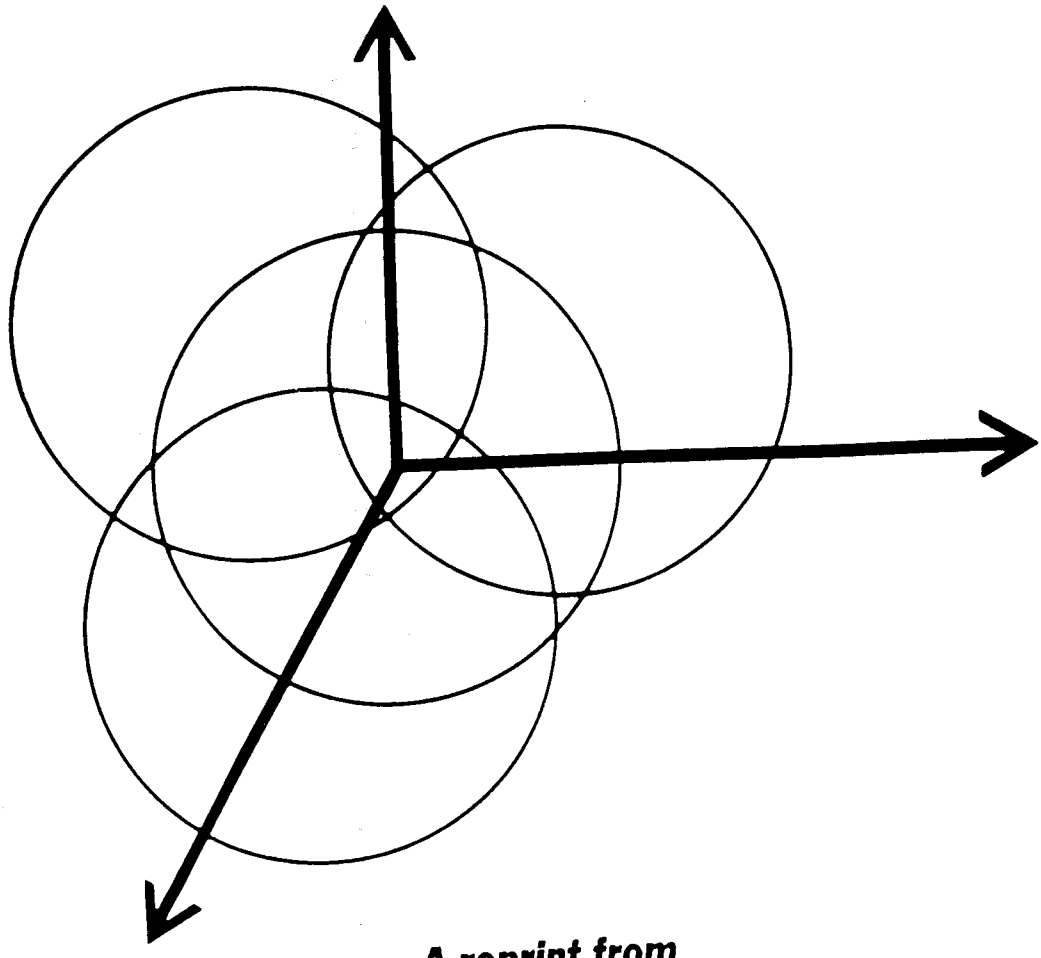

A reprint from

communications in statistics

COMMUN. STATIST.-THEORY METH., 20(4), 1153-1161 (1991)

REGRESSION COMPONENT DECOMPOSITION IN STRUCTURAL ANALYSIS

Klaus Haagen
Istituto di Statistica
Università di Trento
Via Rosmini 42, 38100 Trento, Italy

Giorgio Vittadini
Istituto di Statistica
Università di Brescia
Corso Mameli 27, 25122 Brescia, Italy

Keywords and Phrases: structural model, factor analysis, Lisrel model, identification, regression component decomposition

ABSTRACT

A crucial problem in the Lisrel model remains that of finding necessary and sufficient conditions for the identification of the parameters. But even if the parameters in a particular model are identifiable, there remains an indeterminacy of the scores of the latent variables. To avoid this problem an alternative approach to the Lisrel model is here proposed, one that is based on a decomposition of the datamatrix in such a way that the assumptions in the Lisrel model are satisfied.

1. INTRODUCTION

It is well known that there are no necessary and sufficient conditions for identification of the parameters in structural equation models with latent variables as in the Lisrel model. In this case the estimation of the parameters makes no sense, since the identification of parameters and their estimation are dual aspects of the same problem. Even for particular cases in which the parameters are identified there remains an indeterminacy problem if one is interested in the estimation of the scores of the latent variables. (Elffers, Bethlehem and Gill, 1978, Haagen, 1987,1990, Schönemann and Haagen, 1987, Vittadini, 1988,1989). As Elffers et.al.(1978) observed for the case of the factor model, the solution of the indeterminacy problem is essential to attach a real meaning to the mathematically possible factors. The

same problem arises in the Lisrel model with unknown latent variables. To avoid this arbitrariness in interpreting the latent variables in the Lisrel model, the data matrix is decomposed in components that have the properties of the variables in the Lisrel model. For the factor model Schönemann and Steiger (1976) proposed a decomposition of the datamatrix in two components: the common factors are defined as linear combinations of the observable variables and the factor loadings as the regression coefficients, regressing the observed variables on the so defined common factors (regression components). In our case, however, the problem is more complex. Writing the Lisrel model in the form of a common factor model, the factorloading matrix has quite a different structure than those in the common factor model. The scores of the latent factors are defined as regression components with the same structure as the latent variables in the Lisrel model.

2. THE LISREL MODEL

The Lisrel model is composed of one structural and two measurement equations

$$\begin{aligned} \eta_{(t)} &= B'\eta_{(t)} + \Gamma'\xi_{(t)} + \varepsilon_{(t)} \\ y_{(t)} &= \Lambda_Y'\eta_{(t)} + \delta_{(t)} \\ x_{(t)} &= \Lambda_X'\xi_{(t)} + \nu_{(t)} \qquad t = 1,\ldots,T \end{aligned} \tag{1}$$

where the vectors $\eta, \xi, y,$ and x have m, k, q and p components, with $p>k$, $q>m$, $T>p+q$.

It is assumed that all the random variables have zero mean and finite variance. B' is a triangular matrix with zero on the main diagonal, and

$(\xi_{(t)}', \nu_{(t)}', \delta_{(t)}', \varepsilon_{(t)}')$, $t = 1,\ldots,T$ are
identically and independently distributed
(V', Δ, E) and Ξ are independent
(V', Δ) and E are independent
V' and Δ are independent

where
$E' = (\varepsilon_{(1)}, \ldots, \varepsilon_{(T)})$, $V' = (\nu_{(1)}, \ldots, \nu_{(T)})$ $\Delta' = (\delta_{(1)}, \ldots, \delta_{(T)})$ $\Xi' = (\xi_{(1)}, \ldots, \xi_{(T)})$

In order to show the relatinship between the Lisrel model and the common factor model we write (1) in the form

$$z = A_{z\mu}\mu + \tau \tag{2}$$

with

$$z=\begin{pmatrix} y \\ x \end{pmatrix}, \qquad \mu=\begin{pmatrix} \xi \\ \varepsilon \end{pmatrix}, \qquad \tau=\begin{pmatrix} \delta \\ \nu \end{pmatrix} \quad \text{and}$$

$$A_{z\mu}=\begin{pmatrix} A_{y\xi} & A_{y\varepsilon} \\ A_{x\xi} & A_{x\varepsilon} \end{pmatrix}=\begin{pmatrix} \Lambda'_y(I-B')^{-1}\Gamma' & \Lambda'_y(I-B')^{-1} \\ \Lambda'_x & 0 \end{pmatrix} \tag{3}$$

where $A_{z\mu}$ is the regression coefficient matrix, regressing z on μ.

For the variance-covariance matrix we have

$$\Sigma_{zz}=A_{z\mu}\Sigma_{\mu\mu}A'_{z\mu}+\Sigma_{\tau\tau} \tag{4}$$

where

$$\Sigma_{\mu\mu}=\begin{pmatrix} \Sigma_{\xi\xi} & 0 \\ 0 & \Sigma_{\varepsilon\varepsilon} \end{pmatrix}, \qquad (\Sigma_{\mu\tau}=0) \tag{5}$$

3. THE INDETERMINACY OF THE FACTOR SCORES

The indeterminacy of the scores of the latent variables is based on the following well known lemma (Kano, 1984).

Lemma

Assume that the random vector z satisfies (4); then there exists a random vector w such that

$$\mu=A'_{z\mu}\Sigma_{zz}^{-1}z+w, \qquad \Sigma_{zw}=0$$

and

$$\Sigma_{ww}=\Sigma_{\mu\mu}-A'_{z\mu}\Sigma_{zz}^{-1}A_{z\mu}. \tag{6}$$

$A'_{z\mu}\Sigma_{zz}^{-1}z$, which is usually used as an "estimator" for the latent factors (Haagen, 1986), is called the regression part and w the arbitrary part of z. Kano (1984) calls the covariance matrix of w arbitrariness. The arbitrariness cannot be eliminated for a finite number of variables (Kano, 1984; Haagen, 1990), consequently this can make the prediction of the latent factors meaningless (Schönemann and Haagen, 1987). Williams (1978) proposed a redefinition of the common factor model, in which the limit of the sequence of arbitrariness vanishes for infinite observational variables. However, in the empirical research, where only a limited number of variables is possible, the arbitrariness remains a crucial problem in interpreting the latent factors.

Therefore, we propose a less ambiguous method to analize data structures, decomposing the observable data matrix into components which have analogous properties such as the latent variables in the Lisrel model, but we do not claim that they must be "causal" factors like those in the Lisrel model.

4. DECOMPOSITION OF THE DATA MATRIX

Let

$$H' = (\eta_{(1)}, \ldots, \eta_{(T)}); \quad X' = (x_{(1)}, \ldots, x_{(T)}); \quad Y' = (y_{(1)}, \ldots, y_{(T)})$$

we have for (1)

$$H = HB + \Xi\Gamma + E \tag{7}$$

$$Y = H\Lambda_y + \Delta \tag{8}$$

$$X = \Xi\Lambda_x + V \tag{9}$$

We assume now that X, Y, E, Ξ, V and Δ are mean centered data matrices, where only X and Y are observed. Given X and Y, we construct components H, Ξ, Δ and V and coefficients B, Γ, Λ_x and Λ_y such that the these components have analogous properties like the variables in the Lisrel model.

Let $S(\Xi)$ be the vector space generated by the columns of Ξ.

$$P_\Xi = \Xi(\Xi'\Xi)^{-1}\Xi'$$

is the orthorgonal projector on $S(\Xi)$

$$Q_\Xi = I - P_\Xi$$

is the projector on the orthorgonal complement and

$$P_{X/\Xi} = Q_\Xi X(X'Q_\Xi X)^{-1}X'Q_\Xi \tag{10}$$

is the orthogonal projector on the space generated by those columns of X, which are elements of the orthogonal complement of the space spanned by the columns of Ξ.

Let $P_{\Xi \cup X}$ be the orthogonal projector onto the linear sum of $S(\Xi)$ and $S(X)$

than we obtain (Rao & Yanai, 1979)

$$P_{X/\Xi} = P_\Xi + P_{X/\Xi}$$

With this notation we have the following decomposition:

$$S(\Xi, X, E, Y) = S(\Xi) \oplus S(Q_\Xi X) \oplus S(Q_{\Xi \cup X} E) \oplus S(Q_{\Xi \cup X \cup E} Y), \tag{11}$$

for Z = (Y,X) we have

$$Z = P_\Xi Z + P_{X/\Xi} Z + P_{E/\Xi \cup X} Z + P_{Y/\Xi \cup X \cup E} Z, \tag{12}$$

To distinguish the components which result from the decomposition of the parameters and the variables in the stochastic model we use the symbol " ~ ".

Using $\tilde{A}_{XY}$ to indicate the regression pattern, regressing Y on X, we can rewrite equation (12) in the form

$$Z = \tilde{\Xi}\tilde{A}_{\Xi Z} + \tilde{E}^{*}\tilde{A}_{E^{*}Z^{*}} + (\tilde{\Delta}, \tilde{V}) \tag{13}$$

where

$$\tilde{E}^{*} = Q_{\Xi \cup X}\tilde{E}, \quad Z^{*} = Q_{\Xi \cup X}Z \tag{14}$$

With

$$\tilde{A}_{\Xi Z} = (\tilde{A}_{\Xi Y}, \tilde{A}_{\Xi X}) \quad \text{and} \quad \tilde{A}_{E^{*}Z^{*}} = \tilde{A}_{E^{*}Z} = (\tilde{A}_{E^{*}Y}, 0) \tag{15}$$

(note that $Q_{\Xi \cup X}X = 0$)

we obtain

$$(Y, X) = \tilde{\Xi}(\tilde{A}_{\Xi Y}, \tilde{A}_{\Xi X}) + \tilde{E}^{*}(\tilde{A}_{E^{*}Y}, 0) + (\tilde{\Delta}, \tilde{V}) \tag{16}$$

5. DETERMINATION OF THE WEIGHTS AND THE FACTOR SCORES

First we calculate the component matrix $\tilde{\Xi}$. From (11) and (15), it follows that we must find a basis of the subspace $S(\tilde{\Xi})$ such that

$$P_{\tilde{\Xi}}Z = \tilde{\Xi}(\tilde{A}_{\Xi Y}, \tilde{A}_{\Xi X}) \tag{17}$$

We therefore define $\tilde{\Xi}$ as a linear combination of the columns of Z $\tilde{\Xi} = ZL'$ where the covariance matrix $\Sigma_{\tilde{\Xi}\tilde{\Xi}} = L\Sigma_{ZZ}L'$ is positive definite and Z is decomposed by

$$Z = \tilde{\Xi}\tilde{A}'_{\Xi Z} + (Z - \tilde{\Xi}\tilde{A}'_{\Xi Z}) \tag{18}$$

where

$$\tilde{A}'_{\Xi Z} = \Sigma^{-1}_{\tilde{\Xi}\tilde{\Xi}}\Sigma_{\Xi Z} = \Sigma^{-1}_{\tilde{\Xi}\tilde{\Xi}}L\Sigma_{ZZ}.$$

$$\tilde{A}'_{\Xi Z}\Sigma^{-1}_{ZZ}\tilde{A}_{\Xi Z} = \Sigma^{-1}_{\tilde{\Xi}\tilde{\Xi}}(L\Sigma_{ZZ}L')\Sigma^{-1'}_{\tilde{\Xi}\tilde{\Xi}} = \Sigma_{\tilde{\Xi}\tilde{\Xi}}$$

We have

$$L = (\tilde{A}'_{\Xi Z}\Sigma^{-1}_{ZZ}\tilde{A}_{\Xi Z})^{-1}A'\Sigma^{-1}_{ZZ}, \quad \forall \tilde{A}_{\Xi Z} \tag{19}$$

and therefore

$$\tilde{\Xi} = Z\Sigma_{ZZ}^{-1}\tilde{A}_{\Xi Z}(\tilde{A}'_{\Xi Z}\Sigma_{ZZ}^{-1}\tilde{A}_{\Xi Z})^{-1} \tag{20}$$

The decomposition of Z given in (18) is called Regression Component Decomposition (RCD) (Schönemann and Steiger, 1976). In the case of the common factor model where we must estimate the factor loadings and the common factor scores, given a data matrix Z, Schönemann and Steiger propose the decomposition (18) as an alternative to the estimation of the factor loadings and the factor scores in order to eliminate the indeterminacy of the common factor scores. The components thus defined have the same properties (except the full rank condition of the covariance matrix of the specific factors) as the factors in the common factor model.

From (18-20) it follows that we can take the factor loading matrix, determined by a factor extraction method, in order to get L and $\tilde{\Xi}$. Given $\tilde{\Xi}$, X and Y we obtain

$$\tilde{A}'_{\Xi X} = (\tilde{\Xi}'\tilde{\Xi})^{-1}(\tilde{\Xi}'X) = \tilde{\Lambda}_X \tag{21}$$

and

$$\tilde{A}'_{\Xi Y} = (\tilde{\Xi}'\Xi)^{-1}(\tilde{\Xi}'Y) = \tilde{\Gamma}(I - \tilde{B})^{-1}\tilde{\Lambda}_Y \tag{22}$$

From (12), (13) we have that $\tilde{\Gamma}$ is given by

$$\tilde{\Gamma} = P_{X/\Xi}Z = Q_{\Xi}Z = Q_{\Xi}X(X'Q_{\Xi}X)^{-1}X'Q_{\Xi}Z^* \tag{23}$$

To get E* we note that the columns $\tilde{E}^*$ must be orthogonal to the columns of $\tilde{\Xi}$, $\tilde{\Delta}$ and $\tilde{\Gamma}$.

From (12), defining

$$Z^{**} = Z - P_{\Xi}Z - P_{X/\Xi}Z \tag{24}$$

we have

$$Z^{**} = E^*(E^{*\prime}E^*)^{-1}E^{*\prime}Z + P_{Y/E^*,X,\Xi} \tag{25}$$

where $E^* = Q_{X\cup\Xi}E$

Applying RCD on the Z** we get the linear combination $E^* = Z^{**}L'_{Z^{**}}$, with

$$L_{Z^{**}} = (\tilde{A}'_{E^*Z^{**}}\Sigma_{Z^{**}Z^{**}}^{-1}\tilde{A}_{E^*Z^{**}})^{-1}\tilde{A}'_{E^*Z^{**}}\Sigma_{Z^{**}Z^{**}}^{-1} \tag{26}$$

Up to now we have calculated $\tilde{\Xi}, \tilde{A}_{\Xi X}, \tilde{A}_{\Xi Y}, \tilde{V}, \tilde{E}^*, \tilde{A}_{E^*Z^{**}}$ (note that $\tilde{A}_{E^*Z^{**}} = \tilde{A}_{E^*Z}$).

From the orthogonality of the columns of E* and X we have

$$\tilde{A}_{E^*X} = 0 \tag{27}$$

Finally we calculate $\tilde{\Delta}$, using

$$\tilde{\Delta} = P_{Y/\Xi \cup X \cup E^*} Z \tag{28}$$

From

$$\tilde{A}'_{\Xi Y} = \tilde{\Gamma}(I - \tilde{B})^{-1}\tilde{\Lambda}_Y \tag{29}$$

$$\tilde{A}'_{E^*Y} = (I - \tilde{B})^{-1}\tilde{\Lambda}_Y \tag{30}$$

we have

$$\tilde{A}'_{\Xi Y} = \tilde{\Gamma}\tilde{A}'_{E^*Y} \tag{31}$$

or

$$\tilde{\Gamma} = \tilde{A}'_{\Xi Y}(\tilde{A}_{E^*Y}\tilde{A}'_{E^*Y})^{-1} \tag{32}$$

$\tilde{\Lambda}_x$ is given by

$$\tilde{A}'_{E^*X} = \tilde{\Lambda}_x. \tag{33}$$

From

$$\tilde{H} = \tilde{H}\tilde{B} + \tilde{\Xi}\tilde{\Gamma} + \tilde{E}$$

we obtain

$$(I - \tilde{B})'\tilde{H}'H(I - \tilde{B}) = (\tilde{\Gamma}'\tilde{\Xi} + \tilde{E}')(\tilde{\Xi}\tilde{\Gamma} + \tilde{E}) = :S \tag{34}$$

Let $\tilde{H}'\tilde{H} = I$

then we obtaint $(I - \tilde{B})$ by a Colesky Factorisation of S.

From (30) we therefore have

$$\tilde{\Lambda}_y = (I - \tilde{B})\tilde{A}'_{E^*Y} \tag{35}$$

6. THE EQUIVALENCE TO THE LISREL MODEL

From the above derivation it follows that the given decomposition leads to components, that is to "latent variables" which have the same properties postulated by the assumptions in the stochastic Lisrel model.
But there is a main difference: the indeterminacy problem in the Lisrel model is due to the fact that the partial covariance matrix $\Sigma_{\mu\cdot z}$ does not vanish. For the proposed decomposition, however, we have the following result:

$$\tilde{\Xi}'\tilde{\Xi} - (\tilde{\Xi}'\tilde{\Xi})\tilde{A}_{\Xi Z}(Z'Z)^{-1}\tilde{A}'_{\Xi Z}(\tilde{\Xi}'\tilde{\Xi}) = 0. \tag{36}$$

7. SUMMARY

Because of the arbitrariness of the scores of the latent variables in the Lisrel model, the commonly used least squares "estimation" of these scores is not valid. This "estimation" implies a definition of the latent variables. In other words, the definition of the undetermined latent variables depends on the estimation criteria.

To resolve the indeterminacy of the latent scores, a decomposition of the observed variables according to the particular structure between the observed and the latent variables is proposed, where these "latent" variables (components) are defined as linear combinations of the the observed variables. The difference between these components and the latent variables in the Lisrel model is analogous to the difference between the principal components in a component analysis and the common factors in the factor analysis model.

BIBLIOGRAPHY

Elffers, H., Bethlehem, J.,Gill, R.(1978). Indeterminacy problems and the interpretation of factor analysis results. *Statistica Neerlandica*, 32, 181-199.

Haagen, K. (1983). Il problema dell'identificazione nell'analisi fattoriale e le consequenze nelle applicazioni empiriche. *Quaderni di Statistica Matematica applicata alle scienze economiche-sociali*, Trento, 6, 21-28.

Haagen, K. (1986). The indeterminacy problem in factor analysis and its consequences for prediction. *Österreichische Zeitschrift für Statistik und Informatik*, 15, 197-205.

Haagen, K. (1990). On necessary and sufficient conditions for vanishing the arbitrariness in a common factor model. *Submitted for publication.*

Kano, Y. (1984). Construction of additional variables conforming to a common factor model. *Statistics & Probability Letters*, 2, 241-244.

Jöreskog, K.G. (1982). The Lisrel V. Uppsala University Press.

Schönemann, P.H., Steiger, J.H. (1976). Regression component analysis. *British Journal of Mathematical and Statistical Psychology*, 29, 175-189.

Schönemann, P.H., Haagen, K. (1987), On the use of factor scores for prediction. *Biometrical Journal*, 29, 835-847.

Rao C.R., Yanai,H., (1979), General definition and decomposition of projectors and some applications to statistical problems. *Journal of Statistical Planing and Inference*, 3, 1-17.

Vittadini, G. (1988), On the validity of the indeterminate latent variables in the Lisrel model. *Communications in Statistics*, 17, 3, 861-874.

Vittadini, G. (1989), Indeterminacy problem in the Lisrel model. To be published in *Multivariate Behavioral Research*, 3.

Williams, J.S. (1978), A definition for the common-factor analysis model and the elimination of problems of factor score indeterminacy, *Psychometrika*, 43, 293-306.

Received August 1990; Revised January 1991.

Recommended by A. M. Kshirsagar, University of Michigan, Ann Arbor, MI.