Improper Solutions in Exploratory Factor Analysis: Causes and Treatments

Yutaka Kano
Faculty of Human Sciences, Osaka University
Suita, Osaka 565, Japan. email: kano@hus.osaka-u.ac.jp

Abstract: There are many causes of occurrence of improper solutions in factor analysis. Identifying potential causes of the improper solutions gives very useful information on suitability of the model considered for a data set.

This paper studies possible causes of improper solutions in exploratory factor analysis, focusing upon (A) sampling fluctuations, (B) model underidentifiable and (C) model unfitted, each having several more detailed items. We then give a checklist to identify the cause of the improper solution obtained and suggest a method of reanalysis of the data set for each cause.

Keywords: Covariance Structure Analysis, Checklist, Underidentifiability, Sample Fluctuations.

1. Factor Analysis Model and Improper Solution

In factor analysis, an observed random p-vector $\boldsymbol{x}$ is assumed to have the following form: $\boldsymbol{x} = \Lambda\boldsymbol{f} + \boldsymbol{u}$, where $\Lambda = (\lambda_{ij})$ is a $p \times k$ matrix of factor loadings, $\boldsymbol{f} = [F_1, \cdots, F_k]'$ is a k-vector of common factors, $\boldsymbol{u} = [U_1, \cdots, U_p]'$ being a p-vector of unique factors. Here k is the number of factors. Assume further that $\mathrm{Var}(\boldsymbol{f}) = I_k$, $\mathrm{Cov}(\boldsymbol{f}, \boldsymbol{u}) = 0$, $\mathrm{Var}(\boldsymbol{u}) = \Psi = \mathrm{diag}(\psi_1, \cdots, \psi_p)$. The covariance matrix Σ $(= (\sigma_{ij}))$ of the observed vector $\boldsymbol{x}$ is representable as $\Sigma = \Lambda\Lambda' + \Psi$. Each diagonal element ψ_i of Ψ is a variance of U_i, so that it should be estimated as a positive value. It is said to be an improper solution or a Heywood case when some elements ψ_i are not positively estimated.

2. Cause, Identification and Treatment

Following van Driel (1978), we distinguish among three types of causes of improper solutions as in Table 1: (A) sampling fluctuations, (B) model underidentifiable, and (C) model unfitted. We shall make brief comments on these causes. Since the parameter space of ψ_i is the finite interval $(0, \sigma_{ii})$ and estimation methods naturally do not require estimates $\hat{\psi}_i$ to be in the interval, $\hat{\psi}_i$ can be outside the parameter space or can be at its boundary at a positive probability, because of sampling variations. A typical example is an improper solution which takes place in a simulation study under a true identified model. Ihara and Okamoto (1985) compared estimation methods such as ML and LS in terms of frequency of improper solutions due to sampling fluctuations. Anderson and Gerbing (1984), Boomsma (1985)

- A: sampling fluctuation
- B: underidentifiability
  - B1: $\Lambda = \begin{bmatrix} & \lambda_{1k} \\ \Lambda_1 & 0 \\ & \vdots \\ & 0 \end{bmatrix}$ (only <u>one</u> nonzero element in a column of Λ)
  - B2: $\Lambda = \begin{bmatrix} & \lambda_{1k} \\ & \lambda_{2k} \\ \Lambda_1 & 0 \\ & \vdots \\ & 0 \end{bmatrix}$ (only <u>two</u> nonzero elements in a column of Λ)
  - B3: others
- C: factor model unfitted
  - C1: some true unique variances $\psi_i < 0$
  - C2: inconsistent variables X_i included
  - C3: others (e.g., minor factors)
- D: others (e.g., outliers)

Table 1: *Types of causes of improper solutions*

and Gerbing and Anderson (1985) have studied how model characteristics influence on the frequency of occurrence of improper solutions in the context of confirmatory factor analysis. In my experience, however, improper solutions due to sampling fluctuations are not so often met in practice. When sampling fluctuations cause an improper solution, it may be useful to constrain uniqueness estimates $\hat{\psi}_i$ to be nonnegative. Gerbing and Anderson (1987) discussed interpretability of constrained estimates for improper solutions caused by sampling fluctuations in confirmatory factor analysis.

The causes (B) and (C) are important in model inspection. Anderson and Rubin (1956) gave a necessary condition for identification that there be at least three nonzero elements in each column in Λ. The cases (B1) and (B2) in Table 1 violate the necessary condition.

In (B1), the parameters (λ_{1k}, ψ_1) can take any values as far as they meet $\lambda_{1k}^2 + \psi_1 = \sigma_{11} - \sum_{r=1}^{k-1} \lambda_{1r}^2$, so that ψ_1 can be negative. The location of the nonzero loading in the k-th column is arbitrary. The k-th common factor is not a common factor but a unique factor. Thus, it is actually a $(k-1)$-factor model. Accordingly, an improper solution takes place when the number of factors is overestimated. Why is it overestimated? In many cases, a test of goodness of fit suggests it, that is, the test rejects a $(k-1)$-factor model. There would be two potential cases yielding this anomaly ($(k-1)$-factor model is rejected; examination of a k-factor solution suggests a $(k-1)$-factor model):

(B11) Sample size n is large enough to reject reasonably well-fitted models;

in other words, because of large samples, the statistical test becomes too sensitive to small deviation from the model, possibly caused by minor factors;

(B12) Distribution assumptions such as normality are violated, and then the distribution of test statistics can not be approximated by a chi-square distribution.

The both cases have been pointed out in the context of covariance structure analysis. For (B11), researchers are advised to use goodness of fit indices such as GFI, CFI and RMSEA to measure the distance of the population from the model (e.g., Jöreskog and Sörbom 1993 section 4.5.2; Bentler 1995 chapter 5). They would then get useful information concerning acceptability of the model. The cause (B11) is also interpreted as an effect of minor factors. In (B12), researchers can take elliptical theory or a type of asymptotically distribution-free (ADF) method. Kano (1990) suggests a noniterative estimation procedure which prevents a unique factor from being reinterpreted as a common factor.

Checking Item	Cause of Improper Solutions			
	A	B1	B2	C1
(1) Does iteration converge?	yes	no	no	yes
(2) Is the solution stable? (not depending largely on estimation methods, starting values, nor optimization algorithms)	yes	unstable in one X_i	unstable in two particular X_i's	yes
(3) Are SE's† of $\hat{\psi}_i$ almost the same in magnitude?	yes	one large SE	one or two large SE('s)	yes
(4) Does the confidence interval of ψ_i, not positively estimated, contain zero?	yes	no	no	no
(5) Are residual elements of $S - (\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi})$ almost the same in magnitude?	yes	yes	no for solution reducing k by one	yes

Table 2: *Checklist for identifying the cause of improper solutions. See text for examining C2. †SE denotes standard error.*

In (B2) an improper solution occurs in the first or second variables only, because the loadings λ_{1k} and λ_{2k} can take any value as long as they satisfy

$\lambda_{1k}\lambda_{2k} = \sigma_{12} - \sum_{r=1}^{k-1} \lambda_{1r}\lambda_{2r}$. One can not make exploratory factor analysis when the population factor loading matrix Λ has the form (B2). In the case, the researchers can not help making a constraint to remove the indefiniteness, such as $\lambda_{1k} = \lambda_{2k}$ or $\psi_1 = \psi_2$. It is optional whether to impose $\lambda_{ik} = 0$ ($i = 3, \cdots, p$). See Kano (1997) for details.

When an improper solution occurs due to (B), underidentifiability, it often happens that iteration does not terminate; the solution depends on starting values, estimation methods (e.g., ML, GLS, LS) or optimization algorithms. In (B1) the location of only one nonzero element, or the variable in which $\psi_i < 0$ is arbitrary and it also depends on starting values etc. As noted above, negative estimates can appear at two particular variables for (B2). These observations will distinguish between (B1) and (B2).

For (C1) or (C2), researchers have to remove all variables inconsistent with the model considered. They can remove the variables with negative unique variances in the case (C1). In (C2) we could examine residuals to identify inconsistent variables. A more sophisticated manner would be to take a likelihood ratio test approach, developed by Kano and Ihara (1994).

The case (D) contains all causes other than (A)-(C). Outliers in samples may cause improper solutions, as pointed out by Bollen (1987). Existence of outliers is classified in the case (D). The other cases in (D) are yet unknown and still need to be studied.

In Table 2, we summarize as a checklist how to identify the cause of improper solutions. Table 3 presents the method of reanalysis for each cause of improper solutions.

Cause	Treatment
A	Obtain a boundary solution with all $\psi_i \geq 0$
B11	Refer to goodness of fit indices such as GFI and CFI
B12	Apply an ADF type of estimation method
B2	Estimate under constraint such as $\lambda_{1k} = \lambda_{2k}$ or $\psi_1 = \psi_2$
C1	Remove the variable X_i with $\psi_i < 0$
C2	Remove inconsistent variables

Table 3: *Treatment after identifying the cause of improper solutions*

3. Example

Maxwell (1961) conducted maximum likelihood factor analysis of each of two samples of 810 normal children, and 148 neurotic children attending a psychiatric clinic, where the first five items of the samples are cognitive tests and the other five are inventories for assessing orectic tendencies (see Table

7 for items). He found that the iterative process for obtaining the MLE does not terminate and the communality of the eighth variable approaches to one for 4-factor model for the normal sample whereas a 3-factor model successfully analyzes the sample of neurotic children. He concluded that the method of maximum likelihood along with goodness-of-fit testing does not always perform well. It later turned out that the 4-factor solution for the normal sample is improper (e.g., Jöreskog 1967).

Here, we shall take the sample of normal children to illustrate our procedure for improper solutions. In the sample, Sato (1987), among others, reported that the improper solution depends on initial estimates for iteration and gave three different improper solutions with $\hat{\psi}_6$, $\hat{\psi}_8$ or $\hat{\psi}_9$ negative, respectively, and that it is difficult to achieve convergence in iteration when uniqueness estimates are not constrained to be nonnegative. Table 4 shows MLE for uniqueness and their standard errors, for each case of $0 \leq \psi_i < \infty$ and $-\infty < \psi_i < \infty$. The analysis was made with a covariance structure program Eqs, developed by Bentler (1995).

		goodness-of-fit χ^2_{11}-value	P-value	ψ_1	ψ_2	ψ_3	ψ_4
$0 \leq \psi_i < \infty$	MLE	18.447	0.072	384	621	302	638
	SE			037	059	039	037
$-\infty < \psi_i < \infty$	MLE	14.589	0.202	397	629	298	644
	SE			037	056	040	037

		ψ_5	ψ_6	ψ_7	ψ_8	ψ_9	ψ_{10}
$0 \leq \psi_i < \infty$	MLE	352	778	287	000	690	599
	SE	096	040	052	003	037	050
$-\infty < \psi_i < \infty$	MLE	324	802	275	-276614	725	609
	SE	108	042	046	156	040	044

Table 4: *Uniqueness estimates* (MLE) *and their standard errors* (SE) *in Maxwell's data* ($n = 810$*;* $k = 4$)*. Values are multiplied by 1000.*

Obviously, the cause of the improper solution is not "A: sampling fluctuations." Maybe we should consider "B: identifiability" as a possible cause. Table 5 shows the list of top five (standardized) residuals in absolute value in 3-factor solution. There is no salient residual in the list, which implies that (B2) is not the cause. To conclude that the cause is (B1), we have to examine (B11) and (B12), but unfortunately, we can not check normality of observations because raw data are not available. We would say that the sample size $n = 810$ is so large that the power of the goodness-of-fit test is raised too much. In fact, other fit indices indicate reasonable fit of the 3-factor model, for instance, GFI=0.981; CFI=0.973. As a conclusion, a

X9-X8	X8-X6	X10-X8	X8-X4	X9-X6
0.095	0.084	-0.050	0.046	-0.044

Table 5: *Top five residuals in 3-factor solution*

probable cause of the improper solution is (B1) and the analysis here suggests a 3-factor model for the sample of normal children, as well as for the sample of neurotic children.

Any model is nothing but an approximation to reality, and deviation of a model from reality always exists. Statistical test can detect the deviation even when it is very small, provided that the sample size gets large. There are two considerations: (i) one employs a slightly misspecified model if one considers the deviation as just an error and so negligible; (ii) one rejects the model and finds a suitable treatment to reduce the deviation.

The treatment above for the improper solution of the normal sample is based on the consideration (i). There is an alternative story based on (ii). The cause of the improper solution is then (C) in this story. The key can be found in the list of the residuals in Table 5.

In Table 5, we can find that the top four residuals are related to the eighth variable. This indicates the possibility that the eighth variable be inconsistent with the model under consideration. The inconsistency could be a cause of the rejection of the 3-factor model. Table 6 shows the goodness-of-fit chi-square test statistics of 10 models, each of which is formed by removing one of 10 variables. The only accepted model is the one in which the eighth variable is removed. As a result, the eighth variable can be regarded as inconsistent with the 3-factor model. See Kano and Ihara (1994) for details.

	Variable Deleted									
	1	2	3	4	5	6	7	8	9	10
LRT	35.32	37.38	48.28	65.37	21.63	55.89	40.73	14.58	45.68	38.47

Table 6: χ^2_{12} *values of 10 models after deletion of one variable.* ($\chi^2_{12}(.05) = 18.55$)

Table 7 shows the MLE in 3-factor model after deletion of the eighth variable.

We have here suggested two possibilities for treatment of the improper solution in Maxwell's data. Which approach is to be taken, in others words,

	Factor 1	Factor 2	Factor 3	Communal.
COGNITIVE TESTS				
X1 Verbal Ability	574	412	323	603
X2 Spatial Ability	095	585	139	371
X3 Reasoning	406	697	225	702
X4 Numerical Ability	325	487	117	356
X5 Verbal Fluency	780	243	092	675
ORECTIC TENDENCIES				
X6 Neuroticism Questionnaire	104	175	396	198
X7 Way to be different	228	143	808	725
X8 Worries and Anxiety	---	---	---	---
X9 Interests	103	180	482	275
X10 Annoyances	000	028	625	391

Table 7: *3-factor solution, rotated by normalized-VARIMAX, of 9 variables after deletion of the 8th variable.* χ^2_{12}*-value=* 14.58 *(*$n = 810$*), P-value=.2712. Estimates are multiplied by 1000.*

whether the deviation of the 3-factor model using 10 variables can be considered small enough or not, may depend on researchers and also on interpretability of those two results of the analyses.

4. Remarks

Users may not implement the procedure described above, if they use usual exploratory factor analysis (EFA) programs only. It is absolutely necessary to use programs with which the user can (i) do analysis under no constraint on ψ_i, that is, ψ_i can take negative values; (ii) get standard errors of estimates, particularly, of $\hat{\psi}_i$; (iii) specify starting values; and (iv) get (standardized) residuals, $S - (\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi})$. For this, covariance structure analysis (CSA) programs (e.g., Amos, Eqs, Lisrel) are very useful, although they do not have the option of factor rotation. Researchers are recommended to use both CSA and EFA programs to make exploratory factor analysis in a proper way.

References

Anderson, J. C. and Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, **49**, 155-173.

Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. *Proceedings of the Third Berkeley Symposium on Mathematical*

Statistics and Probability, **5**, 111-150. Berkeley: University of California Press.

Bentler, P. M. (1995). *EQS Structural Equations Program Manual.* Multivariate Software, Inc.: CA.

Bollen, K. A. (1987). Outliers and improper solutions: A confirmatory factor analysis example. *Sociological Meth. Res.*, **15**, 375-384.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika*, **50**, 229-242.

Gerbing, D. W. and Anderson, J. C. (1985). The effect of sampling error and model characteristics on parameter estimation for maximum likelihood confirmatory factor analysis. *Multi. Behav. Res.*, **20**, 255-271.

Gerbing, D. W. and Anderson, J. C. (1987). Improper solutions in the analysis of covariance structures: Their interpretability and a comparison of alternative respecifications. *Psychometrika*, **52**, 99-111.

Ihara, M and Okamoto, M. (1985). Experimental comparison of least-squares and maximum likelihood methods in factor analysis. *Statist. Probab. Lett.*, **3**. 287-293.

Jöreskog K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, **32**, 443-482.

Jöreskog, K. G. and Sörbom, D. (1993). *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language* . Scientific Software International, Inc.: Chicago, IL.

Kano, Y. (1990). Noniterative estimation and the choice of the number of factors in exploratory factor analysis. *Psychometrika*, **55**, 277-291.

Kano, Y. (1997). Exploratory factor analysis with a common factor with two indicators. *Behaviormetrika*, **24**, 129-145.

Kano, Y. and Ihara, M. (1994). Identification of inconsistent variates in factor analysis. *Psychometrika*, **59**, 5-20.

Maxwell (1961). Recent trends in factor analysis. *Jour. Roy. Statist. Soc.*, **A124**, 49-59.

Sato, M. (1987). Pragmatic treatment of improper solutions in factor analysis. *Ann. Inst. Statist. Math.*, **39**, 443-455.

Van Driel, O. P. (1978). On various causes of improper solutions in maximum likelihood factor analysis. *Psychometrika*, **43**, 225-243.