Mem. Fac. Lit. & Sci., Shimane Univ., Nat. Sci., 2, pp. 14–21, March 10, 1969

On the Centroid Method in Factor Analysis

Yasuhiro ASOO
Department of Mathematics, Shimane University, Matsue, Japan

Yosiro TUMURA
Department of Applied Mathematics, Science University of Tokyo

Kazuo FUKUTOMI
Shibaura Technical College

(Received October 30, 1968)

The centroid method is a popular method in factor analysis. Its mathematical structure, however, is not so well specified. D. N. Lawley (2) treated the efficiency problem of this method in the case of covariance matrix with known residual variances. In the present paper, we treat the case of correlation matrix from another point of view. We specify the mathematical structure in the following two cases; one of them is the case of known residual diagonals, and the other is the case of unknown residual diagonals. For the first case, we give the complete justification of this method, and for the second, we treat the following problems; suppose that we know the number of factors only, then, 1. how are effects of initial estimates of communalities in the iterative centroid method, 2. when they converge to the unique limit, whether is it the true solution or not. In the usual complete centroid method, we always adjust residual diagonals after every factor extraction. However, this adjustment may not be appropriate to see effects of initials. We consider four approaches:

1. the complete centroid method with the highest row element as initials, without adjustment and/or with adjustment,

2. the complete centroid method with arbitrary initial estimates of communalities, without adjustment and/or with adjustment. Our problems are the following;

1. how are effects of initial estimates....when they have unique limit are they true solution? By the initial estimate of 1. 0's, or initial estimate greater than true, can we get the true solution?

2. without adjustment of residual diagonals, can we get real solution for some initial estimates, for example, the highest row element. In addition to these problems, we treat the iterative centroid method with true communalities in the case of adjustment and non-adjustment.

In the centroid method, the most troublesome point is the treatment of

reflection of test vectors. In the present paper we treat this point in the same line with usual complete centroid method. With respect to the maximum likelihood method after the computational method of Jöreskog, including the same type of problems as the above, see Tumura et al. (3). Computations are made with IBM 1620 in Science University of Tokyo.

Chapter 1. Case of known communalities

We set up the model of factor analysis as the following; Let every component of real random p-vector, X_i $(i = 1, 2, \ldots, p)$, belong to the space L_2 (Ω) of some probability space Ω. Without the loss of generality, we suppose that $\int_\Omega X dp = 0$ for every i. Define the scalar product in the space L_2 (Ω) as $(X_i, X_j) = \mathrm{cov}(X_i, X_j) = \int_\Omega X_i X_j \, \mathrm{dp}, (X_i, X_j) = 1$. By the theorem of Beppo-Levi, we have a unique orthogonal decomposition of the space $L_2(\Omega)$; $X_i(\omega) = X_{i(1)}(\omega) + X_{i(2)}(\omega)$, $(X_{i(1)}, X_{i(2)}) = 0$. With the dimension of the space $L_2^{(1)}$, $\dim L_2^{(1)} = k$, we define the orthogonal basis $[f_1, f_2, \ldots, f_k]$ in the space $L_2^{(1)}$; $X_{i(1)} = \sum_{\nu=1}^{k} a_{i\nu} f_\nu(\omega)$, and moreover we could define the basis of the space $L_2^{(2)}$ (Ω), $[s_1, s_2, \ldots, s_p, e_1, e_2, \ldots, e_p, \ldots]$ as satisfying relations $(f, s) = 0$, $(f, e) = 0$, $(s, e) = 0$, and so on. The space $L_2[f_1, \ldots, f_k] = L_2^{(1)}(\Omega)$ spanned by vectors $f_1, \ldots, f_k$ corresponds to the common factor space, and coefficient $a_{i\nu}$ in the expression $X_{i(1)} = \sum_{\nu=1}^{k} a_{i\nu} f_\nu$ is called the factor loading of the i-th test on the ν-th common factor, and the quantity $h_i^2 = \sum_{\nu=1}^{k} a_{i\nu}^2$ is called the communality of the i-th test. The space $L_2[s_1, \ldots, s_p]$ corresponds to the specific- factor space, and the space $L_2[e_1, \ldots, e_p]$ spanned by vectors $e_1, \ldots, e_p$ corresponds to the error space.

As the fundamental problem of the factor analysis, we consider the existence of the (minimal) number k and the space $L_2(\Omega)$ satisfing the above decomposition, and the uniqueness of such solutions. We call the solution consistent with the above model the "proper" solution and in the other cases the "improper" solution, the latter is the formal solution. For the proper solution, the communalities belong to the interval (0, 1), necessarily. In this section, we consider the case of known true communalities and show that by the centroid method we could get the true whether it is proper or not (improper). The case when we could calculate communalities by off-diagonal elements (under the assumption of uniqueness) also is contained in this case.

Let $X_{i(1)} = \sum_{\nu=1}^{k} a_{i\nu} f_\nu$, then the centroid of $X_{i(1)}$' s $(i = 1, 2, \ldots, p)$ is the point $\left(\frac{1}{p}\sum_{i=1}^{p} a_{i1}, \ldots, \frac{1}{p}\sum_{i=1}^{p} a_{ik}\right)$ in the space $L_2[f_1, \ldots, f_k]$. The transformation G; $g_i = G f_i$ $(i = 1, 2, \ldots, k)$ is the orthogonal transformation such that the vector g_1

is on the direction of this centroid. Under the new basis g's, we have $X_{i(1)} = \sum_{\nu=1}^{k} b_{i\nu} g_\nu \, (i = 1, \ldots, p)$, and $\sum_{i=1}^{p} X_{i(1)} = g_1 \sum_{i=1}^{p} b_{i1}$, thus we have the relation $\| \sum_{i=1}^{p} X_i^{(1)} \|^2 = (\sum_{i=1}^{p} b_{i1})^2 = \sum_{i \neq j} \sigma_{ij} + \sum_{i=1}^{p} h_i^2$, which is proportional to the square of the distance of centroid from the origin. Also we have the expression

$$(\sum_{j=1}^{p} X_j^{(1)}, X_i^{(1)}) = b_{i1} \sum_{j=1}^{p} b_{j1},$$

whence $b_{i1} = \pm (h_i^2 + \sum_{j \neq i} \sigma_{ij}) / \sqrt{\sum_{i \neq j} \sigma_{ij} + \sum_{i=1}^{p} h_i^2}$, and we may take the positive sign, if necessary. Next the vector $X_i^{(1)} - b_{i1} g_1 = \sum_{\nu=2}^{k} b_{i\nu} f_\nu \varepsilon L_2^{(1)}(\Omega)$ is perpendicular to the space $L_2[g_1]$, and the vector $-(X_i^{(1)} - b_{i1} g_1) = \sum_{\nu=2}^{k} (-b_{i\nu}) f_\nu$ also belongs to the space $L_2^{(1)}(\Omega)$ and is perpendicular to the space $L_2[g_1]$. In this case, from the relation $(X_i^{(1)} - b_{i1} g_1, X_j^{(1)} - b_{i1} g_1) = \sigma_{ij} - b_{i1} b_{j1} \, (i \neq j), = h_i^2 b_{i1}^2 \, (i = j)$, we have the expression $\| \sum_{i=1}^{p} (X_i^{(1)} - b_{i1} g_1) \|^2 = 0$. Thus, under the appropriate substitution of the vector $X_i^{(1)} - b_{i1} g_1$'s by the vector $-(X_i^{(1)} - b_{i1} g_1)$'s, we colud make the expression $\| \sum_{1=i}^{p} (X_i^{(1)} - b_{i1} g_1) \|$ nonzero, and in the same argument as the above, we have the 2nd factor loading, and so on until all arguments of the residual matrix vanish.

Herewith we note that the so-called sign-change is in a sence arbitrary in this case. When we know communalities completely, we could define factor structure uniquely. On the other hand, in some cases, we could determine communalities formally by the rank of off-diagonals—that is ,by the ideal rank of matrix. We give some examples : Example 1. $p = 6$, $k = 2$ (proper solution)

The true loadings are the following :

$$\begin{pmatrix} .8 & .4 & .2 & .9 & -.7 & .3 \\ .5 & .8 & .9 & .4 & .6 & .7 \end{pmatrix}$$

and the corresponding correlation matrix is the following;

$$\begin{pmatrix} & .72 & .61 & .92 & -.26 & .59 \\ .72 & & .80 & .68 & .20 & .68 \\ .61 & .80 & & .54 & .40 & .69 \\ .92 & .68 & .54 & & -.39 & .55 \\ -.26 & .20 & .40 & -.39 & & .21 \\ .59 & .68 & .69 & .55 & .21 & \end{pmatrix}$$

Using the true communalities, which are obtainable from off-diagonal elements of the above matrix, and the centroid method as the above, we get the loading matrix.

$$\begin{pmatrix} .7999 & .8944 & .8967 & .7538 & .2328 & .7607 \\ .4994 & .0095 & -.2141 & .6328 & -.8910 & .0369 \end{pmatrix}$$

which is gotten by the pre-multiplication of the orthogonal matrix

$$\begin{pmatrix} .4380 & .8990 \\ .8972 & -.4367 \end{pmatrix}$$ to the above true loading matrix.

Example 2. $p = 5$, $k = 2$ (improper solution)

The true loading matrix is the following:

$$\begin{pmatrix} 1.1 & 0.4 & 0.4 & 0.6 & 0.3 \\ 0.3 & -0.2 & 0.5 & -0.2 & 0.9 \end{pmatrix}$$

,and the corresponding correlation matrix is the following;

$$\begin{pmatrix} & .38 & .59 & .60 & .60 \\ .38 & & .06 & .28 & -.06 \\ .59 & .06 & & .14 & .57 \\ .60 & .28 & .14 & & .00 \\ .60 & -.06 & .57 & .00 & \end{pmatrix}$$

Since the above structure is improper and unique for $k \leqslant 2$, we have proper solutions for $k \geqslant 3$ which are not unique. However, in the formal sence, using the true communalities which are obtainable from off-diagonal elements of the above matrix, and by the centroid method as the above, we get the loading matrix

$$\begin{pmatrix} 1.12404 & .27858 & .57336 & .45998 & .65110 \\ .19112 & .34984 & -.28506 & .43407 & -.68997 \end{pmatrix}$$

which is gotten by the pre-multiplication of the orthogonal matrix

$$\begin{pmatrix} .907 & .42112 \\ .42112 & -.907 \end{pmatrix}$$ to the above true loading matrix.

Example 3. $p = 3$, $k = 1$ (improper solution)

Take the correlation matrix $\begin{pmatrix} & 3/16 & -\sqrt{6/16} \\ 3/16 & & \sqrt{6/16} \\ -\sqrt{6/16} & \sqrt{6/16} & \end{pmatrix}$. From off-diagonals

of this matrix, we get $h_1^2 = -3/16$, $h_2^2 = -3/16$, $h_3^2 = -2/16$ for $k = 1$, formally, and the corresponding centroid factor loadings are $(\sqrt{-3/4}, -\sqrt{-3/4}, \sqrt{-2/4})$, which is the true with $k = 1$. In this case also proper structures being not unique are given for $k \geqslant 2$.

Chapter 2. Case of unknown communalities

In the above section we treated the case of known true communalities, and in this section we consider the case when we know the true factor number k only, and consider problems given in the introduction using numerical examples. The theoretical treatment of these problems is very difficult. At first we consider the effect of sigh-change with the maximum row element as the initial estimate of communalities. We consider two cases;

(1) After each factor extraction, we use residual diagonal elements. We call this case the "unadjusted case".

(2) After each factor extraction, we use the maximum row elements in the residual matrix as diagonals. We call this case the "adjusted case".

We consider the following two numerical examples and the 15 types of sign-change for each.

(1) $p = 5$, $k = 2$ (proper)

The true loading matrix is the following:

$$\begin{pmatrix} .8 & .4 & .4 & .6 & .3 \\ .3 & -.2 & .5 & -.2 & .7 \end{pmatrix}$$

(2) $p = 5$, $k = 2$ (improper)

The true loading matrix is the following:

$$\begin{pmatrix} 1.1 & .4 & .4 & .6 & .3 \\ .3 & -.2 & .5 & -.2 & .9 \end{pmatrix}$$

For these two examples, the unadjusted case is inconsistent for getting the real loading for the 2nd factor. With the adjustment, on the contrary, we get the real loadings for the 2nd factor in most types of sign-change, and the effects of sign-changes are seen in both examples, in the first iteration cycle at least. The sign-change may not be arbitrary in the present situation, althogh we can not give the principle of determination of optimum sign-change. To make the centroid as distant from the origin as possible may be a rule. By taking maximum row elements as the initial estimate of communalities, we may have negative diagonals in the residual correlation matrix after each factor as in the above examples. To get rid of such an inconsistency, we consider two alternative iterative centroid method, one of them is the usual adjusted complete centroid method, another is the unadjusted complete centroid method.

For these two method, we also consider the effects of initial estimate of communalities. Note that, with the complete centroid method, we can not get the improper solution as in the example 3, p. 4, since we have real solutions only with this method. In the following, we describe the algorithm of the unadjusted complete centroid method (for the adjusted complete centroid method, see Thurstone(1), pp. 161).

Example 1. $p = 5$, $k = 2$ (proper solution)

The true loading matrix is given by

$$\begin{bmatrix} .5 & .3 & -.6 & .4 & -.8 \\ -.8 & .2 & .1 & .9 & .1 \end{bmatrix}$$

and the corresponding correlation matrix is the following:

$$\begin{bmatrix} & -.01 & -.38 & -.52 & -.48 \\ -.01 & & -.16 & .30 & -.22 \\ -.38 & -.16 & & -.15 & .49 \\ -.52 & -.30 & -.15 & & -.23 \\ -.48 & -.22 & .49 & -.23 & \end{bmatrix}$$

Using true communalities we get the following result with this method:

1) Calculate column sums of the above matrix; these are $W(1) = -1.39$, $W(2) = -.09$, $W(3) = -.20$, $W(4) = -.60$, $W(5) = -.44$

2) Let $V(i) = 1.0$ for all i, $J = 1$ and compare $W(i)V(i)$'s. $W(1) - W(2) < 0$, $W(1) - W(3) < 0$, $W(1) - W(4) < 0$, $W(1) - W(5) < 0$,

so that J remains to be 1, and $DW = W(1)V(1) = -1.39$, whence let $V(1) = -1$, other V's are 1.

3) $W(i) = W(i) + 2R(i, 1)V(1) = W(i) - 2R(i, 1)$, note that $R(1, 1) = 0$, so that $W(1)$ is unchanged.

We have results $W(1) = -1.39$, $W(2) = -.07$, $W(3) = .56$, $W(4) = .44$ and $W(5) = .52$. In this case, $W(1)V(1) - W(2)V(2) > 0$, then put $J = 2$, and $W(2)V(2) - W(3)V(3) < 0$, $W(2)V(2) - W(4)V(4) < 0$, and $W(2)V(2) - W(5)V(5) < 0$, so that $J=2$, and $DW = W(2)V(2) = -.07$, since DW is negative, let $V(2) = -1$, at this step we have $V(1) = -1$, $V(2) = -1$, $V(3) = 1$, $V(4) = 1$, $V(5) = 1$. With these, we calculate the new W's, $W(1) = W(1) + 2R(1, 2)V(2) = W(1) - 2R(1, 2)$, and so on.

We continue this until we get nonnegative DW. In our case, we have final result, $V(1) = -1$, $V(2) = -1$, $V(3) = 1$, $V(4) = -.1$, and $W(5) = 1$; $W(1) = -.33$, $W(2) = -.67$, $W(3) = 1.18$, $W(4) = -.16$, and $W(5) = 1.42$, and $DW = .16$.

4) To the above W's we add diagonal elements with sign of V's;

$W(1)=W(1)+(.89)V(1)=-1.22$, $W(2)=W(2)+(.13)V(2)=-.80$,
$W(3)=W(3)+(.37)V(3)=1.55$, $W(4)=W(4)+(.97)V(4)=-1.13$,
and $W(5)=W(5)+(.65)V(5)=2.07$. Let $T=\sum|W(i)|$. $(=6.77)$

5) We get the first loadings as $a_{i1}=W(i)/\sqrt{T}$.
$a_{11}=-.46888$, $a_{21}=-.30746$, $a_{31}=.59571$, $a_{41}=-.43429$,
and $a_{51}=.79557$

6) The (i, j) element of the residual matrix is given by
$\sigma_{ij}-a_{i1}a_{j1}$ $(1\leqslant i\leqslant 5,\ 1\leqslant j\leqslant 5)$, including diagonals.

6) Apply the above steps to residual matrix.

In this example, our estimate of loadings are the following;

$$\begin{pmatrix} -.46888, & -.30746, & .59571, & -.43429, & .79557 \\ -.81862, & .18832, & .12299, & .88396, & .13067 \end{pmatrix}.$$

Thus, we get the true factor loadings.

Example 2. $p=5$, $k=2$ (improper)

The true loading matrix is given by

$$\begin{pmatrix} 1.1 & 0.4 & 0.4 & 0.6 & 0.3 \\ 0.3 & -0.2 & 0.5 & -0.2 & 0.9 \end{pmatrix}.$$

With the above method we get the estimate of loadings as

$$\begin{pmatrix} 1.12404 & .27858 & .57336 & .45998 & .65110 \\ .19112 & .34984 & -.28506 & .43407 & -.68997 \end{pmatrix}$$

, which gives us the true factor loadings (cf. p. 4).

Now we mention a numerical result using these method. Further analysis using other numerical examples will be published in the other place. The true loadings and communalities are the followings:

$$\begin{pmatrix} .60 & .40 & .30 & .20 & .10 \\ .00 & -.20 & .30 & -.40 & -.50 \end{pmatrix}$$

and .36, .20, .18, .20, .26. This is the unique proper structure for $k\leqslant 2$. At first we mention results of the adjusted complete centroid method.

1) Initial estimate of communalities are the maximum row elements of given correlation matrix, that is, .24, .24, .18, .22, .22, respectively. In this case we get the stable result after 35 iterations; communalities are .4186438, .1960549, .1562591, .2222112, and .2419636, respectively. The corresponding loadings are the following:

$$\begin{pmatrix} .5663354 & .4425859 & .1202355 & .3681717 & .3013179 \\ .3129025 & -.0131418 & .3765667 & -.2943821 & -.3888073 \end{pmatrix}.$$

2) Initial estmate of communalities are the true communalities. In this case we get the stable result after 25 iterations ; communalities are .4185317, .1960650, .1562953, .2222034, .2419770, respectively.
The corresponding loadings are the following:

$$\begin{pmatrix} .5662625 & .4425982 & .1202626 & .3681705 & .3013310 \\ .3128554 & -.0131058 & .3766061 & -.2943703 & -.3888144 \end{pmatrix}.$$

Next we mention results of the unadjusted complete centroid method.

1) Initial estimate of communalities are 1.0 for all.
In this case we get the stable result after 43 iterations ; communalities are .3648799, .1998192, .1768116, .2000856, .2598534, respectively. The corresponding loadings are the following:

$$\begin{pmatrix} .5392621 & .4470113 & .1323519 & .3577379 & .3128964 \\ .2721695 & -.0003262 & .3991176 & -.2685315 & -.4024292 \end{pmatrix}.$$

2) Initial estimate of communalities are any larger numbers than the true communalities, for example, .6, .8, .8, .8, and .7, respectively. In this case we get the stable result after 52 iterations.
The communalities are .3599047, .2000133, .1800600, .2000845, and .2598724 respectively. The corresponding loadings are the following:

$$\begin{pmatrix} .5366077 & .4472285 & .1341970 & .3578390 & .3129587 \\ -.2682478 & .0000044 & -.4025558 & .2683947 & .4024043 \end{pmatrix}.$$

3) Initial estimate of communalities are the maximum row element.
In this case we get the stable result after 53 iterations. The communalities are .3599388, .2000070, .1800402, .2000397, .2599392, respectively.
The corresponding loadings are the following:

$$\begin{pmatrix} .5366249 & .4472214 & .1341867 & .3578031 & .3130064 \\ -.2682769 & .0000020 & -.4025347 & .2683592 & .4024503 \end{pmatrix}.$$

From these results we see that we get the true structure for the last two cases, at least. The unadjusted complete method will be the appropriate method. To put these conclusion to be definitive, however, we need further examination.

References

(1) L. L. THURSTONE ; MULTIPLE-FACTOR ANALYSIS, THE UNIVERSITY OF CHICAGO PRESS, 1947
(2) D. N. LWALEY, "A STATISTICAL EXAMINATION OF THE CENTROID METHOD," Proc. Roy. Soc. Edin. A64, 1955, pp. 175–189
(3) Y. TUMURA, K. FUKUTOMI, Y. ASOO, "ON THE UNIQUE CONVERGENCE OF ITERATIVE PROCEDURE IN FACTOR ANALYSIS", TRU Mathematics, vol. 4, 1968