Universidade Federal do Rio Grande do Sul
Instituto de Matemática

Heywood Cases in Unrestricted Factor Analysis

Jandyra M.G. Fachel

Cadernos de Matemática e Estatística

Série A, nº 12, JAN/90
Porto Alegre, janeiro de 1990

HEYWOOD CASES AND IMPROPER SOLUTIONS IN UNRESTRICTED FACTOR ANALYSIS

A Heywood solution is known in the literature of factor analysis as the occurrence of a negative or zero estimate of the error variance for one or more variables in any factor analysis solution. Occurrences of Heywood cases have been reported in the literature since the first observation of this kind of particular solution by Heywood (1931). Heywood cases may occur in any factor analysis method, they also occur in confirmatory factor analysis and there is some evidence in the literature that the maximum likelihood factor analysis method is particularly prone to the occurrence of Heywood cases. The causes for such occurrences are not still clearly understood and some new studies have tried to show, through empirical evidence, in which situations the occurrence of Heywood cases are more frequent.

We shall distinguish, in this paper Heywood solutions and improper solutions in factor analysis. The improper solutions in factor analysis that occur frequently are Heywood solutions, but not all Heywood solutions are improper solutions, and not every improper solution is a Heywood solution. Suppose we have a one-factor with one or more of the factor loading parameters very high or conversely, suppose one or more of the error variance parameters in factor analysis model are positive but very near zero. A factor analysis solution that yelds an exact (and no negative) zero error variance estimate, when the corresponding parameter is approximately zero, cannot be considered an improper solution. In this particular case, the only cause for the zero variance is the sampling variation and any small difference between the estimate and the parameter is only to be expected. From the practical point of view, we can have situations in which the one-factor model fits the data and one of the variables is perfectly correlated with te single factor, meaning that this variable itself could be a good indicator of the factor. Suppose a situation where the corresponding factor loading parameter for that variable is 0.98, say. A solution that yelds a factor loading estimate as 1.00 is not an "improper solution". It will be a Heywood case because the variance estimate of the error term for that variable is zero. But this is a proper solution, given the model.

There are also improper solutions in factor analysis that are not Heywood cases. If the true number of factors is known, any factor analysis solution, that has not the same numbers of factors as is assumed in the model, is an "improper solution". Unless we know the model for a particular factor analysis solution (as is the case in simulation studies), we cannot distinguish, in practical work, an improper solution from a Heywood solution, but, very frequently, when the number of factors is not that of the hypothesized model a Heywood case will indicate an improper solution, as we shall see in a simulation study to be presented in this paper.

Although we shall consider only unrestricted factor analysis in this study, a review of the early research about Heywood cases will be made, considering also confirmatory factor analysis.

Martin and McDonald (1975) distinguish two types of Heywood solution: an exact Heywood solution when at least one unique variance is zero but none are negative and ultra-Heywood solution where at least one unique variance is negative. Ultra Heywood cases are, obviously, improper solutions, because we cannot have negative variances. But an exact Heywood solution may not be an improper solution as we explained before.

Most of the factor analysis programs available in the statistical analysis packages do not allow the communalities of the variables to exceed one. That is the case for the BMDP and SPSS packages. Some of the factor analysis programs in the SAS package have the option for ultra-Heywood cases, that is, they allow communalities to exceed one. Therefore on using either BMDP, SPSS (or SPSS-X) an ultra Heywood case will not be observed, although SPSS will print "the communality is greater than one" and will stop the iteration process.

In past research, there are some simulation studies relevant to the present study, although some of them are concerned with the confirmatory factor analysis model. We now review these studies.

Tumura and Fukutomi (1970) have presented some numerical experiments to investigate the occurrence of Heywood cases in six different cases, where the uniqueness of the solution is considered and also where the given number of factors (m) for the solution is different from the true number of factors of the model. Joreskog's unrestricted maximum likelihood factor analysis method was considered in the study, which is limited in the sense that only one or two experiments per case was analysed. Nevertheless, the authors conclude that for the case where Λ is unique and m=k, Heywood cases "occur occasionally if Λ contains some row vector with their length equal to nearly one" (see also Tumura, Fukutomi and Asoo, 1968).

A Monte Carlo study is presented by Boomsma (1985) to assess the problems of nonconvegence, improper solutions and starting values in LISREL maximum likelihood ratio chi-square statistic for goodness-of-fit are also presented. Twelve factor analysis models were studied, all having two factors (correlated and not correlated factors). The factor pattern Λ (px2), where p is the number of observed variables, was chosen such that half of the observed variables had a non zero loading on the first factor and a zero loading on the second one, and the reverse for the other half (p = 6 or 8). The sizes of the factor loadings were chosen as small (0.4; 0.6); medium (0.6; 0.8) and large (0.8; 0.9). The sample sizes were 25, 50, 100, 200 and 400 (with 300 replications of each). In this study, Boomsma considers only the ultra-Heywood cases (negative estimates of the error variance). She concludes that "there is a real danger of improper solutions" with small sample size. In the simulation results, the occurrence of improper solutions increased as 1) sample size decreased; 2) the number of variables in the model was six rather than eight and 3) the population values of the error variance were close to zero.

Anderson and Gerbing (1984) also present a Monte Carlo study for the LISREL confirmatory factor analysis method. They analyse 54 models, with 2, 3 or 4 factors, for sample sizes of 50, 75, 100, 150 and 300 (with 100 replications of each). The proportion of nonconvergent and improper solutions that occurred in obtaining 100 good solutions per cell is presented. They conclude that a sample size of 150 for models with three or more indicators per factor (6 or more variables in the model) will usually be sufficient for a convergent and proper solution. In this study the solutions are defined as improper when one or more of the unique variances is less than a positive, arbitarily small prescribed number such as 0.005. Anderson and Gerbing also observed that the occurrence of improper solutions increased as 1) sample size decreased; 2) the number of indicators per factor (and consequently the number of variables in the model) decreased; 3) correlation between factors were 0.3 rather than 0.5. For the models analysed, they also observe that with two indicators per factor (small number of variables), loadings of 0.9 give the largest proportion of improper solutions, whereas for larger numbers of variables no improper solutions occurred for models with loadings 0.9. Results on goodness-of-fit indices are also presented in this Monte Carlo study.

Seber (1984) reports some results from a simulation study by Francis (1973, 1974). Francis' analysis is based on exploratory or unrestricted factor analysis models with two or three factors. The sample size is 50. Twelve models were generated with different factor patterns. Again in this case the solution is said to be improper if the error variances are less than an arbitrary small positive number (e.g., 0.005). Several cases of improper solutions were observed when the number of factors for a particular solution was greater than the true number of factors of the model.

Other researchers have proposed methods to avoid the occurrence of Heywood cases for detecting the causes of Heywood cases. We now review briefly these methods.

Joreskog (1967) proposes a procedure to deal with improper solutions for the maximum likelihood factor analysis method. He defines the problem of improper (Heywood) solution as follows: "Since the diagonal elements of ψ are variances the function $f_k(\psi)$ is defined in the region where all the diagonal elements of ψ are positive" (k is the number of factors). "We have no guarantee, however, that all partial derivatives of f_k vanish at a point where all the diagonal elements of ψ are positive. This suggests that we shall define $f_k(\psi)$ in the region R_ε, where $\psi_{ii} \geq \varepsilon$ for all i = 1, 2, ...p and where ε is a positive, arbitrarily small, prescribed number. The problem, then, is to find the minimum of $f_k(\psi)$ in the region R_ε. Since R_ε is a closed region, the minimum is found either in the interior of R or in the boundary. If the minimum is found in the interior of R_ε, we shall say that the minimum is a proper solution. If on the other hand, the minimum is found on the boundary of R_ε, the solution is improper".

We have transcribed Joreskog's text because it seems to be the origin of the term "improper solution", which has been used frequently. Joreskog (1967, p.443) also says that "such improper (Heywood) solutions occur more often than is usually expected." The procedure that he proposes to avoid such improper solutions is to eliminate partially the variables with unique variances equal to ε and the analysis continues from the conditonal dispersion matrix. The solution finally accepted in this process is combined with the principal components of the eliminated variables, to give a complete solution for all the original variables.

Martin and McDonald (1975) propose a Bayesian procedure for estimation in unrestricted factor analysis. The procedure has as one of its objectives to avoid inadmissible estimates of unique variances. A choice of the form of the prior distribution is justified and empirical examples are shown.

Finally, we will review the paper by Van Driel (1978) which has been cited in almost all studies about Heywood cases. Van Driel has identified some of the causes of Heywood solutions, by dropping the constraints of positive definiteness of the matrices containing the parameters of the factor analysis model. He proposes a method, which he calls "the nonclassical approach" and analyses some artifical data drawn from 5 populations, corresponding to five factor analysis models. The models are called: "Close to zero" (one of the unique variances is close to zero and the others are equal to 0.5); "Close to one" (one of the unique variancesis close to one and the others are equal to 0.5); "Dwarf" (all unique variances are equal to 0.5 and the second factor has loadings very small comparing with the first factor); "Heywood" (the classical one-factor model example where one of the unique variances iis supposed to be negative) and "Anderson and Rubin" (a three-factor model with unique variances equal to 0.5; the factor matrix for this population is in accordance with the Anderson and Rubin identification condition). In this study five samples are drawn from each population, each with sample size 800, and each sample is analysed with the classical and non-classical approach for every appropriate number of factors.

Van Driel (1978) referring to Joreskog's paper calls attention to the "subtle" difference between the terms "improper solution" and Heywood cases, but he uses the term improper as meaning Heywood solutions (that is, at least one unique variance negative or zero - small values of the variance, such as 0.004 are considered proper by Van Driel, as for example in the "close to zero" example). Van Driel identifies three causes for Heywood cases:

1) sampling fluctuations combined with true values of ψ close to zero;

2) there does not exist any factor analysis model that fits the data;

3) indefiniteness of the model (e.g. too many true factor loadings are zero)

Starting from the results of the previous studies two main questions arise:

1) How "close to zero" should be the unique variance parameters in the factor analysis model to cause Heywood cases?

2) How often do Heywood cases occur as a consequence of chosing a given number of factors different from the true number of factors of the model?

The first question is approached by Boomsma (1985) when she generates models with "large", "medium" and "small" factor loadings leading to different magnitudes of the unique variances of the model. Boomsma's results are, however, for confirmatory factor analysis using the LISREL program. We shall present some results for unrestricted factor analysis.

Concerning the second question, suppose m is the given number of factors for a particular factor analysis solution and q is the true number of factors of the model. We observed that Tumura and Fukutomi (1970) did not obtain Heywood cases when $m>q$ in their numerical experiments, but on the other hand several cases of Heywood solutions are reported by Seber with reference to Francis' results when m.q (see Seber 1984, p.232, Table 5.20). We also observed that several numerical examples presented by Joreskog show the occurrence of Heywood solutions when icreasing the number of factors for a particular example (see Joreskog, 1967, p.474, Table 8). We then suppose that another possible cause of Heywood cases is the inappropriateness of the solution for a given number of factors.

To assess the effect of sampling variation and model characteristics on the occurrence of Heywood cases for unrestricted factor analysis using the maximum likelihood method, a Monte Carlo study was designed. As a by-product of the study some results about the goodness-of-fit test of the model are also obtained. This simulation study is described in the next section.

The effect of sampling variation and model characteristics on the occurrence of Heywood cases for maximum likelihood factor analysis: a simulation study.

Our first objective is to study how the normal theory estimators for maximum likelihood unrestricted factor analysis perform regarding the occurrence of Heywood cases for models with specified characteristics. Estimates of the MLFA model are provided by the BMDP factor analysis program using an algorithm developed by Jenrich and Sampson (see Dixon et al, 1983).

The normal random variates are created using the Random Number Generator of the BMDP package.

Simulation design

For this Monte Carlo study three one-factor models were chosen for different magnitudes of the first factor loading. The one-factor model for variables with mean zero is given by

$$x_i = \lambda_i z + e_i \qquad i = 1,2,\ldots,p$$

such that $\mathrm{var}(x_i) = 1$, $\mathrm{var}(e_i) = \psi_i$, z and e_i are the normal generated variables and

$$h_i^2 + \psi_i = 1$$

where h_i^2 is the communality of the i-th observed variable, and ψ_i unique variance or error variance. The first factor pattern Λ (px1), where p is the number of observed variables was chosen such that the first observed variable had a "close to zero" unique variance or a very high loading ($\lambda_1 = 0.98$) and all other loadings equal to 0.5 ($\lambda_j = 0.5$, $j \neq 1$). This model will be called Model I.

The other models are similar, but the idea was to vary the first loading in such a way that we had three different degrees of "close to zero" variances. The last model having far from zero but not "close to one" unique variance. The three models are:

Model I $\quad \Lambda_I = \langle\ \lambda_1 = 0.98;\ \lambda_i = 0.5\ (i \neq 1)\ \rangle$ or

$$\psi_I = \langle\ \psi = 0.0396;\ \psi_i = 0.75\ (i \neq 1)$$

$\underline{\text{Model II}}$ $\Lambda_{II} = \{ \lambda_1 = 0.90;\ \lambda_i = 0.5\ (i \neq 1) \}$ or

$$\psi_{II} = \{ \psi_1 = 0.19;\ \psi_i = 0.75\ (i \neq 1) \}$$

$\underline{\text{Model III}}$ $\Lambda_{III} = \{ \lambda_1 = 0.70;\ \lambda_i = 0.447\ (i \neq 1) \}$ or

$$\psi_{III} = \{ \psi_1 = 0.51;\ \psi_i = 0.8\ (i \neq 1) \}$$

For each model three different numbers p of observed variables were analysed so as to represent a range of values typically encountered in practice (p=5; p=10 and p=20). Sample sizes were chosen according with the criterion: small (N=50); medium (N=100) and large (N=500). For each cell of this design, 100 replications were generated.

Finally, to assess the effect of having m>p on Heywood cases, where m is the given number of factors in one solution and q is the true number of factors of the model (q=1 in this case), we chose to analyse the correlation matrices generated by Model III (where the occurrence of Heywood cases is assumed to be very small with a two-factor solution.

Although the above design produced 3600 separate analyses, this is a limited Monte Carlo study, with respect to different models studied, different sample sizes and number of replications. Nevertheless, the study should give a good deal of important information related to the occurrence of Heywood cases, standard erros of the MLFA estimators, results on the likelihood criterion given by th MLFA (BMDP) program and results about the empirical frquency distribution of the eigenvalues. Our results are, however, limited to the cases here studied, no generalizations beyond these models will be made.

Results

The MLFA/BMDP program produces factor loadings estimates and unique variances within the parameter space or on the boundary. No ultra Heywood cases can be observed, because of the constraints in the program. The results to be presented in this section are related to the proportion of exact Heywood cases for each model. In table 1.1 we present the percentage of Heywood solutions in each cell of the simulation design for Models I and II. For each cell we observed 100 replications.

Table 1.1 - Proportion of Heywood solutions for Models I and II in the Monte Carlo study (100 replications per cell).

Model	No of var	Sample Size 50	100	500
	5	.23	.32	.36
MODEL I	10	.18	.04	.01
[$\lambda_1 = 0.98$]	20	.20	.16	.00
	5	.04	.03	.00
MODEL II	10	.00	.02	.00
[$\lambda_1 = 0.90$]	20	.00	.00	.00

The proportion of Heywood solutions decreases as the sample increases, in general, although when p=5 this was not always observed. A greater proportion of Heywood solutions is observed for a small number of variables in the model. For Model I and small sample sizes a greter proportion of Heywood solutions is observed when the number of variables is 5 or 20 rather than 10. A greater number of replications per cell would be necessary to confirm this tendency. The results in Table 1.1 are in accordance with the findings of Van Driel (1978), that is the "close to zero" population is one of the causes for Heywood cases combined with sampling variation. For Model I the Heywood cases were observed always for the first variable ("close to zero case"), therefore these solutions are very similar to the true model (we observe a factor loading $\hat{\lambda}_1 = 1.0$ where the parameter is $\lambda_1 = 0.98$. Due to sampling variation, solutions with the first loading equal to one (and consequently unique variance equal to zero) are expected to occur and such a solution cannot be called "improper". They are exact Heywood cases, but the solution is proper.

In table 1.2 we present the proportion of Heywood solutions out of 100 replications for Model III for different sample sizes. We also show the proportion of Heywood solutions that occur as a result of a simulated misspecification of the model, that is, we knew that the model had one factor, but we asked the program to produce the two-factor solution. We then observed a very high proportion of Heywood solutions for two factors and even more than one variable with zero variance. The Heywood cases were observed for any variable, not always for the first as in the case of Models I and II. In this case we have improper solutions. We then conclude that another cause for Heywood cases is the inclusion of too many factors in the solution. We believe that many of the Heywood solutions observed in the literature are due to the fact that they are over-factored (e.g. too many factors). When analysing empirical data, it is impossible to know the true number of factors of the model. In the simulation studies we know the model, but this is an artificial situation. We suggest that, in empirical situations, when the researcher is using factor analysis and obtains a Heywood solution, he should reanalyse the data decreasing the number of factors one by one. If the goodness-of-fit indices are good , that should be the best solution for factor analysis.

Table 1.2 - Proportion of Heywood solutions for Model III using one-factor solution and two-factor solution (100 replications per cell).

Model	No of var	Sample Size 50	100	500
MODEL III	5	.02	.00	.00
	10	.00	.00	.00
$[\lambda_1 = 0.70]$	20	.00	.00	.00
MODEL III	5	.67	.71	.58
	10	.59	.54	.36
(with two factors)	20	.48	.24	.14

As it can be seen in Table 1.2, the proportion of Heywood solutions indicating an improper solution for two factor solutions is very high even for large sample size. The proportion seems to decrease as the number of variables increases. Model III is a one-factor model with the following error variances $[\psi_1 = 0.51$ and $\psi_i = 0.80,\ i \neq 1]$. It is interesting to observe that for the one-factor solution and for sample size 50, we observe cases with communality very near zero, or variances very near one, producing negative estimates of loadings, which could be considered as another kind of improper solution; the proportion of these cases was very small (1 case for p=5 and p=20 and two cases for p=10, all for N=50).

As a by-product of this simulation study we shall now present results about the Chi-square test which can be obtained from the likelihood criterion (LC) to be minimized. (The BMDP/MLFA program only prints the likelihood criterion). The Chi-square statistic can then be obtained by $X^2 = n'LC$, where n' is given by

$$n' = N-1-(2p+5)/6-2q/3$$

The X^2 statistic for the unrestricted factor analysis model is tested as a chi-square variable with degrees of freedom given by

$$df = 1/2[(p-q)^2 - (p+q)].$$

In Table 1.3 we present the proportion of significant chi-square values for $\alpha = 0.05$, for 100 replications in each cell for the three models analysed. We also include in Table 1.3 the results for the two-factor solutions for Model III.

Table 1.3 - Proportion of significant chi-square statistics (5%) for models I, II and III and model III with two factors (100 replications per cell).

Model	No of var	Sample Size 50	100	500
	5	.03	.11	.14
MODEL I	10	.01	.12	.03
	20	.04	.00	.10
	5	.03	.13	.12
MODEL II	10	.04	.13	.03
	20	.03	.03	.06
	5	.04	.05	.05
MODEL III	10	.05	.03	.02
	20	.00	.02	.04
	5	.01	.03	.03
MODEL III	10	.00	.00	.00
(with two factors)	20	.00	.01	.02

We observe in Table 1.3 that for small samples the observed proportion of significant chi-square statistics is higher than the expected proportion of 0.05 for the one-factor model. When analysing the two factor solution the test accepts the model with two factors, which should not be accepted. But this is a known fact of this goodness-of-fit test, because it depends on the residual correlations, if we include more factors in the model the residuals become smaller, and consequently, the chi-square statistic. The factor analysis user should, for this reason, use more than one goodness-of-fit indice, including in the analysis other criteria such as Akaike's Information Criterion and Schwarz's Bayesian criterion (see Section 4.3).

Another interesting result from this simulation study is the empirical frequency distribution of the number of eigenvalues greater than one ($\gamma > 1$) of the correlation matrices. In Tables

1.4 to 1.6 we present these empirical frequencies, for each of the one-factor models studied, according to sample size and number of variables in the model.

Table 1.4 - Empirical distribution (proportion) of the numbers of eigenvalues greater than one, for Model I (100 replications per cell).

No of var	No. of eigenvalues > 1	Sample Size 50	100	500
5	1	.75	.89	1.00
	2	.25	.11	-
10	1	.05	.29	1.00
	2	.36	.60	-
	3	.57	.11	-
	4	.02	-	-
20	1	-	-	.19
	2	-	-	.63
	3	-	-	.18
	4	.03	.16	-
	5	.23	.40	-
	6	.51	.40	-
	7	.20	.04	-
	8	.03	-	-

Table 1.5 - Empirical distribution (proportion) of the numbers of eigenvalues greater than one for Model II (100 replications per cell).

No of var	No. of eigenvalues > 1	Sample Size 50	100	500
5	1	.80	.92	1.00
	2	.20	.08	-
10	1	.03	.25	1.00
	2	.36	.61	-
	3	.54	.14	-
	4	.07	-	-
20	1	-	-	.27
	2	-	-	.54
	3	-	.02	.19
	4	-	.14	-
	5	.18	.47	-
	6	.52	.31	-
	7	.29	.06	-
	8	.01	-	-

Table 1.6 - Empirical distribution of the number of eigenvalues greater than one for Model III (100 replications per cell).

No of var	No. of eigenvalues > 1	Sample Size 50	100	500
5	1	.40	.79	1.00
	2	.60	.21	-
10	1	.01	.04	.90
	2	.16	.35	.10
	3	.48	.55	-
	4	.35	.06	-
20	1	-	-	-
	2	-	-	.07
	3	-	-	.56
	4	-	-	.34
	5	.07	.11	.03
	6	.28	.48	-
	7	.53	.38	-
	8	.12	.03	-

The Tables 1.4 to 1.6 show that small sample size and for a number of variables in the model such as 20, several eigenvalues of the correlation are greater than one for the one-factor models I, II and III. If the factor analysis user chooses the number of factors by this criterion, as is still very common, with small samples and large number of variables, the inclusion of too many factors in the solution would occur. Even for a moderate sample size such as 100, that would be the case. On the other hand, for all models and cases, the scree test would be more appropriate since the magnitude pattern of the eigenvalues always shows a very high first eigenvalue compared with the others.

Finally, as another by-product of the simulation study we now present the results related to the parameter estimates of the models. In Table 1.7 we present the mean and standard deviation of the first factor loading for the three models, for each cell of the simulation design, based on 100 replications. We have included the Heywood solutions in all calculations. In Tables 1.8 and 1.9 we present the mean and standard deviation of the second and third factor loading for each model, respectively.

Table 1.7 - Mean and standard deviation of the first factor loading estimates for models I, II and III (100 replications per cell).

Model	No of var	Sample Size 50	100	500
	5	.942	.964	.984
		(.058)	(.033)	(.017)
MODEL I	10	.967	.964	.979
		(.033)	(.021)	(.009)
	20	.984	.985	.984
		(.017)	(.011)	(.005)
	5	.847	.893	.903
		(.106)	(.061)	(.025)
MODEL II	10	.887	.940	.902
		(.051)	(.046)	(.017)
	20	.899	.910	.903
		(.066)	(.024)	(.011)
	5	.617	.671	.699
		(.143)	(.095)	(.041)
MODEL III	10	.661	.620	.706
		(.106)	(.076)	(.031)
	20	.739	.699	.715
		(.075)	(.061)	(.027)

Note: Estimates are based on averaging the estimates in each cell. In parenthesis is the empirical satandard deviation.

Table 1.8 - Mean and standard deviation of the second factor loading estimates for models I, II and III (100 replications per cell).

		Sample Size		
Model	No of var	50	100	500
	5	.643	.590	.509
		(.060)	(.072)	(.029)
MODEL I	10	.529	.523	.474
		(.108)	(.056)	(.034)
	20	.483	.460	.487
		(.115)	(.071)	(.031)
	5	.661	.590	.518
		(.090)	(.070)	(.053)
MODEL II	10	.517	.516	.473
		(.113)	(.062)	(.029)
	20	.504	.469	.492
		(.095)	(.072)	(.032)
	5	.609	.594	.462
		(.167)	(.098)	(.042)
MODEL III	10	.487	.449	.412
		(.108)	(.091)	(.039)
	20	.446	.395	.434
		(.122)	(.090)	(.032)

Table 1.9 - Mean and standard deviation of the third factor loading estimates for models I, II and III (100 replications per cell)

Model	No of var	Sample Size 50	100	500
	5	.494	.508	.489
		(.084)	(.077)	(.032)
MODEL I	10	.484	.528	.494
		(.101)	(.068)	(.032)
	20	.430	.428	.487
		(.108)	(.082)	(.029)
	5	.477	.523	.486
		(.117)	(.077)	(.036)
MODEL II	10	.465	.526	.497
		(.110)	(.073)	(.029)
	20	.499	.443	.490
		(.096)	(.076)	(.035)
	5	.395	.462	.432
		(.141)	(.094)	(.042)
MODEL III	10	.421	.466	.444
		(.118)	(.091)	(.040)
	20	.463	.476	.433
		(.106)	(.074)	(.036)

Discussion

In this simulation study the effect of sampling variation and model characteristics on the occurrence of Heywood cases was analysed. Two main causes of Heywood solutions in factor analysis were observed for the models analysed:

1) sampling variation combined with unique variance parameters close to zero, which is in accordance with Van Driel (1978);

2) misspecification of the model - too many factors are

Universidade Federal do Rio Grande Sul
Reitor: Professor Tuiskon Dick

Instituto de Matemática
Diretor: Professor Aron Taitelbaum
Núcleo de Atividades Extra Curriculares
Coordenador: Professora Jandyra G. Fachel
Secretária: Rosaura Monteiro Pinheiro

Os Cadernos de Matemática e Estatística publicam as seguintes séries:

Série A: Trabalho de Pesquisa
Série B: Trabalho de Apoio Didático
Série C: Colóquio de Matemática SBM/UFRGS
Série D: Trabalho de Graduação
Série E: Dissertações de Mestrado
Série F: Trabalho de Divulgação
Série G: Textos para Discussão

Toda correspondência com solicitação de números publicados e demais informações deverá ser enviada para:

NAEC - Núcleo de Atividades Extra Curriculares
Instituto de Matemática - UFRGS
Av. Bento Gonçalves, 9500
91.500 - Agronomia - POA/RS
Telefone: 36.11.59 ou 36.17.85 Ramal: 252

included in one particular solution causing improper solutions

The occurrence of Heywood cases is much more frequent for small sample sizes. Factor analysis based on fifty or less observations should certainly be avoided, not only because of a higher possibility of Heywood cases, but because the sampling fluctuations may lead to solutions that differ substantially from the true model.

It was observed, generally, that the occurrence of Heywood cases increases as the number of the variables of the model decreases.

Our results, for unrestricted factor analysis, are in accordance with the findings of Boomsma (1985) and Anderson and Gerbing (1984) for confirmatory factor analysis, concerning the occurrence of Heywood cases.

For normal theory, the chi-square test has been shown to behave well although a higher proportion than the expected rejects the model for small sample sizes and moderate number of variables (p=20).

The results from the simulation study also show that the Kaiser criterion ("eigenvalues greater than one") choosing the number of factors should not be used as it may lead to the occurrence of Heywood cases, caused by the inclusion of too many factors in a solution, mainly if the sample size is small and the number of variables large. With large sample sizes and small number of variables, the criterion may be useful if used together with other criteria.

As a final comment we strongly advise that sample sizes of 100 or more are needed for reasonable factor analysis results.

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.

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

Dixon, W.J. and Brown, M.B. (eds) (1983). BMDP Statistical Software (1983 Printing with additions). Berkeley: University of California Press.

Francis, I. (1973) . Factor analysis: its purpose, practice and packaged programs. Invited paper, American Statistical Association. NY, December 1973.

Francis I. (1974). Factor analysis: fact or fabrication. Mathematical Chronicle, 3, 9-44.

Gweke, J.F. and Singleton, K.J. (1980). Interpreting the likelihood ratio statistic in factor models when sample size is small. Journal of the American Statistical Association, 75, 133-137.

Harman, H.H. and Fukuda, Y. (1966). Resolution of the Heywood case in the MINRES solution. Psychometrika, 31, 563-571.

Harman H.H. and Jones, W.H. (1966). Factor analysis by minimizing residuals (MINRES). Psychometrika, 31, 351-368.

Heywood, H.B. (1931). On finite sequences of real numbers. Proceedings of the Royal Society of London, 134, 486-501.

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

Martin, J.V. and McDonald, R.P. (1975). Bayesian estimation in unrestricted factor analysis: A treatment for Heywood cases. Psychometrika, 40, 505-517.

SAS Institute (1985). SAS User's guide: Statistics, Version 5 edition. Cary, NC: SAS Institute Inc.

Tumura, Y. and Fukutomi, K. (1970). On the improper solutions in factor analysis. TRU Mathematics, 6, 63-71.

Tumura, Y.; Fukutomi, K. and Asoo, Y. (1968). On the unique convergence of iterative procedures in factor analysis. TRU Mathematics, 4, 52-59.

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

Publicações do Instituto de Matemática da UFRGS
Cadernos de Matemática e Estatística

Série A: Trabalho de Pesquisa

1. Marcos Sebastiani - Transformation des Singularités - MAR/89.
2. Jaime Bruck Ripoll - On a Theoremof R. Langevin About Curvature and Complex Singularities - MAR/89.
3. Eduardo Cisneros, Miguel Ferrero e Maria Inés Gonzales - Prime Ideals of Skew Polynomial Rings and Skew Laurent Polynomial Rings - ABR/89.
4. Oclide José Dotto - ϵ - Dilations - JUN/89.
5. Jaime Bruck Ripoll - A Characterization of Helicoids - JUN/89.
6. Mark Thompson, V. B. Moscatelli - Asymptotic Distribution of Liusternik-Schnirelman Eigenvalues for Elliptic Nonlinear Operators -JUL/89.
7. Mark Thompson - The Formula of Weyl for Regions with a Self- Similar Fractal Boundary - JUL/89.
8. Jaime Bruck Ripoll - A Note on Compact Surfaces with Non Zero Constant Mean Curvature - OUT/89.
9. Jaime Bruck Ripoll - Compact ϵ - Convex Hypersurfaces - NOV/89.
10. Jandyra Maria G. Fachel - Coeficientes de Correlação Tipo -Contingência - JAN/90.
11. Jandyra Maria G. Fachel - The Probability of Ocurrence of Heywood Cases - JAN/90.
12. Jandyra Maria G. Fachel - Heywood Cases in Unrestricted Factor Analysis - JAN/90.