such models, we propose a pseudo likelihood approach based on a hypothetical normal mixture assumption on the latent variables. To obtain pseudo likelihood parameter estimates, a Monte Carlo EM algorithm is developed. Standard errors for the structural parameter estimates are obtained by combining an empirical observed information matrix and a bootstrap estimated covariance matrix. For nonlinear SEMs with latent variables with various distributions, we conduct simulations to show our approach produced unbiased parameter estimates and confidence intervals with nominal coverage.

Keywords: Bootstrap, deconvolution, latent variable modeling, MCEM algorithm, normal mixture, standard error estimation

1. Introduction

Structural equation analysis is useful in the areas where some latent variables of interest are not directly observable, but related indicators are available. The structural equation system consists of two sub-models; a measurement model relating observed measurements to latent variables, and a structural model representing relationships among the latent variables. A system with linear measurement model and linear structural model has been studied extensively in the literature. For general introduction, see, e.g., Bollen [5], Bentler [4], and Joreskog and Sorbom [9]. For continuous type observations taking more than a few discrete values, the use of a model that is nonlinear in latent variables has started to receive attention in applications.

The literature on nonlinear structural equation analysis started with Kenny and Judd [10] who proposed a fitting method for structural model with one quadratic or cross-product term. For the same model, Bollen [6] considered instrumental variable estimation, and Arminger and Muthen [3] suggested a Bayesian approach. Wall and Amemiya [14] introduced a distribution-free method which generalized and modified the approach by Kenny and Judd [10]. Extending these methods to more general situations is not straightforward. For a general polynomial structural model, Wall and Amemiya [13] developed an approach that produces consistent parameter estimates and accurate inference procedures without specifying the distributional form of the latent variables. But, their method does not directly apply to non-polynomial structural models. Therefore, a simultaneous general nonlinear (non-polynomial) structural model has not been treated in the literature.

In this paper, a general system of simultaneous nonlinear structural models is considered. The models do not have to be polynomial in the latent variables. As in the literature, a standard linear measurement model is assumed. Section 2 introduces a system with a general nonlinear system of structural models. Section 3 presents our pseudo maximum likelihood approach and the MCEM algorithm. We propose standard error estimation in Section 4. Simulation studies are reported in Section 5. The paper is concluded with a discussion in Section 6.

*Corresponding author: Yan D. Zhao, 5323 Harry Hines Blvd, Room ND3.101E, Dallas, Texas 75390-8590, USA. E-mail: Daniel.Zhao@utsouthwestern.edu.
2. Nonlinear structural equation system

Let $i$ be the index for the individuals in a sample of size $n$, and let $z_i$ be a $k \times 1$ underlying latent vector for the $i^{th}$ individual. We consider a general implicit simultaneous structural system in the following form,

$$G_0(z_i) = e_i,$$  \hspace{1cm} (1)

where $e_i$ is a $q \times 1$ equation error with zero mean, and $q$ is the number of non-redundant relationships among the components of $z_i$. In general, the covariance matrix of $e_i$ is unrestricted, although some elements are assumed zero if desired. Because model (1) can be multiplied by a full rank matrix without altering the essential meaning of the model, this model is not identifiable. Recall that a linear simultaneous system is typically made identifiable by setting some coefficients to be zeros and ones, and assuming the existence of a nonsingular coefficient matrix, i.e., assuming the existence of a uniquely specified reduced form. We use the same idea for identifying the general nonlinear model (1), and consider only those models that can be expressed in an explicit reduced form. That is, we consider model (1) which can be solved explicitly for $q$ elements of $z_i$ in terms of the other $k - q$ elements. To express such a reduced form representation, we write

$$z_i = \begin{pmatrix} y_i \\ x_i \end{pmatrix}, \hspace{1cm} w_i = \begin{pmatrix} x_i \\ e_i \end{pmatrix},$$  \hspace{1cm} (2)

where $y_i$ is $q \times 1$, $x_i$ is $(k - q) \times 1$, and $w_i$ is $k \times 1$. We consider model (1) can be expressed in the identifiable reduced form

$$y_i = G(w_i; \beta),$$  \hspace{1cm} (3)

where $\beta$ is a vector of relationship parameters. Note that, if a model based on subject-matter meaning has the form Eq. (1) with a zero mean error $e_i$, then the reduced form Eq. (3) will generally be nonlinear in $e_i$.

Two examples of general nonlinear structural models are as follows:

1. A simultaneous quadratic model (1) might take a form

$$z_{1i} = \beta_1 + \beta_2 z_{2i} + \beta_3 z_{3i} + \beta_4 z_{3i}^2 + e_{1i}, \hspace{1cm} z_{2i} = \beta_5 + \beta_6 z_{1i} + \beta_7 z_{4i} + \beta_8 z_{4i}^2 + e_{2i},$$

which can be written in the reduced form Eq. (3) by the usual linear transformation.

2. A two factor log-linear model can be

$$\log z_{1i} = \beta_0 + \beta_1 z_{2i} + e_i,$$

which can be expressed in the reduced form Eq. (3) as

$$z_{1i} = \exp(\beta_0 + \beta_1 z_{2i} + e_i).$$

The unobservable latent variables $z_i$ are related to observed indicators through a measurement model. We assume that the linear measurement model with the separate indicator form holds. Under this model, two separate sets of observed variables $Y_i$ and $X_i$ are assumed to be related to $y_i$ and $x_i$, respectively, and satisfy linear measurement models in the errors-in-variables parameterization

$$Y_i = \begin{pmatrix} \gamma_y \\ 0 \end{pmatrix} + \begin{pmatrix} \Lambda_y \\ I_q \end{pmatrix} y_i + u_{yi},$$  \hspace{1cm} (4)

$$X_i = \begin{pmatrix} \gamma_x \\ 0 \end{pmatrix} + \begin{pmatrix} \Lambda_x \\ I_{k-q} \end{pmatrix} x_i + u_{xi},$$  \hspace{1cm} (5)

where $(\gamma_y, \Lambda_y)$ and $(\gamma_x, \Lambda_x)$ are unrelated, $\gamma_y, \gamma_x, \Lambda_y,$ and $\Lambda_x$ may contain additional restricted elements, and $u_{yi}$ and $u_{xi}$ are independent zero-mean measurement error vectors. The model with Eqs (4) and (5) is an identifiable factor analysis model treating $y_i$ and $x_i$ as unrestricted factors. Although the structural model (3) may place restrictions on $y_i$, this identification approach using unrestricted $y_i$ is practical, especially for exploring various structural models for a given measurement model.
We assume that $x_i$, $e_i$, $u_{yi}$, and $u_{xi}$ are mutually independent, and that the error terms $e_i$, $u_{yi}$, and $u_{xi}$ are normally distributed, i.e.,

$$
\begin{align*}
    e_i & \sim N(0, \Sigma_e), \\
    u_{yi} & \sim N(0, \Psi_y), \\
    u_{xi} & \sim N(0, \Psi_x).
\end{align*}
$$

Clearly, the distribution of latent vector $y_i$ will be completely specified by the distributions of $x_i$ and error terms $e_i$, $u_{yi}$, and $u_{xi}$. The normality of the measurement errors and the equation error can be considered reasonable in most applications. The distribution of the latent factor $x_i$ characterizing the target population may not be normal, and its shape may be of interest in some applications. In the next section, we propose the use of a flexible distribution for $x_i$, and describe the corresponding estimation procedure.

3. An estimation procedure

In the nonlinear structural equation system comprised of model (3), (4), and (5), under the distributional assumption Eq. (6), the only unspecified distributional assumption is on the latent factor $x_i$. We aim to develop estimators of $\beta$ and $\Sigma$ that work well for various unspecified distributions of $x_i$. We propose an approach using a normal mixture as hypothetically assumed distribution of $x_i$. We consider normal mixtures for the following reasons. First, a normal mixture distribution with many components can approximate a large class of distributions reasonably well. Also, the use of a normal mixture makes our overall estimation procedure feasible and practical. In particular, we use a mixture of $J \times (k - q)$ dimensional normal distributions with different means and common covariance matrix. We denote the density for the normal mixture distribution as $f(x; \theta_x)$, and we let $\theta_{mn}$ consist of all the measurement model parameters. Then the parameter for the whole model is $\theta = (\theta'_1, \theta'_2)'$, where $\theta_2 = (\theta'_m, \theta'_z)'$. In this notation, $\theta_1$ is the structural parameter of interest, and $\theta_2$ contains all of the nuisance parameters.

Our approach for developing an estimate for $\theta_1$ utilizes the pseudo maximum likelihood estimation procedure proposed by Gong and Samaniego [8] and Parke [12]. In this approach, instead of maximizing the likelihood with respect to $\theta_1$ and $\theta_2$, some consistent estimator of the nuisance parameter $\theta_2$ is substituted into the likelihood, and the resulting function is maximized with respect to $\theta_1$.

We start with describing our estimator $\hat{\theta}_2$ of $\theta_2$. First, we apply the maximum likelihood estimation to the two factor analytic measurement models (4) and (5) separately, treating $y_i$ and $x_i$ as unrestricted normal vectors, and obtain $\theta_{mn}$. It has been shown that the maximum normal likelihood estimators of the factor loadings and error variances in the linear factor analysis are consistent and have attractive properties for nearly any unspecified distribution of the factor vector. See, e.g., Amemiya, Fuller, and Pantula [1], Anderson and Amemiya [2], and Browne and Shapiro [7]. These estimators are consistent even when the factor vector has non-normal distributions. To obtain an estimate of $\theta_x$ in the latent variable normal mixture distribution, we use a method referred to as a measurement error deconvolution. This method starts with obtaining the so-called factor score estimator $\hat{x}_i$ of each $x_i$ based on the x-part measurement model (5) and its estimated parameters. If we ignore the errors of $O_p(n^{-1/2})$ in estimation of parameters in Eq. (5), then

$$
\hat{x}_i = x_i + e_i,
$$

where $e_i \sim N(0, \Sigma_e)$, and a consistent estimator $\hat{\Sigma}_x$ of $\Sigma_x$ is available in Wall and Amemiya [13]. If $x_i$ has a normal mixture distribution with a common covariance matrix, and if $r_i$ is normally distributed, then the observed factor score $\hat{x}_i$ also has a normal mixture distribution with a common covariance matrix. Thus, by fitting a normal mixture distribution to $\hat{x}_i$ (using an EM algorithm), and by subtracting $\hat{\Sigma}_x$ from the estimated common covariance matrix with certain adjustment for the parameter space, we obtained $\hat{\theta}_x$. This is our approach to the measurement error deconvolution problem. The resulting $\hat{\theta}_x$, combined with $\hat{\theta}_{mn}$ forms an easily obtainable $\hat{\theta}_2$.

The pseudo maximum likelihood estimator (PMLE) $\hat{\theta}_1$ of $\theta_1$ is obtained by maximizing the likelihood evaluated at $\hat{\theta}_2$ with respect to $\theta_1$. Since the likelihood function does not have an explicit expression, we consider performing the maximization using a form of a Monte Carlo EM (MCEM) algorithm. Treating the latent variable $x_i$ as missing,
with \( Z_i = (Y_i', X_i')' \) and by suppressing terms not depending on \( \theta_1 \), the logarithm of the complete pseudo likelihood is
\[
l_p(\theta_1) = \sum_{i=1}^{n} \left[ \log f(Y_i|w_i; \beta, \hat{\theta}_m) + \log f(e_i; \Sigma_e) \right].
\]

Since the measurement model (4) for \( Y_i \) is the standard linear factor analysis model, we can linearly transform \( Y_i \), using \( \theta_m \), into a residual vector not involving \( w_i \) and \( \beta \), and the factor score estimator, \( \hat{y}_i \). Deleting terms not involving \( \beta \) and \( \Sigma_e \), we write
\[
l_p(\theta_1) = \sum_{i=1}^{n} \left[ \log f(\hat{y}_i|w_i; \beta, \hat{\theta}_m) + \log f(e_i; \Sigma_e) \right],
\]
where the first density is the normal density with mean \( G(w_i; \beta) \) and covariance matrix \( \hat{\Psi}_y \). Here, \( \hat{\Psi}_y \) is an estimate of the variance matrix of \( \hat{y}_i - y_i \), and can be computed in a way similar to \( \Sigma_y \).

Let \( \theta^{(j)} \) denote the value of \( \theta_1 \) at the \( j^{th} \) iteration for the pseudo likelihood MCEM algorithm. The E-step in the \((j + 1)^{st}\) iteration is to obtain the conditional expectation of \( l_p(\theta_1) \) given \( Z_1, \ldots, Z_n \), under the distribution specified by \( \theta^{(j)}_1 \) and \( \theta_1 \). By the argument used to obtain \( \hat{y}_i \) and \( \hat{x}_i \), this conditional expectation can be taken given \( \hat{y}_i \) and \( \hat{x}_i \), instead of \( Z_1, \ldots, Z_n \).

\[
E^{(j)}(\theta_1) = E \{ l_p(\theta_1)|Z_1, \ldots, Z_n \} = E \{ l_p(\theta_1)|\hat{y}_1, \ldots, \hat{y}_n, \hat{x}_1, \ldots, \hat{x}_n \} = \sum_{i=1}^{n} \int \left[ \log f(\hat{y}_i|w_i; \beta, \hat{\theta}_m) + \log f(e_i; \Sigma_e) \right] f(w_i|\hat{y}_i, \hat{x}_i; \theta^{(j)}_1, \hat{\theta}_2) \, dw_i.
\]

Our approach expresses the conditional density \( f(w_i|\hat{y}_i, \hat{x}_i; \theta^{(j)}_1, \hat{\theta}_2) \) in the following way
\[
f(w_i|\hat{y}_i, \hat{x}_i; \theta^{(j)}_1, \hat{\theta}_2) = \frac{g(w_i)}{\int \int g(w_i) \, dx_i \, de_i},
\]
where
\[
g(w_i) = f(\hat{y}_i|w_i; \beta^{(j)}, \hat{\theta}_m) f(\hat{x}_i|x_i; \hat{\theta}_m) f(x_i; \hat{\theta}_2) f(e_i; \Sigma_e^{(j)})
\]

When Eq. (8) is substituted into Eq. (7), the \( i^{th} \) term in the summation is a ratio of two integrals. We propose to compute these integrals by Monte Carlo simulation, generating pseudo random vectors from the normal distribution \( f(e_i; \Sigma_e^{(j)}) \) and the normal mixture distribution \( f(x_i; \hat{\theta}_2) \). We choose to use this approach, because these random vectors can be generated readily, and because the normal densities \( f(\hat{y}_i|w_i; \beta^{(j)}, \hat{\theta}_m) \) and \( f(\hat{x}_i|x_i; \hat{\theta}_m) \) are simple to evaluate. For each individual \( i \), we generate a pseudo random sample \( \{ \hat{w}_{im} : m = 1, 2, \ldots, M^{(j)} \} \) of size \( M^{(j)} \), where \( \hat{w}_{im} = (\hat{x}_{im}', e_{im}')' \). To decrease the simulation error in the eventual estimate, we generate \( \{ \hat{w}_{im} : m = 1, 2, \ldots, M^{(j)} \} \) separately and independently for different \( i \) and for different iterations. But, we use the same set of size \( M^{(j)} \) for the two integral computations involved for each \( i \), so that the denominator integral plays the role of normalizing constant for weights adding up to one.

At the \( j^{th} \) E-step, the expected conditional likelihood is decomposed into two terms with the first term only includes \( \beta \) and the second term only includes \( \Sigma_e \). Both terms are approximated using Monte Carlo methods. Then, at the \((j + 1)^{st}\) M-step, \( \beta^{(j+1)} \) is obtained by the multivariate nonlinear weighted least squares and \( \Sigma_e^{(j+1)} \) is obtained using a closed-form formula. More details on the E-step and M-step are provided in the Appendix.

To obtain initial estimates for \( \beta \) and \( \Sigma_e \) in our MCEM pseudo maximum likelihood algorithm, we use \( x_i \) and \( \hat{y}_i \). Transforming back from the reduced structural model (3), we write the original structural model as
\(G_0(z_i; \beta) = e_i.\)

For initial values, we suggest using \(\beta^{(0)}\) obtained by minimizing
\[
\sum_{i=1}^{n} G_0'(z_i; \beta) G_0(z_i; \beta),
\]
and \(\Sigma^{(0)}\) given by
\[
\Sigma^{(0)} = \frac{1}{n} \sum_{i=1}^{n} G_0(z_i; \beta^{(0)}) G_0'(z_i; \beta^{(0)}).
\]

4. Standard error estimation

The computation of the estimated covariance matrix for the PMLE was discussed in Parke [12]. Let \((\theta_1^0, \theta_2^0)\) be the true values for \((\theta_1, \theta_2)\), and let the information matrix for \((\theta_1, \theta_2)\) at \((\theta_1^0, \theta_2^0)\) for the full likelihood be denoted by
\[
\Sigma = \left(\begin{array}{cc}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{array}\right),
\]
partitioned corresponding to \((\theta_1, \theta_2)\). Parke [12] showed that if
\[
\sqrt{n}(\hat{\theta}_2 - \theta_2^0) \xrightarrow{L} N(0, \Omega), \quad \text{as} \quad n \to \infty,
\]
then the PMLE \(\hat{\theta}_1\) satisfies
\[
\sqrt{n}(\hat{\theta}_1 - \theta_1^0) \xrightarrow{L} N(0, \Omega),
\]
where \(n\) is the sample size, and
\[
\Omega = \Sigma_{11}^{-1} + \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22} \Sigma_{11}^{-1}.
\]

Since an estimator \(\hat{\Upsilon}\) of \(\Upsilon\) can not be obtained easily, we suggest using a nonparametric bootstrap covariance matrix to estimate \(\Upsilon\). To estimate \(\Sigma_{11}\) and \(\Sigma_{12}\) in Eq. (9), we use an approximation to the expected information matrix, as described in McLachlan and Krishnan [11, pp. 120–122]. The observed data log-likelihood is
\[
\sum_{i=1}^{n} \log f(Z_i; \theta) = \sum_{i=1}^{n} \log \int f(Z_i|w_i; \theta)f(w_i; \theta) dw_i,
\]
and the corresponding individual score vector is
\[
s(Z_i; \theta) = \partial \log f(Z_i; \theta)/\partial \theta.
\]

We propose to use an estimator of the form
\[
\sum_{i=1}^{n} s(Z_i; \hat{\theta}) s(Z_i; \hat{\theta}),
\]
using our estimator \(\hat{\theta}\), and to extract \(\hat{\Sigma}_{11}\) and \(\hat{\Sigma}_{22}\) parts. It can be shown that \(s(Z_i; \theta) = E\{\partial l_i(\theta)/\partial \theta|Z; \theta\}\), where \(l_i(\theta) = \log f(Z_i, w_i; \theta)\). Thus \(\hat{s}(Z_i; \hat{\theta})\) can be computed using Monte Carlo method, with \(\{\hat{w}_{im}: m = 1, 2, \ldots, M^{(j)}\}\) and \(\{\hat{a}_{im}: m = 1, 2, \ldots, M^{(j)}\}\) obtained in the last step of the MCEM algorithm. Then, \(\hat{\Sigma}_{11}\) and \(\hat{\Sigma}_{12}\) can be obtained in
\[
\sum_{i=1}^{n} \hat{s}(Z_i; \hat{\theta}) s(Z_i; \hat{\theta}).
\]

Combining \(\hat{\Sigma}_{11}, \hat{\Sigma}_{12}, \text{and} \hat{\Upsilon}\), we obtain our estimate of the asymptotic covariance matrix of the PMLE \(\hat{\theta}_1\) as
\[
n^{-1} \left[\hat{\Sigma}_{11}^{-1} + \hat{\Sigma}_{11}^{-1} \hat{\Sigma}_{12} \hat{\Upsilon} \hat{\Sigma}_{22} \hat{\Sigma}_{11}^{-1}\right].
\]


Table 1

<table>
<thead>
<tr>
<th></th>
<th>normal</th>
<th>mixture</th>
<th>uniform</th>
<th>chi-square</th>
</tr>
</thead>
<tbody>
<tr>
<td>bias</td>
<td>0.004</td>
<td>0.053</td>
<td>0.071</td>
<td>−0.033</td>
</tr>
<tr>
<td>rmse</td>
<td>0.247</td>
<td>0.290</td>
<td>0.273</td>
<td>0.424</td>
</tr>
<tr>
<td>cp</td>
<td>0.975</td>
<td>0.965</td>
<td>0.939</td>
<td>0.952</td>
</tr>
</tbody>
</table>

Table 2

<table>
<thead>
<tr>
<th></th>
<th>ML</th>
<th>PML</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>0.025 (0.222)</td>
<td>0.004 (0.247)</td>
</tr>
<tr>
<td>Chi-squared</td>
<td>−0.866 (1.548)</td>
<td>−0.033 (0.424)</td>
</tr>
</tbody>
</table>

5. Simulation study

We considered a quadratic structural model

\[ y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \tag{11} \]

with one observed indicator \( Y_i \) for \( y_i \). For \( x_i \), there were three observed indicators satisfying Eq. (5). The sample size was \( n = 200 \), and all error terms were generated as normal random variables. For \( x_i \), we used four distributions with similar first moments; normal, normal-mixture, uniform and chi-square. For each distribution type, 1000 samples were generated.

Two model fitting procedures were applied to each sample. One was the pseudo likelihood MCEM approach using a normal mixture with nine components as the assumed distribution of \( x_i \). The other method was the maximum likelihood estimation assuming \( x_i \) is normal. The computation for the latter was not trivial, and used a MCEM algorithm developed following the development of this paper.

Table 1 presents the empirical bias, root mean square error, and the empirical coverage probability of the nominal 95% confidence interval for \( \beta_2 \) using our standard error estimate. Here we see that the standard error estimates are reasonably accurate.

Table 2 presents the empirical bias and root mean squared error for the two estimators of the quadratic coefficient \( \beta_2 \). From the results for normal case, we see that the efficiency loss due to the use of our PMLE is negligible when the true distribution happens to be normal. This may be partly due to the numerical stability of the PML achieved by performing optimization in a much smaller dimension. For the case with chi-square distribution, the full maximum likelihood estimator under normality is badly biased and inefficient.

6. Discussion

In this article we formulated a general structural equation model with a linear measurement model and a nonlinear and non-polynomial structural model. For such a model, a parameterization useful for identification and interpretation is presented. For model fitting and parameter inferences, the pseudo maximum likelihood approach is considered. A method for obtaining parameter estimates and their asymptotic covariance matrix estimate is developed based on a new version of the Monte Carlo EM algorithm. The performance of the algorithm is examined using simulation studies. Our simulation shows that the performance of our algorithm is reasonably good.

The PMLE algorithm is somewhat time-consuming. For the quadratic model with sample size \( n = 200 \) in the Section 5, on the average each sample required approximately 4 hours on a DEC Alpha workstation with 256 MB RAM and 333 MHZ CPU. Our simulation program was written in C Language with some of the functions called in from NAG. The C programs are available upon request.

Recently, Zhao and Rahardja [15] proposed a similar model but assuming \( x_i \) have a normal distribution. The validity of such a normality assumption can be checked by examining its indicators \( X_i \). If \( X_i \) is found to be not
normally distributed (e.g., skewness), then transformations of $X_i$ should be performed so that the new variables can be assumed to be normally distributed. Then Zhao and Rahardja [15] method can be applied. If no transformation can be found, then the method developed in this paper should be utilized to analyze this particular data.

Acknowledgment

The authors thank an anonymous referee for the careful review and insightful comments which have improved the presentation of this paper.

Appendix

In this section we provide details on the E-step and M-step in our MCEM algorithm. First, our MCEM E-step gives

$$
\hat{E}^{(j)}(\theta_1) = \hat{E}^{(j)}_1(\beta) + \hat{E}^{(j)}_2(\Sigma_e),
$$

where

$$
\hat{E}^{(j)}_1(\beta) = -\frac{1}{2} \sum_{i=1}^{n} \sum_{m=1}^{M^{(j)}} \tilde{d}'_{im} \Psi^{-1}_{m} \tilde{d}_{im} a_{im},
$$

$$
\hat{E}^{(j)}_2(\Sigma_e) = \sum_{i=1}^{n} \sum_{m=1}^{M^{(j)}} (\log f(\tilde{e}_{im}; \Sigma_e)) a_{im},
$$

$$
a_{im} = \frac{f(\tilde{y}_i|\tilde{w}_{im}; \beta^{(j)}; \hat{\theta}_m) f(\tilde{x}_i|\tilde{x}_{im}; \hat{\theta}_m)}{\sum_{m=1}^{M^{(j)}} f(\tilde{y}_i|\tilde{w}_{im}; \beta^{(j)}; \hat{\theta}_m) f(\tilde{x}_i|\tilde{x}_{im}; \hat{\theta}_m)},
$$

$$
\tilde{d}_{im} = \tilde{y}_i - G(\tilde{w}_{im}; \beta).
$$

The M-step in the $(j + 1)^{st}$ iteration finds $\beta^{(j+1)}$ and $\Sigma_e^{(j+1)}$ that maximize $E^{(j)}(\theta_1)$ in Eq. (12). Our way of formulating the E-step also simplifies the M-step, because the parameters $\beta$ and $\Sigma_e$ in Eq. (12) are separated into two terms, and because each of $\hat{E}^{(j)}_1(\beta)$ and $\hat{E}^{(j)}_2(\Sigma_e)$ is a weighted sum of the log-transformed normal densities with given weights $a_{im}$. Thus, $\beta^{(j+1)}$ is obtained by the multivariate nonlinear weighted least squares minimizing the summation term in $\hat{E}^{(j)}_1(\beta)$, and $\Sigma_e^{(j+1)}$ can be explicitly obtained as

$$
\Sigma_e^{(j+1)} = \frac{1}{n} \sum_{i=1}^{n} \sum_{m=1}^{M^{(j)}} \tilde{e}'_{im} \tilde{e}_{im} a_{im},
$$

where we have used $\sum_{m=1}^{M^{(j)}} a_{im} = 1$ for every $i$. If some covariance elements of $\Sigma_e$ are known to be zero (e.g., diagonal), then the corresponding elements of $\Sigma_e^{(j+1)}$ are set to be zero. The above E-step and M-step are to be iterated to obtain the PMLE $\beta$ and $\Sigma_e$.

References