Sankhyā : The Indian Journal of Statistics
1983, Volume 45, Series B, Pt. 3, pp. 447-465.

A NOTE ON TWO STAGES AGGREGATION OF VARIABLES THROUGH EQUAL WEIGHTING IN EVERY STAGE*

By R. N. DE

Indian Statistical Institute

SUMMARY. A two-stage method of aggregation of variables has been suggested here for controlling associations between the index to be constructed and the variables according to some *a priori* hypotheses which indicate the purpose of the index.

1. Introduction

In aggregating variables there are two systems—unequal weighting system (Kendall, 1939) and equal weighting system (Pal, 1971, 1974). When one decides to aggregate certain variables into an index, it is necessarily with some object in view. That is, one decides *a priori* what the index will indicate. An order among the representations of the variables into the index to be constructed is decided upon according to some *a priori* economic considerations or objectives. By representations of the variables into the index we mean the associations between the variables and the index. Representation values obtained by the two methods which do not fulfil the *a priori* objectives cannot be accepted. Even if the *a priori* order among the representations of the variables is maintained by any of the two methods, representation values relating to some variables for the unequal weighting system and to all variables for the equal weighting system would be low and thus unacceptable in case of a dispersed correlation matrix where there are very high as well as very low values of correlation coefficients.

So, when the two methods cannot yield results which satisfy the *a priori* objectives and the condition that the representation values are not low, we would propose a two-stage use of the equal weighting system by judicious choice of subgroups of the variables. In the first stage indices for the subgroups are derived using the system of equal weighting and in the second stage the indices are aggregated into a final index by the same method. Now

* The article was presented in the 4th World Congress of Econometric Society during August-September 1980 held in Aix-en-Provence, France. The work was done under the Research Fellowship of the Indian Statistical Institute.

one can evaluate the associations between the variables and the final index. The associations are the representations of the variables into the final index. The use of the unequal weighting system in the two stages of aggregation may not achieve the desired goal as the representations of the variables in each of the stages remain unequal. Thus, the equal weighting system can regulate the representations of the variables more effectively than the unequal weighting system.

We prove (i) some general results relating to the equal weighting systems, (ii) some general results relating to the two-stage use of the equal weighting system where we shall consider two subgroups. We shall further purpose—for certain types of correlation matrix—sufficient and necessary conditions for total representation by the two-stage method exceeding that by single stage use of the equal weighting system. We shall also give a sufficient condition in the case of the general correlation matrix. Lastly we shall illustrate the proposed method with an example.

2. Determination of results

Let x_i $(i = 1, \ldots, n)$ be $n\,(> 2)$ standardised variables with the associated correlation matrix $R_{n,n} = ((r_{ij}))_{n,n}, r_{ij} = 1$ for $i = j = 1, \ldots, n$ and $0 \leqslant r_{ij} < 1$ for $i \neq j = 1, \ldots, n$ where r_{ij} stands for the correlation coefficient between x_i and x_j. Replication of variable is not entertained for which $r_{ij} < 1$ for $i \neq j = 1, \ldots, n$.

Define

(i) $$R_{n,n} = \left[\begin{array}{c|c} R_{11_{k,k}} & R_{12_{k,n-k}} \\ \hline R_{21_{n-k,k}} & R_{22_{n-k,n-k}} \end{array}\right]_{n,n}$$

for some $1 \leqslant k < n$, where

$R_{11} = ((r_{ij}))_{k,k}$ for $1 \leqslant i, j \leqslant k$;

$R'_{21} = R_{12} = ((r_{ij}))_{k,n-k}$ for $1 \leqslant i \leqslant k$ and $k < j \leqslant n$;

and $R_{22} = ((r_{ij}))_{n-k,n-k}$ for $k < i, j \leqslant n$.

(ii) $Q_{n,n}$ = the matrix obtained by (a) deleting k-th row of $R_{n,n}$; (b) subtracting k-th row of $R_{n,n}$ from other rows of $R_{n,n}$; and (c) introducing an unit vector of order n as the first row of $Q_{n,n}$ (unit vector is defined by V'_n in $(v)) = ((q_{ij}))_{n,n}$

$$= \left[\begin{array}{c|c} Q_{11_{k,k}} & Q_{12_{k,n-k}} \\ \hline Q_{21_{n-k,k}} & Q_{22_{n-k,n-k}} \end{array}\right]_{n,n}$$

where

$$Q_{11} = ((q_{ij}))_{k,\,k} \qquad \text{for } 1 \leqslant i,\, j \leqslant k;$$

$$Q'_{21} = Q_{12} = ((q_{ij}))_{k,\,n-k}, \quad \text{for } i \leqslant j \leqslant k \text{ and } k < j \leqslant n;$$

and $$Q_{22} = ((q_{ij}))_{n-k,\,n-k}, \qquad \text{for } k < i,\, j \leqslant n.$$

(iii) $S_{n,n}$ = the matrix obtained by (a) deleting n-th row of $R_{n,n}$; (b) subtracting n-th row of $R_{n,n}$ from other rows of $R_{n,n}$; and (c) introducing an unit vector of order n as the $(k+1)$-th row of $S_{n,n}$

$$= ((s_{ij}))_{n,n}$$

$$= \left[\begin{array}{c|c} S_{11_{k,\,k}} & S_{12_{k,\,n-k}} \\ \hline S_{21_{n-k,\,k}} & S_{22_{n-k,\,n-k}} \end{array}\right]_{n,\,n}$$

where

$$S_{11} = ((s_{ij}))_{k,\,k} \qquad \text{for } 1 \leqslant i,\, j \leqslant k;$$

$$S'_{21} = S_{12} = ((s_{ij}))_{k,\,n-k} \quad \text{for } 1 \leqslant i \leqslant k \text{ and } k < j \leqslant n;$$

and $$S_{22} = ((s_{ij}))_{n-k,\,n-k} \qquad \text{for } k < i,\, j \leqslant n.$$

(iv) $R^T_{n,n}$ = the matrix obtained by subtracting k-th row of $R_{n,n}$ from other rows of $R_{n,n}$.

(v) $V'_n = (1, \ldots, 1)_{1,n}$

and

(vi) $e'_n(j) = (0, 0, \ldots, 1, \ldots, 0, 0)_{1,n}$ where unity occurs in j-th position.

Here two groups are considered—one with first k variables and the other with remaining $(n-k)$ variables. Let I, I_1, I_2 and I_3 be four indices dervied from n variables, first k variables, last $(n-k)$ variables and from I_1 and I_2 respectively using the equal weighting system. They are given by the following equations.

$$I = \sum_{i=1}^{n} \alpha_i x_i, \;\; \ni \sum_{i=1}^{n} \alpha_i = 1,$$

with $0 \leqslant \text{corr}(x_i, I) = \text{corr}(x_j, I) = r = s(I) < 1$ for $1 \leqslant i \neq j \leqslant n$ and

$$I_1 = \sum_{i=1}^{k} W_i x_i, \ni \sum_{i=1}^{k} W_i = 1,$$

with $0 \leqslant \text{corr}(x_i, I_1) = \text{corr}(x_j, I_1) = r_1 = s(I_1) < 1$ for $1 \leqslant i \neq j \leqslant k$

and

$$I_2 = \sum_{i=k+1}^{n} W_i x_i, \ni \sum_{i=k+1}^{n} W_i = 1,$$

with $0 \leqslant \mathrm{corr}(x_i, I_2) = \mathrm{corr}(x_j, I_2) = r_2 = s(I_2) < 1$ for $k < i \neq j \leqslant n$
lastly

$$I_3 = l_1 \frac{I_1}{r_1} + l_2 \frac{I_2}{r_2} \ni l_1 + l_2 = 1 \text{ and } l_1 = l_2 = \frac{1}{2}$$

with $r_3 = \mathrm{corr}(I_3, I_1) = \mathrm{corr}(I_3, I_2)$

$$= \sqrt{\frac{1}{2}\left(1 + \sum_{i=1}^{k} \sum_{j=k+1}^{n} W_i W_j r_{ij}/r_1 r_2\right)} = s(I_3) < 1$$

where

$$(\alpha_1, \ldots, \alpha_n) = e_n'(1).(Q^{-1})' = e_n'(k+1).(S^{-1})';$$

and

$$(W_1, \ldots, W_k) = e_k'(1).(Q_{11}^{-1})'; \text{ and } (W_{k+1}, \ldots, W_n) = e_{n-k}'(1).(S_{22}^{-1})';$$

and $s(I)$, $s(I_1)$, $s(I_2)$, $s(I_3)$ are standard deviations of I, I_1, I_2 and I_3 respectively.

Proposition 1 : $r^2 = \dfrac{|R|}{|Q|} = \dfrac{|R|}{|S|}$.

Proof : In the equal weighting system,

$$\alpha_1 r_{k1} + \alpha_2 r_{k2} + \ldots + \alpha_n r_{kn} = r^2 \text{ for any } 1 \leqslant k \leqslant n$$

i.e.,

$$r^2 = \frac{r_{k1}.\text{cofactor of } q_{11} \text{ in } Q}{|Q|} + \frac{r_{k2}.\text{cofactor of } q_{12} \text{ in } Q}{|Q|}$$

$$+ \ldots + \frac{r_{kn}.\text{cofactor in } q_{1n} \text{ in } Q}{|Q|}$$

$$= \frac{r_{k1}.\text{cofactor of } r_{k1} \text{ in } R^T}{|Q|} + \frac{r_{k2}.\text{cofactor of } r_{k2} \text{ in } R^T}{|Q|}$$

$$+ \ldots + \frac{r_{kn}.\text{cofactor of } r_{kn} \text{ in } R^T}{|Q|}$$

$$= \frac{|R^T|}{|Q|}$$

But note that $|R^T| = |R|$ as R^T is obtained by one elementary operation on R as defined. Therefore, $r^2 = \dfrac{|R|}{|Q|}$ and similarly $r^2 = \dfrac{|R|}{|S|}$. Thus $|Q| = |S|$.

Corollary 1 : *R being correlation matrix with $r_{ij} < 1$ for $1 \leqslant i \neq j \leqslant n$, $|R| > 0$, therefore $r^2 > 0$; r^2 tends to unity as $|Q|$ decreases to $|R| > 0$.*

Proposition 2 : $r_1 \geqslant r$, $r_2 \geqslant r$, $r_3 > r$ and $\operatorname{corr}(I_3, I) > r$.

Proof :

$$\operatorname{corr}(I_1, I) = \operatorname{cov}\left(\frac{I_1}{r_1}, \frac{I}{r}\right) = \frac{1}{r_1} \operatorname{corr}\left(\sum_1^k W_i x_i, \frac{I}{r}\right)$$

$$= \frac{1}{r_1} \sum_{i=1}^k W_i \operatorname{corr}(x_i, I) = \frac{1}{r_1} \sum_{i=1}^k W_i r$$

$$= \frac{r}{r_1} \text{ since } \sum_{i=1}^k W_i = 1.$$

Similarly, $\operatorname{corr}(I_2, I) = \dfrac{r}{r_2}$.

Since

$$\operatorname{corr}(I_1, I) \leqslant 1 \text{ and } \operatorname{corr}(I_2, I) \leqslant 1$$

therefore $r_1 \geqslant r$ and $r_2 \geqslant r$.

Now

$$\operatorname{corr}(I_3, I) = \operatorname{cov}\left(\frac{I_3}{r_3}, \frac{I}{r}\right) = \frac{1}{r_3} \operatorname{cov}\left(I_3, \frac{I}{r}\right)$$

$$= \frac{1}{r_3} \operatorname{cov}\left(\frac{l_1 I_1}{r_1} + \frac{l_2 I_2}{r_2}, \frac{I}{r}\right), \; l_1 + l_2 = 1$$

$$= \frac{1}{r_3}\left(l_1 \operatorname{corr}(I_1, I) + l_2 \operatorname{corr}(I_2, I)\right)$$

$$= \frac{r}{r_3}\left(\frac{l_1}{r_1} + \frac{l_2}{r_2}\right).$$

Since

$$r_1 < 1 \quad \text{and} \quad r_2 < 1$$

therefore

$$\frac{l_1}{r_1 r_3} > l_1 \text{ and } \frac{l_2}{r_2 r_3} > l_2$$

so,

$$\frac{l_1}{r_1 r_3} + \frac{l_2}{r_2 r_3} > l_1 + l_2 = 1,$$

$$\operatorname{corr}(I_3, I)/r = \frac{l_1}{r_1 r_3} + \frac{l_2}{r_2 r_3} > 1.$$

Therefore,

$$\operatorname{corr}(I_3, I) > r.$$

Again,

$$\frac{l_1}{r_1} + \frac{l_2}{r_2} > l_1 + l_2 = 1$$

i.e.,

$$\frac{r}{r_3}\left(\frac{l_1}{r_1} + \frac{l_2}{r_2}\right) > \frac{r}{r_3}$$

$$1 \geqslant \operatorname{corr}(I_3, I) = \frac{r}{r_3}\left(\frac{l_1}{r_1} + \frac{l_2}{r_2}\right) > \frac{r}{r_3}$$

so

$$r_3 > r.$$

Proofs are so designed that it holds good for more than two subgroups.

Proposition 3 : $|R+g.V_n V_n'| = |R|+g|Q_{11}|$ for any scalar g.

Proof :

$$\begin{aligned} |R+g.V_n V_n'| &= |R^T+g.(e_n(k), \ldots, e_n(k))| \text{ for any } k \\ &= |R^T|+g|Q_{11}| \\ &= |R|+g|Q_{11}|. \end{aligned}$$

Proposition 4 : If $R_{12} = R_{21}' = c.V_k.V_{n-k}'$ for $0 \leqslant c < 1$ then

$$\begin{aligned} \alpha_i &= W_i \frac{r_2^2 - c}{r_1^2 + r_2^2 - 2c}, \qquad \text{for } i \leqslant k \\ &= W_i \frac{r_1^2 - c}{r_1^2 + r_2^2 - 2c}, \qquad \text{for } k < i \leqslant n. \end{aligned}$$

Proof: For $i < k$,

$$\alpha_1 = \frac{|Q^{ii}|}{|Q|} = \frac{|Q_{11}^{ii}| \; |Q_{22} - Q_{21}^{\cdot i}(Q_{11}^{ii})^{-1}Q_{12}^{i\cdot}|}{|Q_{11}| \; |Q_{22} - Q_{21}(Q_{11})^{-1}Q_{12}|}$$

$$= W_i \frac{|Q_{22} - Q_{21}^{\cdot i}(Q_{11}^{ii})^{-1}Q_{12}^{i\cdot}|}{|Q_{22} - Q_{21}(Q_{11})^{-1}Q_{12}|}$$

where Q^{ij} denotes the matrix obtained by deleting i-th row and j-th column of the matrix Q; and $Q^{\cdot j}$ denotes the matrix obtained by deleting j-th column of the matrix Q and $Q^{i\cdot}$ denotes the matrix obtained by deleting i-th row of the matrix Q.

If $R_{12} = c.V_k V'_{n-k}$ then

(i) $Q_{12} = e_k(1).V'_{n-k}$;

(ii) $Q_{12}^{1\cdot} = ((0))_{k-1,\, n-k}$;

(iii) $Q_{21} = V_{n-k}.(c.V'_k - R'_{11}(k))$, $R'_{11}(k)$ being k-th row of R_{11};

and

(iv) $Q_{22} = R_{22} - c.V_{n-k}V'_{n-k}$,

therefore

$$|Q_{22}| = |R_{22} - c.V_{n-k}\, V'_{n-k}|$$

$$= |R_{22}| - c|S_{22}| \text{ [due to Proposition 3]}$$

$$= |S_{22}|\left(\frac{|R_{22}|}{|S_{22}|} - c\right)$$

$$= |S_{22}|(r_2^2 - c) \text{ [due to Proposition 1]}.$$

Therefore,

$$Q_{21}(Q_{11})^{-1}Q_{12} = Q_{21}(Q_{11})^{-1}e_k(1)V'_{n-k}$$

$$= Q_{21}U_k V'_{n-k}, \text{ where } U_k = (W_1, \ldots, W_k)' = (Q_{11})^{-1}e_k(1)$$

[according to the equal weighting system]

$$= V_{n-k}(cV'_k - R'_{11}(k))U_k V'_{n-k}$$

$$= V_{n-k}(cV'_k U_k - R'_{11}(k)U_k)V'_{n-k} \text{ since } V'_k U_k = 1$$

$$= V_{n-k}(c - r_1^2)V'_{n-k}, \text{ since } R'_{11}(k)U_k = r_1^2$$

[according to the equal weighting system]

$$= (c - r_1^2)V_{n-k}V'_{n-k}.$$

And finally,

$$|Q_{22}-Q_{21}(Q_{11})^{-1}Q_{12}| = |R_{22}-cV_{n-k}V'_{n-k}-(c-r_1^2)V_{n-k}V'_{n-k}|$$

$$= |R_{22}+(r_1^2-2c)V_{n-k}V'_{n-k}|$$

$$= |R_{22}|+(r_1^2-2c)|S_{22}| \text{ [due to Proposition 3]}$$

$$= |S_{22}|\left(\frac{|R_{22}|}{|S_{22}|}+(r_1^2-2c)\right)$$

$$= |S_{22}|(r_1^2+r_2^2-2c) \text{ [due to Proposition 1].}$$

For $i \leqslant k$,

$$\alpha_i = W_i \frac{|Q_{22}-Q_{21}^{\cdot i}(Q_{11}^{ii})^{-1}Q_{12}^{i\cdot}|}{|Q_{22}-Q_{21}(Q_{11})^{-1}Q_{12}|}$$

$$= W_i \frac{|Q_{22}|}{|Q_{22}-Q_{21}(Q_{11})^{-1}Q_{12}|} \quad \text{sinoe } Q_{12}^{i\cdot} = ((0))_{k-1,\,n-k}$$

$$= W_i \frac{|S_{22}|(r_2^2-c)}{|S_{22}|(r_1^2+r_2^2-2c)} = W_i \frac{(r_2^2-c)}{(r_1^2+r_2^2-2c)}.$$

Similarly, for $k < i \leqslant n$,

$$\alpha_i = W_i \frac{(r_1^2-c)}{(r_1^2+r_2^2-2c)}.$$

Corollary 2 : $(r_1^2+r_2^2-2c) > 0$, *where* $R_{12} = R'_{21} = cV_kV'_{n-k}$, $0 \leqslant c < 1$

Proof : From Proposition 1 and Corollary 1,

$$|Q| > 0 \qquad \text{for all } r < 1$$

$$|Q_{11}| > 0 \qquad \text{for all } r_1 < 1$$

and

$$|S_{22}| > 0 \qquad \text{for all } r_2 < 1.$$

$$|Q| = |Q_{11}|\,|Q_{22}-Q_{21}(Q_{11})^{-1}Q_{12}| > 0 \text{ since } r < 1.$$

Therefore

$$|Q_{22}-Q_{21}(Q_{11})^{-1}Q_{12}| > 0 \quad \text{since } |Q_{11}| > 0$$

i.e.

$$|S_{22}|(r_1^2+r_2^2-2c) > 0 \text{ [due to Proposition 4]}$$

so

$$(r_1^2+r_2^2-2c) > 0 \quad \text{since } |S_{22}| > 0.$$

Corollary 3 ·

$$r^2 = \frac{r_1^2r_2^2-c^2}{(r_1^2+r_2^2-2c)} \quad \textit{for } R_{12} = R'_{21} = c.V_kV'_{n-k},\ 0 \leqslant c < 1.$$

Proof: Note that

$$r^2 = \alpha_1 + r_{12}\alpha_2 + \ldots + r_{1k}\alpha_k + c(\alpha_{k+1} + \ldots + \alpha_n)$$

$$= \frac{(r_2^2 - c)}{(r_1^2 + r_2^2 - 2c)} \cdot [W_1 + r_{12}W_2 + \ldots + r_{1k}W_k]$$

$$+ \frac{c(r_1^2 - c)}{(r_1^2 + r_2^2 - 2c)} [W_{k+1} + \ldots + W_n]$$

$$= \frac{r_1^2(r_2^2 - c)}{(r_1^2 + r_2^2 - 2c)} + \frac{c(r_1^2 - c)}{(r_1^2 + r_2^2 - 2c)};$$

since $\sum_{i=1}^{k} W_i r_{1i} = r_1^2$ and $\sum_{k+1}^{n} W_i = 1$

$$= \frac{r_1^2 r_2^2 - c}{r_1^2 + r_2^2 - 2c}.$$

Corollary 4: *$r^2 \geqslant 0$ if and only if $c \leqslant r_1 r_2$, where $R_{12} = R'_{21} = cV_k V'_{n-k}$, $0 \leqslant c < 1$ and $r < 1$.*

Proof: From Corollary 3,

$$r^2 = \frac{r_1^2 r_2^2 - c}{r_1^2 + r_2^2 - 2c} = \frac{(r_1 r_2 - c)(r_1 r_2 + c)}{(r_1^2 + r_2^2 - 2c)}$$

since $(r_1^2 + r_2^2 - 2c) > 0$ and $(r_1 r_2 + c) > 0$, $r^2 \geqslant 0$ if and only if $c \leqslant r_1 r_2$.

Proposition 5: $\sum_{i=1}^{n} (\text{corr}^2(X_i, I_3) - r^2) > 0$ if and only if either $r_1^2 > r_2^2$ and

$$(r_1^2 - (n-k)r_2^2) > \frac{2c}{(r_1 + r_2)} (kr_1 - (n-k)r_2)$$

or $r_2^2 > r_1^2$ and

$$(kr_1^2 - (n-k)r_2^2) < \frac{2c}{(r_1 + r_2)} \cdot (kr_1 - (n-k)r_2)$$

where $R_{12} = R'_{21} = cV_k V'_{n-k}$, $0 \leqslant c < 1$.

Proof: Note that

$$r_3^2 = \frac{1}{2}\left(\frac{\sum\limits_{i=1}^{k}\sum\limits_{j=k+1}^{n} W_iW_jr_{ij}}{r_1r_2}\right) = \frac{1}{2}\left(1+\frac{c}{r_1r_2}\right) > 0$$

since $r_{ij} = c \;\forall\; i = 1, \ldots, k;\ j = k+1, \ldots, n,$

$$= \frac{1}{2}\,\frac{r_1r_2+c}{r_1r_2}.$$

Again,

$$\operatorname{corr}(x_i, I_3) = \frac{r_1l_1}{r_3}+\frac{l_2}{r_3}\operatorname{corr}(x_i, I_2) \quad \text{for } i \leqslant k$$

$$= \frac{1}{2r_3}\left(r_1+\frac{c}{r_2}\right), \text{ since } \operatorname{corr}(x_i, I_2) = \sum_{j=k+1}^{n} r_{ij}W_j = c$$

$$= \frac{1}{2r_3}\left(\frac{r_1r_2+c}{r_2}\right) = r_1r_3.$$

Similarly for $k < i \leqslant n$,

$$\operatorname{corr}(x_i, I_3) = r_2r_3.$$

Therefore,

$$\sum_{i=1}^{n}\left(\operatorname{corr}^2(x_i, I_3)-r^2\right)$$

$$= \sum_{i=1}^{k}\left(r_1^2r_3^2-\frac{r_1^2r_2^2-c^2}{r_1^2+r_2^2-2c}\right)+\sum_{i=k+1}^{n}\left(r_2^2r_3^2-\frac{r_1^2r_2^2-c^2}{r_1^2+r_2^2-2c}\right)$$

$$= kr_3^2\left(r_1^2-\frac{2r_1r_2(r_1r_2-c)}{r_1^2+r_2^2-2c}\right)+(n-k)r_3^2\left(r_2^2-\frac{2r_1r_2(r_1r_2-c)}{r_1^2+r_2^2-2c}\right)$$

$$= r_3^2\,\frac{kr_1(r_1-r_2)\{r_1(r_1+r_2)-2c\}+(n-k)r_2(r_2-r_1)\{r_2(r_1+r_2)-2c\}}{(r_1^2+r_2^2-2c)}$$

$$= \frac{r_3^2(r_1^2-r_2^2)}{r_1^2+r_2^2-2c}\left\{(kr_1^2-(n-k)r_2^2)-\frac{2c}{(r_1+r_2)}(kr_1-(n-k)r_2)\right\}$$

$(r_1^2+r_2^2-2c) > 0$ when $r_2 < 1$ from Corollary 2; and $r_3^2 > 0$.

So,

$$\sum_{i=1}^{n} (\text{corr}^2(x_i, I_3)-r^2) > 0$$

if and only if either

$$r_1^2 > r_2^2 \text{ and } (kr_1^2-(n-k)r_2^2) > \frac{2c}{(r_1+r_2)}\,(kr_1-(n-k)r_2)$$

or

$$r_1^2 < r_2^2 \text{ and } (kr_1^2-(n-k)r_2^2) < \frac{2c}{(r_1+r_2)}\,(kr_1-(n-k)r_2).$$

Corollary 5 : $\sum_{i=1}^{n} (corr^2(x_i, I_3)-r^2) > 0$ *if and only if*

$$\frac{r_1^2}{r_2^2} \notin \left(1, \frac{n-k}{n}\right) \textit{ or } \left(\frac{n-k}{k}, 1\right)$$

whichever is applicable, where $R_{12} = R'_{21} = ((0))_{k,\, n-k}$.

Proof : $R_{12} = R'_{21} = ((0))_{k,\, n-k} \Longleftrightarrow c = 0$. Therefore,

$$\sum_{i=1}^{n} (\text{corr}^2(x_i, I_3)-r^2) > 0$$

if and only if either

$$\frac{r_1^2}{r_2^2} > 1 \quad \text{and} \quad \frac{r_1^2}{r_2^2} > \frac{n-k}{k} \text{ [due to Proposition 5]}$$

i.e., either

$$\frac{r_1^2}{r_2^2} > \max\left(1, \frac{n-k}{k}\right)$$

or

$$\frac{r_1^2}{r_2^2} < 1 \quad \text{and} \quad \frac{r_1^2}{r_2^2} < \frac{n-k}{k}$$

or

$$\frac{r_1^2}{r_2^2} < \min\left(1, \frac{n-k}{k}\right)$$

combining two results,

$$\sum_{i=1}^{n} (\text{corr}^2(x_i, I_3)-r^2) > 0$$

iff

$$\frac{r_1^2}{r_2^2} \notin \left(1, \frac{n-k}{k}\right) \quad \text{or} \quad \left(\frac{n-k}{k}, 1\right)$$

whichever is applicable.

Corollary 6 : $\sum_{i=1}^{n} (corr^2(x_i, I_3)-r^2) > 0$ *if and only if*

$$\frac{r_1^2}{r_2^2} \notin \left(1, \left(\frac{\sqrt{2c}}{r_2}-1\right)^2\right) \quad or \quad \left(\left(\frac{\sqrt{2c}}{r_2}-1\right)^2, 1\right)$$

where $R_{12} = R'_{21} = cV_kV'_{n-k}$ *and* $k = n-k = n/2,\ 0 \leqslant c < 1.$

Proof : From Proposition 5, we have, if $k = n-k = n/2$,

$$\sum_{i=1}^{n} (\text{corr}^2(x_i, I_3)-r^2) > 0$$

iff either

$$\frac{r_1^2}{r_2^2} > 1 \text{ and } (r_1+r_2)^2 > 2c$$

or

$$\frac{r_1^2}{r_2^2} < 1 \text{ and } (r_1+r_2)^2 < 2c.$$

Now,

$$(r_1+r_2)^2 \gtrless 2c \text{ iff } \frac{r_1^2}{r_2^2} \gtrless \left(\frac{\sqrt{2c}}{r_2}-1\right)^2$$

since $0 \leqslant c < 1$ and $0 \leqslant r_2$. Therefore

$$\frac{r_1^2}{r_2^2} > 1 \text{ and } (r_1+r_2)^2 > 2c \text{ iff } \frac{r_1^2}{r_2^2} > \max\left(1, \left(\frac{\sqrt{2c}}{r_2}-1\right)^2\right)$$

and

$$\frac{r_1^2}{r_2^2} < 1 \text{ and } (r_1+r_2)^2 < 2c \text{ iff } \frac{r_1^2}{r_2^2} < \min\left(1, \left(\frac{\sqrt{2c}}{r_2}-1\right)^2\right)$$

Together says that

$$\sum_{i=1}^{n} (\text{corr}(x_i, I_3)^2-r^2) > 0 \text{ iff } \frac{r_1^2}{r_2^2} \notin \left(1, \left(\frac{\sqrt{2c}}{r_2}-1\right)^2\right),$$

or

$$\left(\left(\frac{\sqrt{2c}}{r_2}-1\right)^2, 1\right)$$

whichever is applicable.

Corollary 7 :

$$\sum_{i=1}^{n} (corr(x_i, I_3)^2 - r^2) > 0 \ \ iff \ \ r_1^2 \neq r_2^2$$

when $R_{12} = R_{21}' = ((0))_{k,\, n-k}$.

Proof : Putting $c = 0$ in Corollary 6, proof becomes obvious.

Proposition 6 : $r^2 > 0$ attains maximum value under

$$R_{12} = R_{21}' = c V_k V_{n-k}'$$

iff $$c = \min(r_1^2, r_2^2),\ r_1^2 \neq r_2^2.$$

Proof : From Corollary 3,

$$r_2 = \frac{r_1^2 r_2^2 - c^2}{r_1^2 + r_2^2 - 2c}$$

$$\frac{dr^2}{dc} = \frac{-2c(r_1^2 + r_2^2 - 2c) + 2(r_1^2 r_2^2 - c^2)}{(r_1^2 + r_2^2 - 2c)^2}$$

$$\frac{dr^3}{dc} = 0 \Longleftrightarrow c^2 - c(r_1^2 + r_2^2) + r_1^2 r_2^2 = 0 \Longleftrightarrow c = r_1^2 \text{ and/or } c = r_2^2.$$

From Corollary 4, $c < r_1 r_2$ for $r > 0$ ($r = 0$ iff $c = r_1^2 = r_2^2$).

Therefore if $c = r_1^2$ then $r_1 < r_2$ and if $c = r_2^2$ then $r_2 < r_1$, so

$$c = \min(r_1^2, r_2^2).$$

Now,

$$\frac{d^2r^2}{dc^2} = \frac{2(2c - (r_1^2 + r_2^2))(r_1^2 + r_2^2 - 2c)^2 + 4(r_1^2 + r_2^2 - 2c)(c^2 - c(r_1^2 + r_2^2) + r_1^2 r_2^2)}{(r_1^2 + r_2^2 - 2c)^4}.$$

Without loss of generality we may assume $r_1 < r_2$ then $c = r_1^2$ at which

$$\frac{dr^2}{dc} = 0$$

$$\left.\frac{d^2r^2}{dc^2}\right|_{c=r_1^2} = \frac{2(2r_1^2 - (r_1^2 + r_2^2))(r_1^2 + r_2^2 - 2r_1^2)^2}{(r_1^2 + r_2^2 - 2r_1^2)^4}$$

$$= \frac{2(r_1^2 - r_2^2)(r_2^2 - r_1^2)^2}{(r_2^2 - r_1^2)^4} = \frac{-2(r_2^2 - r_1^2)^3}{(r_2^2 - r_1^2)^4} = \frac{-2}{(r_2^2 - r_1^2)} < 0.$$

Proposition 7 : A sufficient condition for $\sum_{i=1}^{n} (\text{corr}(x_i, I_3)-r) > 0$ is

$$\frac{1}{n}\left[kr_1+(n-k)r_2+\sum_{i=1}^{k}\text{corr}(x_i, I_2)+\sum_{i=k+1}^{n}\text{corr}(x_i, I_1)\right]-1 > \text{corr}(I_1, I_2),$$

for any correlation matrix R with $0 \leqslant r_{ij} < 1$ for $1 \leqslant i \neq j \leqslant n$.

Proof :

$$\text{corr}(x_i, I_3) = \frac{1}{2r_3}\{r_1+\text{corr}(x_i, I_2)] \quad \text{for } i \leqslant k$$

$$= \frac{1}{2r_3}\{r_2+\text{corr}(x_i, I_1)\} \quad \text{for } i > k.$$

Therefore,

$$\sum_{i=1}^{n}\text{corr}(x_i, I_3)-nr$$

$$= \frac{1}{2r_3}\left[kr_1+\sum_{i=1}^{k}\text{corr}(x_i, I_2)+(n-k)r_2+\sum_{i=k+1}^{n}\text{corr}(x_i, I_1)\right]-nr$$

$$> \frac{1}{2r_3}\left[kr_1+\sum_{i=1}^{k}\text{corr}(x_i, I_3)+(n-k)r_2+\sum_{k+1}^{n}\text{corr}(x_i, I_1)-2nr_3^2\right]$$

(due to Proposition 2)

$$> 0$$

if

$$\left(kr_1+\sum_{i=1}^{k}\text{corr}(x_i, I_2)+(n-k)r_2+\sum_{k+1}^{n}\text{corr}(x_i, I_1)-2nr_3^2\right) > 0$$

i.e.,

$$\frac{1}{n}\left(kr_1+(n-k)r_2+\sum_{i=1}^{k}\text{corr}(x_i, I_2)+\sum_{i=k+1}^{n}\text{corr}(x_i, I_1)\right)-1 > \text{corr}(I_1, I_2)$$

since

$$2r_3^2 = 1+\text{corr}(I_1, I_2).$$

The objective of the proposition is to state that the family of correlation matrix satisfying the sufficient condition yields better results in two stages of aggregation.

3. Illustration

We want to estimate an index of industrial activity for the districts in two time points 1960-61 and 1970-71. There are 334 districts in India, We have chosen six variables denoting the industrial activity. They are given by X_1 to X_6 as follows.

X_1 : Density of labour in secondary sector per sq. km.

X_2 : Share of labour in secondary sector per thousand of total labour.

X_3 : Share of labour in large factories per thousand of labour in secondary sector.

X_4 : Average size of labour in large factory in thousands.

X_5 : Density of labour in large factories per thousand sq. km, and

X_6 : Density of large factories per thousand sq. km.

Here large factories are those which employ at least 100 labour. The data relating to labour in large factories and number of factories in districts have been collected from the Annual Survey of Industries (unpublished). The labour in secondary sector has been compiled from the Census of India. Since the total labour data are not comparable between 1960-61 and 1970-71, comparable estimates have been obtained from the work done by Pal, De and Malakar (1978).

The variables X_1 to X_6 have been transformed to near-normal distributions so as to reduce the skewness of their original distributions. All variables under consideration have been found to be lognormally distributed with different parameters (De, 1981). The transformed forms of X_i's, denoted by Y_i's, are given below :

$$Y_i = \ln(X_i+1) \text{ for } i = 1, 2, 3;\ Y_4 = \ln(10X_4+1);$$

$$Y_5 = \ln(X_5+1) \text{ and } Y_6 = \ln(10X_6+1).$$

The correlation matrix based on 668 observations among Y_i's thus obtained is given in Table 1.

TABLE 1. CORRELATION MATRIX BETWEEN VARIABLES Y_1 TO Y_6

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
Y_1	1.00000	0.73028	0.45327	0.39591	0.78317	0.84132
Y_2		1.00000	0.38263	0.33585	0.60268	0.63520
Y_3			1.00000	0.87500	0.87409	0.70147
Y_4				1.00000	0.77084	0.54094
Y_5					1.00000	0.94346
Y_6						1.00000

The value of r, the corrolation coefficient between I, as defined in the preceeding sections, and Y_i ($i = 1, \ldots, 6$) is found to be only 0·61294 which is low, although significant. We have certain a priori hypotheses regarding the index of industrial activity to be constructed following the method of two-stage aggregation. We have formulated three hypotheses given below.

(i) The solution should be compatible with the notion the variables relating to large factories would be highly correlated with the index to be constructed. Variables on large factories are given more emphasis as they indicate a country's potentiality in industrial development.

(ii) The difference between the maximum and the minimum correlation coefficients among the correlation coefficients (Y_i, I_3) ($i = 1, \ldots, 6$) should preferably be small so as to minimize inclination towards any particular variable.

(iii) $D = \sum_{i=1}^{6} (\text{corr}^2(Y_i,\ I_3) - 0{\cdot}61294^2)/6 = \sum_{i=1}^{6} (\text{corr}^2(Y_i,\ I_3) - 0{\cdot}37570)/6$ is positive and high. We cannot accept I as the index of industrial activity since all the variables are given an equal importance violating our first hypothesis.

In order to obtain an I_3 compatible with our *a priori* hypotheses, we obtain two possible I_3's either of which may be taken as the index of industrial activity. They are obtained from the two groupings (a) (Y_1, Y_2, Y_3, Y_4) and (Y_5, Y_6), and (b) (Y_1, Y_2, Y_6) and (Y_3, Y_4, Y_5). Table 2 discusses their detail structures. In addition, the table also shows the extent of validity of the *a priori* hypotheses. Both of the groupings satisfy the first hypothesis that large industrial activity should be given more importance. The order of representation among the variables remains same under both groupings. The variable—density of labour in large factories per sq. km. which indicates the degree of areal or geographical concentration of large industrial activity has got the highest degree of representation. The next important variable is density of number of large factories per thousand sq. km. It has the second highest degree of representation in both cases. The variable share of labour in large factories per thousand of labour in secondary activity which indicates the importance of industrial activity over other activities in secondary activity has been retained in the third position in the order of representation. Since an index of industrial activity should not only include the activity of large factory but also the other industrial activity, density of labour in secondary activity should have an adequate role. We find the variable

TABLE 2. SHOWING THE BEHAVIOUR OF TWO GROUPINGS

Groupings	subgroup 1 $W(I_1)$	r_1	subgroup 2 $W(I_2)$	r_2	r_3	$\text{corr}(I_3, Y_i)$ $i = 1, \ldots, 6$	D	maxima-minima	interpretational advantage
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Grouping 1	(0.21498, 0.32269 0.12146, 0.34087) corresponding to (Y_1, Y_2, Y_3, Y_4)	0.80040	(0.5, 0.5) corresponding to (Y_5, Y_6)	0.98576	0.97049	0.83688, 0.73585 0.83978, 0.75517 0.97890, 0.93445	0.3491	0.24305	variables on large industries are emphasised. It is indeed necessary for measuring the country's industrial potential. Index from grouping 2 is more preferable to the index from grouping 1.
Grouping 2	(0.13881, 0.45172 0.40947) corresponding to (Y_1, Y_2, Y_6)	0.90177	(0.07976, 0.45996, 0.46028) corresponding to (Y_3, Y_4, Y_5)	0.94051	0.92198	0.82274, 0.75506, 0.83934, 0.76754 0.97147, 0.91792	0.3456	0.21581	

concerned occupies the fourth position in the order of representation. Next two positions—fifth and sixth are successively held by average size of labour in large factory and share of labour in secondary sector per thousand of total labour. The order of representation is quite justifiable as it satisfies the first hypothesis.

Among the two groupings, the index derived from the second grouping is more acceptable to us than the index derived from the first grouping as the difference between the maximum and minimum correlation coefficients (hypothesis ii) is less for the second grouping. It is also to be noted if the unequal weighting system (Kendall, 1939) were operated on all variables directly, the resulting index—call it I_k—in sipte of being the best possible index in the sense that it explains maximum variance, may not be acceptable as the difference between maxima and minima among correlation coefficients (I_k, Y_i), $(i = 1, \ldots, 6)$, i.e., (0·82494, 0·70817, 0·86074, 0·76896, 0·98326, 0·93615) is 0·27509 whereas for grouping 1 and grouping 2 as given in Table 2, it turns out to be 0·24305 and 0·21581 respectively.

Acknowledgement. The author expresses his indebtedness to Dr. P. Bhimsankaran who spared his valuable time to go through the article and gave certain important suggestions. If any error appears, it is absolutely due to the author. The author is also grateful to Dr. D. K. Rengnekar for his encouragement.

REFERENCES

BERRY, B. J. L. (1960) : An inductive approach to the regionalization of economic development, in *Essays on Geography and Economic Development*, N. Ginsburg (ed.). The University of Chicago Press, Chicago.

DE, R. N. (1981) : Economic regionalization of India 1960-61 and 1970-71 : A study in quantitative methods. Doctoral Dissertation, (unpublished), Indian Statistical Institute, Calcutta.

GIRSHICK, M. A. (1936) : Principal components. *Jour. Amer. Statist. Assoc.*, **31**.

HOLZINGER, K. and HARMON, H. H. (1941) : *Factor Analysis : A Synthesis of Factorial Methods*, The University of Chicago Press, Chicago.

HOTELLING, H. (1933) : Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**.

KENDALL, M. G. (1939) : The geographical distribution of crop productivity in England. *J. Roy. Statist. Soc.*, **102**.

PAL, M. N. (1971) : Quantitative techniques for regional planning. . *Indian Journal of Regional Science*, **3**.

——— (1974) : Regional information, regional statistics and regional planning in India, in *Regional Information and Regional Planning*, A. Kuklinski (ed.), Vol. 6, UNRISD Publication Series on Regional Planning, Mouton, The Hague.

PAL, M. N., DE, R. N. and MALAKAR, B. (1978): A statistical revision of 1961 agricultural workers for the comparability with 1971 estimates by Indian districts. *Indian Journal of Regional Science*, **10**.

RHODES, E. C. (1936a): The precision of index numbers. *J. Roy. Statist. Soc.*, **99**.

——— (1936b): The precision of index numbers-II. *J. Roy. Statist. Soc.*, **99**.

——— (1937): The construction of an index of business activity. *J. Roy. Statist. Soc.*, **100**.

THURSTONE, L. L. (1931): Multiple factor analysis. *Psychological Review*, **38**.

WILKS, S. S. (1938): Weighting system for linear functions of correlated variables when there is no independent variables. *Psychometrika*, **3**.

Paper received : *December*, 1979.

Revised : *February*, 1983.