COVARIANCE STRUCTURE ANALYSIS IN SEVERAL POPULATIONS

Sik-Yum Lee
THE CHINESE UNIVERSITY OF HONG KONG

Kwok-Leung Tsui
UNIVERSITY OF WISCONSIN-MADISON

This paper is concerned with the study of covariance structural models in several populations. Estimation theory of the parameters that are subject to general functional restraints is developed based on the generalized least squares approach. Asymptotic properties of the constrained estimator are studied; and asymptotic chi-square tests are presented to evaluate appropriate model comparisons. The method of multipliers and the standard reparametrization technique are discussed in obtaining the estimates. The methodology is demonstrated by a set of real data.

Key words: generalized least squares, asymptotic distributions, goodness-of-fit test, multiplier method, reparametrization, idempotent matrix.

1. Introduction

This paper is concerned with the study of covariance structure analysis in several populations. A significant initial contribution towards the several groups problem was made by Jöreskog [1971]. He considered the confirmatory factor analysis model with three kinds of parameters (i) fixed parameters that have been assigned given values, (ii) constrained parameters that are unknown but equal to one or more other parameters, and (iii) free parameters. The model was estimated by the maximum likelihood method, and various null hypotheses about the structure of the model were tested by a large sample chi-square test. Sörbom [1974] generalized these ideas to models which also allow one to estimate the population means. Their work was implemented into computer programs which are widely distributed under the names of COFAMM and LISREL. These programs are extremely useful in analyzing covariance structures. However, the theoretical basis for multiple population structural analysis is as yet incomplete. For example, the above cited work has not discussed the asymptotic distribution of the constrained estimates, the formulas used to compute standard error estimates are not presented, and no proofs have been given on the distribution of the goodness-of-fit $\chi^2$ statistic. In this paper, we extend the work of Jöreskog [1971] in two aspects: (1) we consider a general covariance structure model instead of the factor analytic model, and (2) we allow more complex constraints on parameters than (i) and (ii). Moreover, basic asymptotic properties of the estimation theory will be stated and proved.

We are considering a set of $m$ populations. For $g = 1, \ldots, m$, it is assumed that the distribution of the $g$th population is multivariate normal with covariance matrix $\Sigma_g^* = \Sigma_g(\beta_g^*)$ whose elements are differentiable functions of the unknown $\theta_g$ by 1 parameter vector $\beta_g^*$. In general, we may have different covariance structures and hence the density functions for different populations may be distinct. Let $\theta^* = (\beta_1^*, \ldots, \beta_m^*)'$ be the overall parameter
vector of order \( q = t_1 + \cdots + t_m \). We suppose that \( \Theta^* \) satisfies \( r \leq q \) relationships: \( h(\Theta^*) = (h_1(\Theta^*), \ldots, h_r(\Theta^*)) = 0 \), where \( h_1, \ldots, h_r \) are independent differentiable functions. Thus, the restrictions imposed on \( \Theta^* \) are very general, and they can be across populations. The constrained estimate of \( \Theta^* \) will be obtained based on \( S_g, g = 1, \ldots, m \), the sample covariance matrix based on a random sample of size \( (n_g + 1) \). The method of generalized least squares estimation will be studied, and its relation to the maximum likelihood approach will be discussed. Asymptotic properties of the constrained generalized least squares estimates and the goodness-of-fit statistics will be stated and proved. Expressions for computing the standard error estimates are obtained as a byproduct. To the best of our knowledge, these asymptotic properties in the several populations' context have not been given in the literature.

Two methods for obtaining the constrained generalized least squares estimates will be discussed. For simple constraints, e.g., linear constraints, the estimates are obtained via the reparameterization technique which is similar to that given in McDonald [1980] in a single sample case. The method of multipliers (see, e.g., Bertsekas, 1976; Lee, 1981) is used to cope with more complicated constraints that cannot be handled easily by reparameterization. To implement the theory for both methods, computer programs have been written in the context of the confirmatory factor analytic model, that, by the theory of Lee and Jennrich [1979], also provide maximum likelihood estimates. These programs yield estimates and can handle complex constraints that COFAMM and LISREL cannot handle. It would seem that no other packaged programs can serve these functions.

The approach to be employed in developing the statistical and computational theory for multiple population covariance structures is given here as an extension of the work of Browne [1974] and Lee and Bentler [1980], that is, the distribution of sample covariance matrices serves as the basis for theoretical and computational development. Because of the psychometric tradition of working with covariance structure models, it would seem that this is a logical approach. However, we make no attempt to claim that the present approach is unique, or even the best approach, although it is based on standard multivariate theory. Other approaches, such as those based on nonlinear regression, might yield similar or identical results.

It might seem that the multiple population problem is only a slight variant of the single population case. Indeed, some reviewers have suggested that an approach based on the conglomeration of data into a single larger population would allow past theory (e.g., Aitchison & Silvey, 1958; Lee & Bentler, 1980) to be directly applied to the problem so as to make it quite trivial. Our point of view is different. We believe that the multiple sample case contains many subtleties that may lead to traps and errors unless systematic proofs are obtained. For example, when the samples are of equal size, say \( n + 1 \), McDonald [1980] pointed out that his one-population computer program COSAN can be applied to give the maximum likelihood or generalized least squares estimates for the several samples problem. Essentially, he proposed to fit \( \Sigma^0 \) to \( S^0 \), where

\[
S^0 = \begin{bmatrix} S_1 & \cdots & 0 \\ 0 & \cdots & S_m \end{bmatrix}, \quad \Sigma^0 = \begin{bmatrix} \Sigma_1 & \cdots & 0 \\ 0 & \cdots & \Sigma_m \end{bmatrix}.
\]

Computationally, this method can give maximum likelihood estimates. But the asymptotic theory for a single population cannot be directly applied, because \( nS^0 \) does not have a Wishart distribution. Therefore, even in this special situation, the properties of the multiple samples case are still necessary to making statistical inference.

Lee and Bentler [1980] provided some asymptotic properties for constrained generalized least squares estimation in covariance structure models in the context of a single