Statistica Sinica **8**(1998), 543-557

ANALYSIS OF COVARIANCE STRUCTURES WITH INDEPENDENT AND NON-IDENTICALLY DISTRIBUTED OBSERVATIONS

Sik-Yum Lee and Jian-Qing Shi

Chinese University of Hong Kong

Abstract: The main purpose of this paper is to develop some asymptotic properties in analysis of covariance structures with independent but non-identically distributed observations. Consistency and asymptotic normality for the maximum likelihood estimation will be developed. Analogous results for the generalized least squares estimation will be provided. Applications of the general theory to two specific models are discussed. Finally, results from a simulation study will be presented to illustrate the theory developed.

Key words and phrases: Analysis of covariance structures, asymptotic normality, consistency, information matrix, missing data, multi-level structural equation model.

1. Introduction

Analysis of covariance structures (Jöreskog (1978), Browne (1982), Bentler (1980)) is an important multivariate method in analyzing behavioral, medical and social science data. It has been shown by Jöreskog (1970) that this method covers multivariate analysis of variance, regression, principal component and factor analysis as special cases. In its general setting, the method involves the covariance structure that corresponds to a system of structural equations which relate the observed variables with the latent variables. This covariance structure $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is a function of a vector of parameters $\boldsymbol{\theta}$, which contains all the unknown parameters in the model: the unknown coefficients of the system and the variances/covariances of the latent variables. Traditionally, one of the main objectives is to estimate $\boldsymbol{\theta}$ based on a random sample of independently and identically distributed (i.i.d.) observations from $N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$. However, in many situations, the assumption of identically distributed observations is not satisfied. In the literature, there are asymptotic results for the independent non-identically distributed (i.n.i.d.) case in the general context of maximum likelihood estimation (see for example Hoadley (1971)). However, these results were established under certain conditions which unfortunately are difficult to verify. Hence, in dealing with special models, it is necessary to identify more flexible conditions that are easier to verify (see, for examples, Wu (1981) for nonlinear regression models; Fahrmeir and Kaufmann (1985) for the generalized linear model; Nordberg (1980) for estimation of exponential family models, etc). In this paper, the corresponding problem will be investigated in the context of covariance structures analysis. We will first develop some asymptotic properties such as consistency and asymptotic normality for the maximum likelihood estimation for i.n.i.d. observations. Then, analogous results for the generalized least squares estimation will be provided. Applications of the general theory to two specific models are discussed and results from a simulation study are presented to illustrate the theory developed.

In the following sections, for any p by p matrix $\boldsymbol{A}$, $\mathrm{Vec}(\boldsymbol{A})$ represents the p^2 by 1 vector which formed by stacking all the elements of $\boldsymbol{A}$ row by row sequentially.

2. Maximum Likelihood Estimation

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_G$ be i.n.i.d. observations such that $\boldsymbol{X}_g$ is distributed as $N(\mathbf{0}, \boldsymbol{\Sigma}_g(\boldsymbol{\theta}_0))$, where $\boldsymbol{\theta}_0$ is an unknown $q \times 1$ parameter vector, and the dimension of $\boldsymbol{X}_g$ is N_g. The maximum likelihood (ML) estimate of $\boldsymbol{\theta}_0$, $\hat{\boldsymbol{\theta}}_G$, is the vector that minimizes the following log-likelihood function

$$l_G(\boldsymbol{\theta}) = \frac{1}{2}\sum_{g=1}^{G}\Big[\log\Big|\boldsymbol{\Sigma}_g(\boldsymbol{\theta})\Big| + \boldsymbol{X}_g'\boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\theta})\boldsymbol{X}_g\Big]. \tag{1}$$

Differentiating $l_G(\boldsymbol{\theta})$ twice, it can be shown that the gradient vector and the Hessian matrix are respectively equal to

$$\dot{\boldsymbol{l}}_G(\boldsymbol{\theta}) = \frac{\partial l_G(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\frac{1}{2}\sum_{g=1}^{G}\Big[\boldsymbol{\Delta}_g(\boldsymbol{\theta})[\boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\theta})]\mathrm{Vec}[\boldsymbol{X}_g\boldsymbol{X}_g' - \boldsymbol{\Sigma}_g(\boldsymbol{\theta})]\Big], \tag{2}$$

and

$$\begin{aligned}\boldsymbol{H}_G(\boldsymbol{\theta}) = \frac{\partial \dot{\boldsymbol{l}}_G(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{1}{2}\sum_{g=1}^{G}\Big[\boldsymbol{\Delta}_g(\boldsymbol{\theta})[\boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\theta})]\boldsymbol{\Delta}_g'(\boldsymbol{\theta})\Big] \\ &\quad -\frac{1}{2}\sum_{g=1}^{G}\Big[\boldsymbol{W}_g(\boldsymbol{\theta})\{\boldsymbol{I}_q \otimes \mathrm{Vec}[\boldsymbol{X}_g\boldsymbol{X}_g' - \boldsymbol{\Sigma}_g(\boldsymbol{\theta})]\}\Big],\end{aligned} \tag{3}$$

where $\boldsymbol{I}_q$ is an $q \times q$ identity matrix,

$$\boldsymbol{\Delta}_g(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{\Sigma}_g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \quad \text{and} \quad \boldsymbol{W}_g(\boldsymbol{\theta}) = \frac{\partial\{\boldsymbol{\Delta}_g(\boldsymbol{\theta})[\boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\theta})]\}}{\partial \boldsymbol{\theta}}.$$

Let $\boldsymbol{J}_G(\boldsymbol{\theta})$ be the information matrix, and $\boldsymbol{R}_G(\boldsymbol{\theta}) = \boldsymbol{J}_G(\boldsymbol{\theta}) - \boldsymbol{H}_G(\boldsymbol{\theta})$. From (3),

$$\boldsymbol{J}_G(\boldsymbol{\theta}) = \mathrm{E}[\boldsymbol{H}_G(\boldsymbol{\theta})] = \frac{1}{2}\sum_{g=1}^{G}\Big[\boldsymbol{\Delta}_g(\boldsymbol{\theta})[\boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\theta})]\boldsymbol{\Delta}_g'(\boldsymbol{\theta})\Big].$$

For simplicity, let $\boldsymbol{J}_G = \boldsymbol{J}_G(\boldsymbol{\theta}_0)$, $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}_g(\boldsymbol{\theta}_0)$ and $\boldsymbol{\Delta}_g = \boldsymbol{\Delta}_g(\boldsymbol{\theta}_0)$.

The asymptotic properties of the maximum likelihood estimation based on i.n.i.d. observations will be derived with the following conditions.

Condition (A). (a) All elements of $\boldsymbol{\Sigma}_g(\theta)$ and all their partial derivatives of the first three orders with respect to elements of $\boldsymbol{\theta}$ are continuous and bounded uniformly for all g in a neighborhood of $\boldsymbol{\theta}_0$; (b) $\boldsymbol{\theta}$ is identifiable, i.e., $\boldsymbol{\Sigma}_g(\boldsymbol{\theta}_1) = \boldsymbol{\Sigma}_g(\boldsymbol{\theta}_2)$ for all g implies that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$; and (c) $\boldsymbol{\Sigma}_g(\boldsymbol{\theta}_0)$ is positive definite for all g.

Condition (B). There exist a neighborhood of $\boldsymbol{\theta}_0$: $N(\delta) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \delta\}$, and positive constants c and G^* such that for all $G > G^*$ and all $\boldsymbol{\theta}$ in $N(\delta)$, $\boldsymbol{\lambda}_{\min}(\boldsymbol{J}_G(\boldsymbol{\theta})/G) \geq c$, where $\lambda_{\min}(\boldsymbol{J}_G(\boldsymbol{\theta})/G)$ is the minimum eigenvalue of the matrix $\boldsymbol{J}_G(\boldsymbol{\theta})/G$.

Condition (A) includes some mild regularity assumptions which are usually satisfied in practice. Since $2^{-1}\boldsymbol{\Delta}_g(\boldsymbol{\theta})[\boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\theta})]\boldsymbol{\Delta}_g'(\boldsymbol{\theta})$ is the information matrix corresponding to $\boldsymbol{X}_g$, $\boldsymbol{J}_G(\boldsymbol{\theta})/G$ can be interpreted as the mean of the information matrices. Under Condition (B), the minimum eigenvalue of this mean of the information matrices is larger than a positive constant within a neighborhood of $\boldsymbol{\theta}_0$ for all $G > G^*$. This means that there is enough information to describe every unknown parameter as the sample size gets sufficiently large.

Condition (C). There exists a neighborhood of $\boldsymbol{\theta}_0$, $N(\delta)$, such that for all $\boldsymbol{\theta}$ in $N(\delta)$, $\lim_{G\to\infty}(\boldsymbol{J}_G(\boldsymbol{\theta})/G) = \boldsymbol{K}(\boldsymbol{\theta})$, where $\boldsymbol{K}(\boldsymbol{\theta})$ is a positive definite matrix.

Under Conditions (A) and (C), there exist a sufficient large number G_1 and a neighborhood $N(\delta)$, such that for any $\varepsilon > 0$ and $\boldsymbol{\eta} \neq \mathbf{0}$,

$$\begin{aligned}
&|\frac{1}{G}\boldsymbol{\eta}'\boldsymbol{J}_G(\boldsymbol{\theta})\boldsymbol{\eta} - \boldsymbol{\eta}'\boldsymbol{K}(\boldsymbol{\theta}_0)\boldsymbol{\eta}| \\
&\leq |\frac{1}{G}\boldsymbol{\eta}'\boldsymbol{J}_G(\boldsymbol{\theta})\boldsymbol{\eta} - \frac{1}{G}\boldsymbol{\eta}'\boldsymbol{J}_G(\boldsymbol{\theta}_0)\boldsymbol{\eta}| + |\frac{1}{G}\boldsymbol{\eta}'\boldsymbol{J}_G(\boldsymbol{\theta}_0)\boldsymbol{\eta} - \boldsymbol{\eta}'\boldsymbol{K}(\boldsymbol{\theta}_0)\boldsymbol{\eta}| < \varepsilon,
\end{aligned}$$

when $G > G_1$ and $\boldsymbol{\theta} \in N(\delta)$. Therefore, we can choose an ε such that $G^{-1}\boldsymbol{\eta}'\boldsymbol{J}_G(\boldsymbol{\theta})\boldsymbol{\eta} > \boldsymbol{\eta}'\boldsymbol{K}(\boldsymbol{\theta}_0)\boldsymbol{\eta} - \varepsilon > 0$. Hence, Condition (C) implies Condition (B).

The consistency of $\hat{\boldsymbol{\theta}}_G$ will be established based on the following lemmas.

Lemma 1. *If $\{y_g, g = 1, 2, \ldots\}$ is a sequence of mutually independent random variables with $E(y_g) = 0$ and $\mathrm{Var}\,(y_g) = \sigma_g^2$, then*

$$\lim_{G\to\infty} \frac{1}{G}\sum_{g=1}^{G} \sigma_g^2 < +\infty \textit{ implies } \lim_{G\to\infty} \frac{1}{G}\sum_{g=1}^{G} y_g = 0, \textit{ a.s..}$$

In the above lemma, 'a.s.' stands for almost surely, which means that $\lim_{G\to\infty}\sum_{g=1}^{G} y_g/G = 0$ with probability one. This lemma can be proved immediately if we take $n = G, A_n = G$ and $\delta = 1/2$ in Wu's (1981) Lemma 2.

Lemma 2. *Under Condition* (A),*there exists a neighborhood of* $\boldsymbol{\theta}_0$*, such that* $\lim_{G\to\infty} G^{-1}\dot{\boldsymbol{l}}_G(\boldsymbol{\theta}) = 0$*, a.s., for any* $\boldsymbol{\theta}$ *in this neighborhood.*

Proof. The jth element of $G^{-1}\dot{\boldsymbol{l}}_G(\boldsymbol{\theta})$ is equal to $G^{-1}\boldsymbol{d}_j'\dot{\boldsymbol{l}}_G(\boldsymbol{\theta}) = -(2G)^{-1}\sum_{g=1}^{G} y_{gj}$, where $\boldsymbol{d}_j$ denotes a $q \times 1$ unit vector with the jth element equal to 1 and zero elsewhere, and $y_{gj} = \boldsymbol{d}_j'\{\boldsymbol{\Delta}_g(\boldsymbol{\theta})[\boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\theta})]\mathrm{Vec}[\boldsymbol{X}_g\boldsymbol{X}_g' - \boldsymbol{\Sigma}_g(\boldsymbol{\theta})]\}$. Hence, y_{gj} are mutually independent for $g = 1, \ldots, G$. Since $\boldsymbol{X}_g$ is distributed as a normal distribution $N(\boldsymbol{0}, \boldsymbol{\Sigma}_g)$, $\boldsymbol{X}_g\boldsymbol{X}_g'$ is distributed as a Wishart distribution, $W(\boldsymbol{\Sigma}_g, N_g, 1)$. The mean and covariance matrix of $\boldsymbol{X}_g\boldsymbol{X}_g'$ are $\boldsymbol{\Sigma}_g$ and $2(\boldsymbol{\Sigma}_g \otimes \boldsymbol{\Sigma}_g)$, respectively (see Eaton (1983), p305). Thus, $E(y_{gj}) = 0$ and

$$\mathrm{Var}\,(y_{gj}) = 2\boldsymbol{d}_j'\boldsymbol{\Delta}_g(\boldsymbol{\theta})[\boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\theta})](\boldsymbol{\Sigma}_g \otimes \boldsymbol{\Sigma}_g)[\boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\theta})]\boldsymbol{\Delta}_g'(\boldsymbol{\theta})\boldsymbol{d}_j.$$

From (a) and (c) of Condition (A), there exist a neighborhood of $\boldsymbol{\theta}_0$ and a real number M such that M is independent of g and $\mathrm{Var}(y_{gj}) \le M$ for all θ in this neighborhood. Therefore, $\sum_{g=1}^{G}\mathrm{Var}(y_{gj})/G \le M < \infty$. From Lemma 1, $\lim_{G\to\infty} G^{-1}\sum_{g=1}^{G} y_{gj} = 0$, a.s.. Since j is taken from 1 to q, this lemma is proved.

Let $\bar{N}(\delta) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \le \delta\}$ and $\partial N(\delta) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| = \delta\}$ be the closure and the boundary of $N(\delta)$ respectively; then based on reasoning similar to above, the following lemma is valid.

Lemma 3. *Under Condition* (A), *there exist a positive integer* G_1 *and a* $\delta > 0$ *such that* $Pr\{\sup_{\boldsymbol{\theta}\in\bar{N}(\delta)} \|\boldsymbol{R}_G(\boldsymbol{\theta})\|/G < \varepsilon\} = 1$ *for any* $G > G_1$*.*

Theorem 1. *Under Conditions* (A) *and* (B), *there exists a sequence* $\{\hat{\boldsymbol{\theta}}_G, G = 1, 2, \ldots\}$ *of random vectors such that* (i) $\lim_{G\to\infty} Pr\{\dot{\boldsymbol{l}}_G(\hat{\boldsymbol{\theta}}_G) = \boldsymbol{0}\} = 1$*; and* (ii) $\lim_{G\to\infty} \hat{\boldsymbol{\theta}}_G = \boldsymbol{\theta}_0$*, a.s..*

Proof. From (3), $G^{-1}\boldsymbol{H}_G(\boldsymbol{\theta}) = G^{-1}\boldsymbol{J}_G(\boldsymbol{\theta}) - G^{-1}\boldsymbol{R}_G(\boldsymbol{\theta})$. Also, from Condition (B) and Lemma 3, there exists a $\delta > 0$ such that

$$\Pr\{\lambda_{\min}(G^{-1}\boldsymbol{H}_G(\boldsymbol{\theta})) \ge c - \varepsilon, G > G_1\} = 1, \tag{4}$$

for any $\boldsymbol{\theta} \in N(\delta)$. It means that $\boldsymbol{H}_G(\boldsymbol{\theta})$ is positive definite in $N(\delta)$. Thus, the event

$$l_G(\boldsymbol{\theta}) - l_G(\boldsymbol{\theta}_0) > 0, \text{ for all } \boldsymbol{\theta} \in \partial N(\delta) \text{ and } G > G_1 \tag{5}$$

implies the existence of a local minimum inside $N(\delta)$. This minimum must be located at $\hat{\boldsymbol{\theta}}_G$, so (5) implies (i). Moreover, since δ is arbitrary, (ii) can be deduced from (5) with a sufficiently small δ. Hence, it suffices to prove that (5) is true with probability one.

Let $\boldsymbol{\tau} = (\boldsymbol{\theta} - \boldsymbol{\theta}_0)/\delta$; then $\boldsymbol{\theta} \in \partial N(\delta)$ if and only if the norm of $\boldsymbol{\tau}$ is equal to 1. By Taylor's expansion, $l_G(\boldsymbol{\theta}) - l_G(\boldsymbol{\theta}_0) = \delta\boldsymbol{\tau}'\dot{\boldsymbol{l}}_G(\boldsymbol{\theta}_0) + 2^{-1}\delta^2\boldsymbol{\tau}'\boldsymbol{H}_G(\hat{\boldsymbol{\theta}}_G^*)\boldsymbol{\tau}$, where

$\hat{\boldsymbol{\theta}}_G^* = t_G\boldsymbol{\theta}_0 + (1 - t_G)\boldsymbol{\theta}$ with $0 \le t_G \le 1$. For simplicity, let $\dot{\boldsymbol{l}}_G = \dot{\boldsymbol{l}}_G(\boldsymbol{\theta}_0)$; hence (5) is equivalent to $|\boldsymbol{\tau}'\dot{\boldsymbol{l}}_G|/G < 2^{-1}\delta\boldsymbol{\tau}'(\boldsymbol{H}_G(\hat{\boldsymbol{\theta}}_G^*)/G)\boldsymbol{\tau}$ for any $\boldsymbol{\tau}$ with $\|\boldsymbol{\tau}\| = 1$ and $G > G_1$. By the Cauchy-Schwarz inequality, $|\boldsymbol{\tau}'\dot{\boldsymbol{l}}_G|/G \le \|\boldsymbol{\tau}\|\|\dot{\boldsymbol{l}}_G\|/G$. Hence, from Lemma 2, $|\boldsymbol{\tau}'\dot{\boldsymbol{l}}_G|/G$ converges to zero with probability one. Combining this result with (4), it can be seen that (5) is true with probability one.

Lemma 4. *Under Conditions* (A) *and* (C), $\boldsymbol{J}_G^{-1/2}\dot{\boldsymbol{l}}_G \xrightarrow{L} N(\boldsymbol{0}, \boldsymbol{I}_q)$, *where* $\xrightarrow{L}$ *denotes convergence in distribution.*

Proof. From (2), it can be shown that $\boldsymbol{J}_G^{-1/2}\dot{\boldsymbol{l}}_G = \sum_{g=1}^G \boldsymbol{B}_g\mathrm{Vec}(\boldsymbol{Y}_g - I_{N_g})$, where $\boldsymbol{B}_g = -2^{-1}\boldsymbol{J}_G^{-1/2}\boldsymbol{\Delta}_g(\boldsymbol{\Sigma}_g^{-1/2} \otimes \boldsymbol{\Sigma}_g^{-1/2})$ and $\boldsymbol{Y}_g = \boldsymbol{\Sigma}_g^{-1/2}\boldsymbol{X}_g\boldsymbol{X}_g'\boldsymbol{\Sigma}_g^{-1/2}$. Clearly, the distribution of $\boldsymbol{Y}_g$ is Wishart $W(I_{N_g}, N_g, 1)$, and its covariance matrix is $2(\boldsymbol{I}_{N_g} \otimes \boldsymbol{I}_{N_g})$. Hence,

$$\mathrm{Var}\,(\boldsymbol{J}_G^{-1/2}\dot{\boldsymbol{l}}_G) = \frac{1}{2}\sum_{g=1}^{G}\boldsymbol{J}_G^{-1/2}\boldsymbol{\Delta}_g(\boldsymbol{\Sigma}_g^{-1} \otimes \boldsymbol{\Sigma}_g^{-1})\boldsymbol{\Delta}_g'\boldsymbol{J}_G^{-1/2} = \boldsymbol{J}_G^{-1/2}\boldsymbol{J}_G\boldsymbol{J}_G^{-1/2} = \boldsymbol{I}_q.$$

Moreover, $\|\boldsymbol{B}_g\| \le (2G^{1/2})^{-1}\|(\boldsymbol{J}_G/G)^{-1/2}\|\|\boldsymbol{\Delta}_g(\boldsymbol{\Sigma}_g^{-1/2} \otimes \boldsymbol{\Sigma}_g^{-1/2})\|$ by the Cauchy-Schwarz inequality. Under Conditions (A) and (C), $\|\boldsymbol{B}_g\|$ tends to zero uniformly for all g as G tends to infinity. So, for every $\varepsilon > 0$,

$$\int_{\|\boldsymbol{B}_g\mathrm{Vec}(\boldsymbol{Y}_g - \boldsymbol{I}_{N_g})\| > \varepsilon}\|\boldsymbol{B}_g\mathrm{Vec}(\boldsymbol{Y}_g - \boldsymbol{I}_{N_g})\|^2 dF_i \le \|\boldsymbol{B}_g\|^2\int\|\mathrm{Vec}(\boldsymbol{Y}_g - \boldsymbol{I}_{N_g})\|^2 dF_i.$$

Since the distribution of $\boldsymbol{Y}_g$ is Wishart, $W(\boldsymbol{I}_{N_g}, N_g, 1)$, and N_g is finite and bounded for all g, $\|\mathrm{Vec}(\boldsymbol{Y}_g - \boldsymbol{I}_{N_g})\|^2$ is uniformly bounded and the right hand side of the above inequality tends to zero uniformly with probability one. Therefore, based on the multivariate central limit theorem (Rao (1973), p147), the lemma is proved.

Theorem 2. *Under Conditions* (A) *and* (C), $G^{1/2}(\hat{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_0) \xrightarrow{L} N(\boldsymbol{0}, \boldsymbol{K}^{-1}(\boldsymbol{\theta}_0))$.

Proof. Using Taylor's expansion, it can be shown that

$$\boldsymbol{J}_G^{-1/2}\dot{\boldsymbol{l}}_G(\boldsymbol{\theta}_0) = [\boldsymbol{J}_G^{-1/2}\boldsymbol{H}_G(\hat{\boldsymbol{\theta}}_G^*)\boldsymbol{J}_G^{-1/2}][\boldsymbol{J}_G^{1/2}(\hat{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_0)],$$

for some $\hat{\boldsymbol{\theta}}_G^*$ between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_G$. Hence,

$$G^{1/2}(\hat{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_0) = (\boldsymbol{J}_G/G)^{-1/2}\boldsymbol{S}_G^{-1}\boldsymbol{J}_G^{-1/2}\dot{\boldsymbol{l}}_G(\boldsymbol{\theta}_0), \tag{6}$$

where $\boldsymbol{S}_G = \boldsymbol{J}_G^{-1/2}\boldsymbol{H}_G(\hat{\boldsymbol{\theta}}_G^*)\boldsymbol{J}_G^{-1/2} = (\boldsymbol{J}_G/G)^{-1/2}[\boldsymbol{J}_G(\hat{\boldsymbol{\theta}}_G^*)/G - \boldsymbol{R}_G(\hat{\boldsymbol{\theta}}_G^*)/G](\boldsymbol{J}_G/G)^{-1/2}$. It follows from Lemma 3 that $\boldsymbol{R}_G(\hat{\boldsymbol{\theta}}_G^*)/G$ tends to zero. Also, since $\hat{\boldsymbol{\theta}}_G$ tends to $\boldsymbol{\theta}_0$, $\hat{\boldsymbol{\theta}}_G^*$ tends to $\boldsymbol{\theta}_0$ as well. Therefore, $\boldsymbol{J}_G(\hat{\boldsymbol{\theta}}_G^*)/G$ tends to $\boldsymbol{K}(\boldsymbol{\theta}_0)$, and $\boldsymbol{S}_G$ tends to $\boldsymbol{I}_q$. From (6), Lemma 4 and Condition (C), the theorem is proved.

It should be noted that if $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_G$ are i.i.d. random observations, we have $\boldsymbol{\Sigma}_g(\boldsymbol{\theta}) = \boldsymbol{\Sigma}(\boldsymbol{\theta})$, $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$, $N_g = p$, $\boldsymbol{\Delta}_g(\boldsymbol{\theta}) = \boldsymbol{\Delta}(\boldsymbol{\theta})$ and $\boldsymbol{K}(\boldsymbol{\theta}) = \boldsymbol{J}_G(\boldsymbol{\theta})/G = 2^{-1}\boldsymbol{\Delta}(\boldsymbol{\theta})[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})]\boldsymbol{\Delta}'(\boldsymbol{\theta})$. Hence, Theorem 2 implies that $G^{1/2}(\hat{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_0)$ converges to $N(\mathbf{0},\ 2[\boldsymbol{\Delta}(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{\Delta}']^{-1})$ in distribution. This gives the classical result in the literature (e.g. Browne (1982)). Moreover, it can be shown that if there are only a finite number of distinct distributions, results developed here reduced back to the classical theory of covariance structures analysis in several populations (e.g. Jöreskog and Sörbom (1989)).

3. Generalized Least Squares Estimation

Another important estimation method in covariance structures analysis is the generalized least squares (GLS) approach. Consider the following generalized least squares objective function:

$$Q_G(\boldsymbol{\theta}) = \frac{1}{2}\sum_{g=1}^{G} \mathrm{tr}\Big[[\boldsymbol{X}_g\boldsymbol{X}_g' - \boldsymbol{\Sigma}_g(\boldsymbol{\theta})]\boldsymbol{V}_g\Big]^2,$$

where $\boldsymbol{V}_g$ is some $N_g \times N_g$ positive definite weight matrix. The generalized least squares (GLS) estimate $\tilde{\boldsymbol{\theta}}_G$ of $\boldsymbol{\theta}_0$ is the vector that minimizes $Q_G(\boldsymbol{\theta})$. The first two derivatives of $Q_G(\boldsymbol{\theta})$ are respectively

$$\dot{\boldsymbol{Q}}_G(\boldsymbol{\theta}) = \frac{\partial Q_G(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\sum_{g=1}^{G}\Big[\boldsymbol{\Delta}_g(\boldsymbol{\theta})(\boldsymbol{V}_g \otimes \boldsymbol{V}_g)\mathrm{Vec}[\boldsymbol{X}_g\boldsymbol{X}_g' - \boldsymbol{\Sigma}_g(\boldsymbol{\theta})]\Big], \tag{7}$$

and $\ddot{\boldsymbol{Q}}_G(\boldsymbol{\theta}) = \partial\dot{\boldsymbol{Q}}_G(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \boldsymbol{J}_{QG}(\boldsymbol{\theta}) - \boldsymbol{R}_{QG}(\boldsymbol{\theta})$, where

$$\boldsymbol{R}_{QG}(\boldsymbol{\theta}) = \sum_{g=1}^{G}\Big\{\nabla_g(\boldsymbol{\theta})\{\boldsymbol{I}_q \otimes [(\boldsymbol{V}_g \otimes \boldsymbol{V}_g)\mathrm{Vec}(\boldsymbol{X}_g\boldsymbol{X}_g' - \boldsymbol{\Sigma}_g(\boldsymbol{\theta}))]\}\Big\},$$

$$\boldsymbol{J}_{QG}(\boldsymbol{\theta}) = \sum_{g=1}^{G}\Big(\boldsymbol{\Delta}_g(\boldsymbol{\theta})(\boldsymbol{V}_g \otimes \boldsymbol{V}_g)\boldsymbol{\Delta}_g'(\boldsymbol{\theta})\Big), \text{ and } \nabla g(\boldsymbol{\theta}) = \partial\boldsymbol{\Delta}_g(\boldsymbol{\theta})/\partial\boldsymbol{\theta}.$$

The consistency and asymptotic normality of $\tilde{\boldsymbol{\theta}}_G$ can be established with the following conditions similar to those in the ML case:

Condition (B'). There exist a neighborhood $N(\delta)$ of $\boldsymbol{\theta}_0$, and positive constants c and G_1 such that $\lambda_{\min}(\boldsymbol{J}_{QG}(\boldsymbol{\theta})/G) \geq c$, for all $G > G_1$ and $\boldsymbol{\theta} \in N(\delta)$.

Condition (C'). There exists a neighborhood $N(\delta)$ of $\boldsymbol{\theta}_0$ such that for all $\boldsymbol{\theta} \in N(\delta)$,

$$\lim_{G\to\infty}[\boldsymbol{J}_{QG}(\theta)/(2G)] = \boldsymbol{M}_1(\boldsymbol{\theta}), \text{ and } \lim_{G\to\infty}[\boldsymbol{J}_{VG}(\theta)/(2G)] = \boldsymbol{M}_2(\boldsymbol{\theta}),$$

where $\boldsymbol{M}_1(\boldsymbol{\theta})$ and $\boldsymbol{M}_2(\boldsymbol{\theta})$ are positive definite matrices, and

$$\boldsymbol{J}_{VG}(\boldsymbol{\theta}) = \sum_{g=1}^{G} \Big[\boldsymbol{\Delta}_g(\boldsymbol{\theta})[(\boldsymbol{V}_g\boldsymbol{\Sigma}_g(\boldsymbol{\theta})\boldsymbol{V}_g) \otimes (\boldsymbol{V}_g\boldsymbol{\Sigma}_g(\boldsymbol{\theta})\boldsymbol{V}_g)]\boldsymbol{\Delta}_g'(\boldsymbol{\theta})\Big].$$

Note that $\dot{\boldsymbol{Q}}_G(\boldsymbol{\theta})$ in (7) and $\dot{\boldsymbol{l}}_G(\boldsymbol{\theta})$ in (2) are very similar, the only difference is that $[\boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\theta}) \otimes \boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\theta})]$ is replaced by $[\boldsymbol{V}_g \otimes \boldsymbol{V}_g]$. Hence, interpretations of the above conditions are similar to those in the ML case. In addition, based on reasoning similar to that in the last section, the following results can be proved.

Lemma 5. *Under Condition* (A), *there exists a neighborhood $N(\delta)$ of $\boldsymbol{\theta}_0$, such that* $\lim_{G\to\infty} G^{-1}\dot{\boldsymbol{Q}}_G(\boldsymbol{\theta}) = 0$, *a.s., uniformly for any $\boldsymbol{\theta}$ in $N(\delta)$.*

Lemma 6. *Under Condition* (A), *there exist a positive integer G_1 and a $\delta > 0$ such that* $\Pr\{\sup_{\boldsymbol{\theta}\in\bar{N}(\delta)} \|\boldsymbol{R}_{QG}(\boldsymbol{\theta})\|/G < \varepsilon\} = 1$ *for any $G > G_1$.*

Theorem 3. *Under Conditions* (A) *and* (B'), *there exists a sequence $\{\tilde{\boldsymbol{\theta}}_G, G = 1, 2, \ldots\}$ of random vectors such that* (i) $\lim_{G\to\infty} \Pr\{\dot{\boldsymbol{Q}}_G(\tilde{\boldsymbol{\theta}}_G) = \boldsymbol{0}\} = 1$; *and* (ii) $\lim_{G\to\infty} \tilde{\boldsymbol{\theta}}_G = \boldsymbol{\theta}_0$, *a.s..*

It should be noted that the proof of Theorem 3 is very similar to the proof of Theorem 1, the only difference is that $l_G, \dot{\boldsymbol{l}}_G, \boldsymbol{H}_G, \boldsymbol{J}_G$ and $\boldsymbol{R}_G$ are replaced by Q_G,$\dot{\boldsymbol{Q}}_G$,$\ddot{\boldsymbol{Q}}_G$, $\boldsymbol{J}_{QG}$ and $\boldsymbol{R}_{QG}$, respectively. Also, under conditions (A) and (C'), $2^{-1/2}\boldsymbol{J}_{VG}^{-1/2}\dot{\boldsymbol{Q}}_G \xrightarrow{L} N(\boldsymbol{0}, \boldsymbol{I}_q)$. Based on this result and reasoning similar to that in the proof of Theorem 2, the following result can be established.

Theorem 4. *Under Conditions* (A) *and* (C'), $G^{1/2}(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_0) \xrightarrow{L} N(\boldsymbol{0}, \boldsymbol{M}(\boldsymbol{\theta}_0))$, *where* $\boldsymbol{M}(\boldsymbol{\theta}_0) = \boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0)\boldsymbol{M}_2(\boldsymbol{\theta}_0)\boldsymbol{M}_1^{-1}(\boldsymbol{\theta}_0)$.

The following theorem shows that as G tends to infinity, $\hat{\boldsymbol{\theta}}_G$ and $\tilde{\boldsymbol{\theta}}_G$ are very close to each other. The proof of this theorem is rather straightforward and hence omitted.

Theorem 5. *Under Conditions* (A), (C) *and* (C'), $\lim_{G\to\infty}[G^{1/2-\gamma}(\tilde{\boldsymbol{\theta}}_G - \hat{\boldsymbol{\theta}}_G)] = \boldsymbol{0}$, *a.s. for any positive real number γ.*

It should be noted that for any arbitrary weight matrix $\boldsymbol{V}_g$, $\boldsymbol{M}(\boldsymbol{\theta}_0) - \boldsymbol{K}^{-1}(\boldsymbol{\theta}_0)$ is positive definite; thus, asymptotically, the GLS estimator is less efficient than the ML estimator. However, if $\boldsymbol{V}_g$ tends to $\boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\theta}_0)$ in probability, it can be shown by reasoning similar to that in Wu (1981), Lemma 2 that $G^{1/2}(\tilde{\boldsymbol{\theta}}_G - \hat{\boldsymbol{\theta}}_G)$ tends to zero in probability, and $\boldsymbol{M}(\boldsymbol{\theta}_0) = \boldsymbol{K}^{-1}(\boldsymbol{\theta}_0)$. Under this situation, $\hat{\boldsymbol{\theta}}_G$ and $\tilde{\boldsymbol{\theta}}_G$ are asymptotically equivalent. In practice, this 'Best' GLS estimator can be obtained via the iteratively reweighted Gauss-Newton algorithm (Lee and Jennrich (1979)) in which the $\boldsymbol{V}_g$ at the jth step was iteratively set equal to

$\boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\theta}^{(j)})$ from iteration to iteration, where $\boldsymbol{\theta}^{(j)}$ is the current parameter vector at the jth step.

In this section, the GLS theory based on the normality assumption is presented. However, we expect that this GLS development can be generalized to the non-normal case as considered in Bentler (1983) and Browne (1984). Based on interpretations similar to those given in Section 2 about Conditions (B) and (C), it can be seen that the corresponding non-normal conditions for consistency and asymptotic normality can be naturally satisfied with social and behavioral science data.

4. Specific Models

In this section, the general results obtained in previous sections will be applied to two special models in covariance structures analysis. It will be seen that the conditions for obtaining the asymptotic properties can be verified to be true for these particular cases. Hence, some previous results in the literature can be generalized with the present development.

4.1. Models with missing data

Firstly, we discuss the model with missing data that are missing completely at random. Consider a p-dimensional multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}_0))$. The problem is to analyze the covariance structure $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ based on complete and incomplete random observations. Let m be the number of distinct missing patterns and $T_i, i = 1, \ldots, m$, be the number of random vectors of pattern i. Let $\boldsymbol{X}_{ij}$ be a random observation in pattern i with dimension p_i, and $\boldsymbol{F}_i$ be an $p_i \times p$ matrix with elements equal to zero or one such that $\boldsymbol{\Sigma}_{0i} = \boldsymbol{F}_i \boldsymbol{\Sigma}_0 \boldsymbol{F}_i'$ is the covariance matrix of $\boldsymbol{X}_{ij}$. Let $\boldsymbol{S}_i$ be the sample covariance matrix obtained from the observations in pattern i. The ML estimator of $\boldsymbol{\theta}$ is the vector that minimizes the following function

$$l(\boldsymbol{\theta}) = 2^{-1} \sum_{i=1}^{m} c_i \Big\{ \log|\boldsymbol{\Sigma}_i| + \mathrm{tr}(\boldsymbol{\Sigma}_i^{-1} \boldsymbol{S}_i) - \log|\boldsymbol{S}_i| - p_i \Big\},$$

where $c_i = T_i/G, G = \sum_{i=1}^m T_i, \boldsymbol{\Sigma}_i = \boldsymbol{F}_i \boldsymbol{\Sigma}(\boldsymbol{\theta}) \boldsymbol{F}_i'$. The GLS approach in estimating $\boldsymbol{\theta}$ is to minimize the function

$$Q(\boldsymbol{\theta}) = 2^{-1} \sum_{i=1}^{m} c_i \mathrm{tr}[(\boldsymbol{S}_i - \boldsymbol{\Sigma}_i)\boldsymbol{V}_i]^2,$$

where $\boldsymbol{V}_i$ is a positive definite weight matrix. Under this situation, the information matrices $\boldsymbol{J}_G(\theta)$ and $\boldsymbol{J}_{QG}(\theta)$ of the ML and GLS approaches are equal to

$$\boldsymbol{J}_G(\boldsymbol{\theta}) = 2^{-1}\boldsymbol{\Delta}\Big\{\sum_{i=1}^{m}[c_i(\boldsymbol{F}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{F}_i)\otimes(\boldsymbol{F}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{F}_i)]\Big\}\boldsymbol{\Delta}',$$

and

$$\boldsymbol{J}_{QG}(\boldsymbol{\theta}) = E(-\ddot{Q}(\boldsymbol{\theta})) = \boldsymbol{\Delta}\Big\{\sum_{i=1}^{m}[c_i(\boldsymbol{F}_i'\boldsymbol{V}_i\boldsymbol{F}_i)\otimes(\boldsymbol{F}_i'\boldsymbol{V}_i\boldsymbol{F}_i)]\Big\}\boldsymbol{\Delta}',$$

respectively, where $\boldsymbol{\Delta} = \partial\boldsymbol{\Sigma}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$. Moreover, for the GLS approach, the matrix $\boldsymbol{J}_{VG}(\boldsymbol{\theta})$ defined in Condition (C') is equal to

$$\boldsymbol{J}_{VG}(\boldsymbol{\theta}) = \boldsymbol{\Delta}\Big\{\sum_{i=1}^{m}\{c_i[(\boldsymbol{F}_i'\boldsymbol{V}_i\boldsymbol{F}_i\boldsymbol{\Sigma}\boldsymbol{F}_i'\boldsymbol{V}_i\boldsymbol{F}_i)\otimes(\boldsymbol{F}_i'\boldsymbol{V}_i\boldsymbol{F}_i\boldsymbol{\Sigma}\boldsymbol{F}_i'\boldsymbol{V}_i\boldsymbol{F}_i)]\}\Big\}\boldsymbol{\Delta}'.$$

From results developed in previous sections, the consistency and the asymptotic normality of the ML and the GLS estimators can be established if Conditions {A,B,C} and {A,B',C'} on the models, $\boldsymbol{J}_G(\boldsymbol{\theta})$ and $\{\boldsymbol{J}_{QG}(\boldsymbol{\theta}), \boldsymbol{J}_{VG}(\boldsymbol{\theta})\}$ respectively are satisfied.

Based on the above model and missing-mechanism, Lee (1986) showed that the ML and the GLS estimators are consistent and asymptotically normal under the following conditions: (i) Condition (A) as given in Section (2), (ii) $\boldsymbol{\Delta}$ is of full rank, and (iii) for all $i = 1,\ldots,m$, c_i tends to a constant $\gamma_i > 0$ as G tends to infinity. The practical implication of the third condition means that the number of observations in each missing pattern must be sufficiently large. For some situations, this may not be true. The results developed in this paper depend on (i) and (ii) but some weaker condition than (iii). For example, Conditions (B) and (C) will be satisfied if $\sum_{i=1}^{m}[c_i(\boldsymbol{F}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{F}_i)\otimes(\boldsymbol{F}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{F}_i)]$ is positive definite as G tends to infinity. Hence, the present development is more general and has more practical value.

Computationally, the ML estimate of $\boldsymbol{\theta}$ can be obtained via the following scoring algorithm: $\boldsymbol{\theta}^{(j+1)} - \boldsymbol{\theta}^{(j)} = -\rho\boldsymbol{\Gamma}(\boldsymbol{\theta}^{(j)})^{-1}\boldsymbol{v}(\boldsymbol{\theta}^{(j)})$, where ρ is the step-size parameter, $\boldsymbol{\Gamma}(\boldsymbol{\theta}) = \boldsymbol{J}_G(\boldsymbol{\theta})$ and $\boldsymbol{v}(\boldsymbol{\theta}) = -\boldsymbol{\Delta}\sum_{i=1}^{m}c_i\mathrm{Vec}[\boldsymbol{F}_i'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{S}_i - \boldsymbol{\Sigma}_i)\boldsymbol{\Sigma}_i^{-1}\boldsymbol{F}_i]$, the gradient vector of $l(\boldsymbol{\theta})$. The GLS estimate can be obtained via the Gauss-Newton algorithm in which the basic step is also defined as above but with $\boldsymbol{\Gamma}(\boldsymbol{\theta}) = \boldsymbol{J}_{QG}(\boldsymbol{\theta})$ and $\boldsymbol{v}(\boldsymbol{\theta}) = -\boldsymbol{\Delta}\sum_{i=1}^{m}c_i\mathrm{Vec}[\boldsymbol{F}_i'\boldsymbol{V}_i(\boldsymbol{S}_i-\boldsymbol{\Sigma}_i)\boldsymbol{V}_i\boldsymbol{F}_i]$, the gradient vector of $Q(\boldsymbol{\theta})$. Clearly, these algorithms also give the estimates of the asymptotic covariance matrices in terms of $\boldsymbol{J}_G^{-1}(\hat{\boldsymbol{\theta}})$ and $\boldsymbol{J}_{QG}^{-1}(\tilde{\boldsymbol{\theta}})$ at the last iteration.

4.2. Multilevel structure equation models

For simplicity, we just consider the two-level case, but the results can be extended to general multilevel models. Suppose

$$\boldsymbol{Z}_{gj} = \boldsymbol{u}_g^* + \boldsymbol{u}_{gj},\ g = 1,\ldots,G \ \text{ and } \ j = 1,\ldots,T_g, \tag{8}$$

where $\boldsymbol{u}_g^*$ are $p \times 1$ group-level random vectors which are independent and identically distributed as $N(\mathbf{0}, \boldsymbol{\Omega}_B(\boldsymbol{\theta}_0))$, and $\boldsymbol{u}_{gj}$ are $p \times 1$ individual-level random vectors which are also independent and identically distributed as $N(\mathbf{0}, \boldsymbol{\Omega}_{gw}(\boldsymbol{\theta}_0))$, and $\boldsymbol{\theta}_0$ is an unknown $q \times 1$ parameter vector. Note that $\boldsymbol{Z}_{gj}$ are not independent due to the presence of $\boldsymbol{u}_g^*$. This kind of two-level data is very common in practice. For examples, data obtained from randomly drawn students (individual-level) from randomly drawn schools (group-level), or randomly drawn officials from randomly drawn companies, etc. McDonald and Goldstein (1989) provided some analyses for a model with balanced sampling designs where T_g are all equal, and Muthén (1989) discussed the application of some package in obtaining the ML solution of their model. Lee (1990) developed the asymptotic theory for the ML and GLS estimations with unbalanced sampling designs by assuming G and T_g are sufficiently large. Under the special case where $\boldsymbol{\Omega}_{gw}(\boldsymbol{\theta}) = \boldsymbol{\Omega}_w(\boldsymbol{\theta})$ for all g, Lee and Poon (1992) derived certain asymptotic results with small individual-level sample sizes. Below, it will be shown that results obtained from Section 2 are generalizations of Lee and Poon (1992) to general models with unequal individual-level covariance structures $\boldsymbol{\Omega}_{gw}(\boldsymbol{\theta})$ and small T_g.

Let $\boldsymbol{X}_g = (\boldsymbol{Z}'_{g1}, \ldots, \boldsymbol{Z}'_{gT_g})'$ be a pT_g by 1 random vector with distribution $N(\mathbf{0}, \boldsymbol{\Sigma}_g(\boldsymbol{\theta}))$, where $\boldsymbol{\Sigma}_g(\boldsymbol{\theta}) = \boldsymbol{E}_g \otimes \boldsymbol{\Omega}_B(\boldsymbol{\theta}) + \boldsymbol{I}_g \otimes \boldsymbol{\Omega}_{gw}(\boldsymbol{\theta})$, with $\boldsymbol{E}_g$ be a T_g by T_g matrix with unit entries and $\boldsymbol{I}_g$ be an identity matrix of order T_g. Hence, $\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_G\}$ is a sample of i.n.i.d. observations. The negative log-likelihood function of this sample is proportional to

$$l_G(\boldsymbol{\theta}) = \frac{1}{2}\sum_{g=1}^{G}\Big[\log|\boldsymbol{\Sigma}_g(\boldsymbol{\theta})| + \boldsymbol{X}_g'\boldsymbol{\Sigma}_g^{-1}(\boldsymbol{\theta})\boldsymbol{X}_g\Big].$$

It can be shown (see, Lee (1990)) that the information matrix, $\boldsymbol{J}_G(\boldsymbol{\theta})$, is equal to

$$\frac{1}{2}\sum_{g=1}^{G}\Big\{(T_g-1)\boldsymbol{\Delta}_{gw}^*(\boldsymbol{\theta})[\boldsymbol{\Omega}_{gw}^{-1}(\boldsymbol{\theta})\otimes\boldsymbol{\Omega}_{gw}^{-1}(\boldsymbol{\theta})]\boldsymbol{\Delta}_{gw}^{*\prime}(\boldsymbol{\theta})+\boldsymbol{\Delta}_g^*(\boldsymbol{\theta})[\boldsymbol{\Omega}_g^{-1}(\boldsymbol{\theta})\otimes\boldsymbol{\Omega}_g^{-1}(\boldsymbol{\theta})]\boldsymbol{\Delta}_g^{*\prime}(\boldsymbol{\theta})\Big\}, \tag{9}$$

where $\boldsymbol{\Delta}_g^*(\boldsymbol{\theta}) = \partial\boldsymbol{\Omega}_g(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, $\boldsymbol{\Delta}_{gw}^*(\boldsymbol{\theta}) = \partial\boldsymbol{\Omega}_{gw}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, and $\boldsymbol{\Omega}_g(\boldsymbol{\theta}) = \boldsymbol{\Omega}_{gw}(\boldsymbol{\theta}) + T_g\boldsymbol{\Omega}_B(\boldsymbol{\theta})$. To establish the asymptotic properties as given in previous sections, similar regularity conditions are required. For example, the condition analogous to Condition (A) is given as below:

Condition (AM). (a) All elements of $\boldsymbol{\Omega}_B(\boldsymbol{\theta})$ and $\boldsymbol{\Omega}_{gw}(\boldsymbol{\theta})$ and their partial derivatives of the first three orders with respect to $\boldsymbol{\theta}$ are continuous and bounded uniformly in a neighborhood of $\boldsymbol{\theta}_0$; (b) $\boldsymbol{\theta}$ is identifiable, i.e., $\boldsymbol{\Omega}_B(\boldsymbol{\theta}_1) = \boldsymbol{\Omega}_B(\boldsymbol{\theta}_2)$

and $\boldsymbol{\Omega}_{gw}(\boldsymbol{\theta}_1) = \boldsymbol{\Omega}_{gw}(\boldsymbol{\theta}_2)$ for all g implies $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$; (c) $\boldsymbol{\Omega}_{gw}(\boldsymbol{\theta}_0)$ and $\boldsymbol{\Omega}_B(\boldsymbol{\theta}_0)$ are positive definite for all g.

Conditions (B) and (C) are defined similarly as before with $\boldsymbol{J}_G(\boldsymbol{\theta})$ given by (9). Clearly, since $\boldsymbol{\Omega}_{gw}^{-1}(\boldsymbol{\theta})$ and $\boldsymbol{\Omega}_g^{-1}(\boldsymbol{\theta})$ are positive definite, these conditions are satisfied with some full rank matrices $\boldsymbol{\Delta}_g^*(\boldsymbol{\theta})$ and $\boldsymbol{\Delta}_{gw}^*(\boldsymbol{\theta})$. Hence, the consistency and asymptotic normality of the ML estimate can be achieved for most practical applications.

Moreover, the GLS analysis of this model can be similarly established based on results developed in previous sections and Lee (1990). The scoring algorithm and the Gauss-Newton algorithm can be used respectively to obtain the ML and the GLS solutions.

5. A Simulation Study

To study the empirical behaviors of the ML and GLS estimators, a simulation study has been conducted based on the confirmatory factor analysis model (Lawley and Maxwell (1971)) with missing data that are missing completely at random. The basic model is defined as $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$, where $\boldsymbol{\Lambda}$ is the factor loading matrix, $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ are covariance matrices of the factors and error measurements, respectively. To obtain the ML and GLS estimates of these parameters, a scoring algorithm and a reweighted Gauss-Newton algorithm were implemented based on the results developed in previous sections.

Part of the data set described in Wheaton, Muthén, Alwin and Summers (1977) was used for this simulated study. The data set consists 6 variables and from previous studies (see, e.g. Jöreskog and Sörbom (1989), Lee (1986)) it is known that the covariance structure can be fitted by a confirmatory factor analysis model, in which $\boldsymbol{\Phi}$ is a 3×3 unknown covariance matrix, and

$$\boldsymbol{\Lambda} = \begin{pmatrix} 1 & 0 & 0 \\ \boldsymbol{\Lambda}_{21} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \boldsymbol{\Lambda}_{42} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \boldsymbol{\Lambda}_{63} \end{pmatrix}, \qquad \boldsymbol{\Psi} = \begin{pmatrix} \boldsymbol{\Psi}_{11} & & & \text{Sym.} & & \\ 0 & \boldsymbol{\Psi}_{22} & & & & \\ \boldsymbol{\Psi}_{31} & 0 & \boldsymbol{\Psi}_{33} & & & \\ 0 & \boldsymbol{\Psi}_{42} & 0 & \boldsymbol{\Psi}_{44} & & \\ 0 & 0 & 0 & 0 & \boldsymbol{\Psi}_{55} & \\ 0 & 0 & 0 & 0 & 0 & \boldsymbol{\Psi}_{66} \end{pmatrix},$$

where the zeros and ones are fixed values and not to be treated as unknown parameters. The total number of unknown parameters in this model is 17. Firstly, the ML estimates of these unknown parameters were obtained based on the original data, then these estimates were taken as the population values to define the population covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$. Random observations were simulated from

$N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}_0))$ via the IMSL (1991) subroutines. Various missing patterns were created by randomly deleting certain entries in the simulated observations. Data structure corresponding to the missing patterns with different T_i for consideration is given in Table 1, in which '1' represents the entry is present, while '0' represents the entry is missing. The ML and GLS estimates of the unknown parameters were obtained, and the process was repeated until 100 replications were completed. The sample means of the estimates from the 100 replications are reported in Table 2. We observed that the ML and GLS estimates are quite accurate. When comparing the means of the estimates with the true values, we found that the ML estimates are slightly better. To give some idea about the variation of the estimates, the deviations $\{\sum_i(\hat{\theta}_i(k)-\theta_0(k))^2/100\}^{1/2}$, $k=1,\ldots,q$, of the parameter estimates around the true values are presented under the column STD_0 in Table 2. The Shapiro-Wilk (1965) W-test was used to test the asymptotic normality of each estimator. The p-values associated with the W-test are also presented in Table 2. These results show that the normality of most parameter estimates is not rejected, and hence basically agree with the theory developed in previous sections. For completeness, the eigenvalues of the matrices $\boldsymbol{J}_G(\boldsymbol{\theta}_0)/G$ and $\boldsymbol{J}_G^{-1}(\boldsymbol{\theta}_0)$ are reported in Table 2 as well. We found that $\lambda_{\min}(\boldsymbol{J}_G(\boldsymbol{\theta}_0)/G) = 0.0015$. Hence, from the results in Theorems 1 to 4, the consistency and the asymptotic normality of the ML and GLS estimates can be established. Moreover, we observed that $\lambda_{\max}(\boldsymbol{J}_G^{-1}(\boldsymbol{\theta}_0)) = 2.00$; so as expected, the standard deviations of the ML and GLS estimates are approximately bounded by $2^{1/2}$.

Table 1. The patterns of missing entries in the first data set

Number	missing pattern						Number	missing pattern					
T_i	y_1	y_2	y_3	y_4	y_5	y_6	T_i	y_1	y_2	y_3	y_4	y_5	y_6
40	1	1	1	1	1	1	15	1	0	1	1	0	1
41	0	1	1	1	1	1	15	1	0	1	1	1	0
41	1	0	1	1	1	1	10	1	1	0	0	1	1
41	1	1	0	1	1	1	10	1	1	0	1	0	1
41	1	1	1	0	1	1	10	1	1	0	1	1	0
41	1	1	1	1	0	1	10	1	1	1	0	0	1
41	1	1	1	1	1	0	10	1	1	1	0	1	0
15	0	0	1	1	1	1	10	1	1	1	1	0	0
15	0	1	0	1	1	1	5	0	0	0	1	1	1
15	0	1	1	0	1	1	4	1	0	0	0	1	1
15	0	1	1	1	0	1	4	1	1	0	0	0	1
15	0	1	1	1	1	0	3	0	0	1	1	1	0
15	1	0	0	1	1	1	2	0	1	1	1	0	0
15	1	0	1	0	1	1	1	1	1	1	0	0	0

Table 2. ML and GLS analyses of the first data set

	True	MLE			GLSE			Eigenvalues	
	Value	Mean	STD_0	p-value	Mean	STD_0	p-value	$\boldsymbol{J}_G(\boldsymbol{\theta}_0)/G$	$\boldsymbol{J}_G^{-1}(\boldsymbol{\theta}_0)$
λ_{21}	0.987	.997	.099	.124	.997	.100	.138	.00100	2.00412
λ_{42}	0.935	.953	.109	.841	.952	.108	.862	.00207	.96540
λ_{63}	0.526	.527	.069	.357	.528	.071	.135	.00384	.52134
ψ_{11}	4.753	4.747	.690	.661	4.527	.691	.592	.00498	.40180
ψ_{22}	2.450	2.421	.738	.356	2.292	.721	.148	.01235	.16188
ψ_{31}	1.658	1.599	.553	.634	1.516	.544	.461	.01279	.15640
ψ_{33}	4.424	4.447	.900	.244	4.230	.867	.444	.01688	.11848
ψ_{42}	0.860	.805	.475	.246	.759	.469	.742	.02376	.08419
ψ_{44}	2.947	2.812	.784	.414	2.696	.785	.155	.03241	.06171
ψ_{55}	2.716	2.648	.839	.094	2.559	.824	.165	.03601	.05554
ψ_{66}	2.632	2.679	.300	.295	2.553	.302	.353	.04624	.04326
ϕ_{11}	7.221	7.221	.957	.644	6.834	.992	.448	.05748	.03479
ϕ_{21}	5.338	5.283	.738	.628	5.001	.786	.483	.06415	.03118
ϕ_{22}	8.273	8.132	1.259	.033	7.712	1.331	.010	.10220	.01957
ϕ_{31}	-3.954	-3.942	.500	.750	-3.720	.540	.658	1.87126	.00107
ϕ_{32}	-3.959	-3.849	.582	.163	-3.654	.625	.657	2.62989	.00076
ϕ_{33}	6.938	6.921	1.000	.122	6.553	1.017	.091	2.74207	.00073

Table 3. Eigenvalues of the information matrices

Complete data		MD2		MD3	
$\boldsymbol{J}_G(\boldsymbol{\theta}_0)/G$	$\boldsymbol{J}_G^{-1}(\boldsymbol{\theta}_0)$	$\boldsymbol{J}_G(\boldsymbol{\theta}_0)/G$	$\boldsymbol{J}_G^{-1}(\boldsymbol{\theta}_0)$	$\boldsymbol{J}_G(\boldsymbol{\theta}_0)/G$	$\boldsymbol{J}_G^{-1}(\boldsymbol{\theta}_0)$
.00152	1.31840	.00020	10.24577	.00007	26.83721
.00328	.60899	.00073	2.72297	.00028	7.17648
.00637	.31377	.00161	1.24122	.00063	3.19110
.00686	.29171	.00193	1.03436	.00117	1.70563
.01694	.11808	.00217	.92104	.00188	1.06417
.01846	.10832	.00345	.58034	.00219	.91162
.02777	.07202	.00817	.24488	.00797	.25079
.04401	.04545	.01468	.13620	.01460	.13700
.04832	.04139	.03223	.06205	.03191	.06269
.07580	.02639	.04343	.04605	.04296	.04655
.08678	.02305	.04617	.04332	.04551	.04394
.09340	.02141	.05460	.03663	.05391	.03710
.09797	.02041	.07594	.02634	.07563	.02644
.18784	.01065	.14508	.01379	.14449	.01384
2.57577	.00078	1.89764	.00105	1.85404	.00108
3.69730	.00054	2.38208	.00084	2.37501	.00084
4.33624	.00046	3.87892	.00052	3.86923	.00052

To study the impact of Condition (C) on the estimation, in addition to the complete data (CD) set and the data set (MD1) with missing patterns given in Table 1, we also included in our study two more data sets with the following three missing patterns: (1,1,1,1,1,1), (0,1,1,1,1,1) and (0,1,1,1,1,0). The sample sizes for each missing pattern in the new data sets are given by MD2: $T_1 = 20$, $T_2 = 240$, $T_3 = 240$; and MD3: $T_1 = 8$, $T_2 = 246$, $T_3 = 246$. The eigenvalues of the matrices $\boldsymbol{J}_G(\boldsymbol{\theta}_0)/G$ and $\boldsymbol{J}_G^{-1}(\boldsymbol{\theta}_0)$ were computed and reported in Table 3. From this table, we observe that the minimum eigenvalues of the matrix $\boldsymbol{J}_G(\boldsymbol{\theta}_0)/G$ corresponding to the complete data set and MD1 are quite close and large. Because more entries are missing in data sets MD2 and MD3, the corresponding minimum eigenvalues are smaller. The behaviors of the eigenvalues for $\boldsymbol{J}_G^{-1}(\boldsymbol{\theta}_0)$ are similar but in the opposite direction. Therefore, according to the results given in Theorems 1 and 2, the ML estimates corresponding to complete data and data set MD1 should be consistent. The ML estimates corresponding to data sets MD2 and MD3 may exist but their standard deviations may be quite large. Moreover, the minimization process in getting the ML estimates for MD2 and MD3 was rather unstable. Similar phenomena are observed in the GLS estimation.

6. Discussion

In this paper, the ML and GLS analyses for covariance structures with i.n.i.d. observations are discussed. Conditions for obtaining the consistency and asymptotic normality of the estimators are identified. In the analyses, it is assumed that the mean vectors of the random observations are zero vectors. For most practical applications in the field, this assumption is satisfied. Moreover, it should be noted that even for the more general case with $\boldsymbol{X}_g$ distributed as $N[\boldsymbol{\mu}_g(\boldsymbol{\beta}_0), \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)]$, where $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$ are distinct parameter vectors, analogous ML and GLS theory can be developed based on an approach similar to that presented in this paper.

Analyses of two special models that have important applications are presented to demonstrate the applicability of the general results. These are the multilevel models and models with certain data missing completely at random. It is shown that the conditions for obtaining the asymptotic results can be verified easily in these cases. Moreover, it is expected that the general results developed can be utilized to analyze other models with non-identically distributed observations.

Acknowledgement

The authors would like to thank the editors and anonymous referees for their helpful comments and suggestions. This research was supported in part by a HK-UGC grant.

References

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modeling. *Annual Review of Psychology* **31**, 419-456.

Bentler, P. M. (1983). Some contributions to efficient statistics in structural models: Specification and estimation of moment structures. *Psychometrika* **48**, 493-517.

Browne, M. W. (1982). Covariance structures. In *Topics in Applied Multivariate Analysis* (Edited by D. M. Hawkins), 72-141, Cambridge: Cambridge University Press.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British J. Math. Statist. Psych.* **37**, 62-83.

Eaton, M. L. (1983). *Multivariate Statistics*. John Wiley.

Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Annals Statist.* **13**, 342-368.

Hoadley, B. (1971). Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *Ann. Math. Statist.* **42**, 1977-1991.

IMSL (1991). *FORTRAN Subroutines for Statistical Analysis*. IMSL Inc.

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. *Biometrika* **57**, 239-251.

Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika* **43**, 443-477.

Jöreskog, K. G. and Sörbom, D. (1989). *LISREL 7: A Guide to the Program and Applications*. (2nd Edition). Chicago, IL: SPSS.

Lawley, D. N. and Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method* (2nd ed). American Elsevier, New York.

Lee, S. Y. (1986). Estimation for structural equation models with missing data. *Psychometrika* **51**, 93-99.

Lee, S. Y. (1990). Multilevel analysis of structural equation models. *Biometrika* **77**, 763-772.

Lee, S. Y. and Jennrich, R. I. (1979). A study of algorithms for covariance structure analysis with specific comparisons using factor analysis. *Psychometrika* **44**, 99-113.

Lee, S. Y. and Poon, W. Y. (1992). Two-level analysis of covariance structures for unbalanced designs with small level-one samples. *British J. Math. Statist. Psych.* **45**, 109-123.

McDonald, R. P. and Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British J. Math. Statist. Psych.* **42**, 215-232.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika* **54**, 557-585.

Nordberg, L. (1980). Asymptotic normality of maximum likelihood estimators based on independent, unequally distributed observations in exponential family models. *Scand. J. Statist.* **7**, 27-32.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. (2nd ed.). John Wiley.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality. *Biometrika* **52**, 591-612.

Wheaton, B., Muthén, B. O., Alwin, D. and Summers, G. (1977). Assessing reliability and stability in panel models. In *Sociological Methodology* (Edited by D. R. Heise), 84-136. Jossey-Bass, San Francisco.

Wu, C. F. (1981). Asymptotic theory of nonlinear least squares estimation. *Ann. Statist.* **9**, 501-513.

Department of Statistics, The Chinese University of Hong Kong Shatin, N.T., Hong Kong.

E-mail: jgshi@cuhk.edu.hk; B003784@idea.csc.cuhk.edu.hk

(Received May 1995; accepted May 1997)