253

WP : 253

Working Paper

FULL INFORMATION BASED COMPOSITE INDICES – A BETTER ALTERNATIVE TO PRINCIPAL COMPONENTS

by

P.N. Misra

विद्याविनियोगाद्विकासः
IIM
AHMEDABAD

INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD

FULL INFORMATION BASED COMPOSITE INDICES - A BETTER ALTERNATIVE TO PRINCIPAL COMPONENTS

by

P.N. Misra

W P No. 253
Nov.1978

INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD

ABSTRACT

The problem of constructing composite indices has most often been tackled by using principal components in several disciplines. The approach, however, has some vital weaknesses.The paper suggests a method of constructing composite indices based upon full information contained in the data set. The method is also free from major defects of the method of principal components. The proposed method is amenable to simple statistical tests and provides a natural extension of the concept of centroid to statistically dependent constituent variables. The method is definitely a better substitute of the methods of principal components and factor analysis.

•••••

FULL INFORMATION BASED COMPOSITE INDICES - A BETTER ALTERNATIVE TO PRINCIPAL COMPONENTS

by

P.N. Misra

1. Introduction:

Sederal attempts have been made to build composite indices fro a given number of variable that are supposed to constitute a composit variate. Uses of such indices are often made in formation of homogene groups, segmentation and regional determination. Major statistical methods used for building composite indices are those of principal components and factor analysis (Harman, 1967). Harman (1971) has also suggested use of distance measure of major factors that are determined by minimising residuals to classify population or sample units into groups consisting of most like units. As the name itself implies the method of principal components or the associated factors involve use of several components that carry different levels of information. Further, one can obtain as many set of indices as the number of components associated with a given set of data. In other words, the indices are not unique.

A common practice is to pick up the component associated with maximum variation in the data. In that case the degree of the fit is shown by Pearson (1901) to be minimum. Obviously, a method providing best fit and associated with full variation in the data at hand is the one to which one looks for. The purpose of this paper is to provide a method that fills up the vaccum in the above sense.

Section two of this paper provides a brief survey of results relating to principal components. This is required to describe the relative advantages of the full information based composite indices to be discussed in Section 3. The last section contains discussion of the properties of the proposed method.

2. Indices Based upon Principal Components

Let us consider that K variables $x_1, \ldots\ldots, x_K$ are direct quantifiable and represent K manifest characteristics of a composit character. The composite character may represent economic developme industrial or agricultural potential, personality of individuals, resource potential of different regions and any other similar concept belonging to any discipline. Further, let us assume that data on n units are available relating to all the K variables so that the same can be expressed by $x_{1i}, \ldots\ldots, x_{Ki}$ for the i-th unit ($i = 1, \ldots, n$). Apriori reasoning can be used to define composite index for the compos character for the i-th unit as

(2.1) $$z_i = \alpha_1 x_{1i} + \ldots\ldots + \alpha_K x_{Ki}$$
$$i = 1, \ldots\ldots, n$$

The problem now boils down to determination of K weights of the variables on the right hand side of relation (2.1). As one may not be sure of relative importance of the individual x variable in the z indi the weights need to be determined objectively from the given data on the variables. Any apriori fixation of weights, for example, all α's being equal, will amount to loading the resulting indices by an assumption that could in fact be wrong.

The method of principal components provides one way of determination of the weights. McDonald (1968) provides a survey of other related weighting problems of almost similar characteristics. However, we will restrict ourselves in what follows to details of the method of principal components only. According to this approach, the weights are determined by considering the following restricted optimisation problem:

(2.2) $$\text{Optimise} \quad \alpha' X' X \alpha = z' z$$
$$\text{subject to} \quad \alpha' \alpha = 1$$

The matrix X represents n x K observations on n units with K variables relating to each unit and the n x 1 vector z represents vector of indices corresponding to n units. The constraint $\alpha'\alpha = 1$ provides a normalising rule for the weights because otherwise in the absence of such a restriction, one can optimise z'z by any arbitrary choice of α-vector.

Optimum solution is provided by the characteristic equation

$$(2.3) \qquad X'X\alpha = \lambda\alpha$$

where λ can assume as many values as the number of nonzero characteristic roots of the matrix X'X and α is the characteristic vector corresponding to each characteristic root. Supposing that $n > K$ and the columns of matrix X are linearly independent, the matrix X'X has K characteristic roots, say $\lambda_1, \ldots\ldots \lambda_K$, and K characteristic vectors $\alpha_1, \ldots\ldots \alpha_K$, respectively. We may compute composite indices or principal components by using estimated vector α_k in relation (2.1) as follows

$$(2.4) \qquad z_j = X\alpha_j \qquad j = 1, \ldots\ldots\ldots, K$$

All these principal components can be represented together as

$$(2.5) \qquad Z = XA$$

where

$$(2.6) \qquad \text{(i)} \quad Z = (z_1, \ldots\ldots\ldots, z_K)$$

$$\text{(ii)} \quad A = (\alpha_1, \ldots\ldots\ldots\ldots, \alpha_K)$$

It is well known (Johnston, 1972) that the principal components satisfy the following relations

(2.7) (i) $z'_j z_j = \lambda_j$; $j = 1, \ldots\ldots, K$

(ii) $\alpha_j{}'\alpha_j = 1$; $j = 1, \ldots\ldots, K$

(iii) $\alpha'_j \alpha_{j'} = 0$; $j \neq j' = 1, \ldots\ldots, K$

(iv) $Z'Z = A'X'XA = \Lambda$; Λ being diagonal and containing λ_i at the i-th place.

(v) $\mathrm{Tr}\,(X'X) = \mathrm{Tr}\,A'X'XA = \mathrm{Tr}\,\Lambda$

(vi) $\sum_{j=1}^{K} \sum_{i=1}^{n} x_{ji}^2 = \sum_j \lambda_j = \sum_j z'_j z_j$

(vii) $\sum_j p_j = 1$

$p_j = (\Sigma \lambda_j)^{-1} \lambda_j$; p_j being proportionate contribution of j-th principal component to total variation of the x's

(viii) $\sum_{j=1}^{K} r_{jk}^2 = 1$; $k = 1, \ldots\ldots, K$ where r_{jk} trepresents correlation between z_j and x_k

(ix) $z'_j z_j = \alpha'_j X'X\alpha_j$ represents squared sum of deviations of observed points from the plane $x'\alpha = 0$ passing through the origin (Anderson, 1972, pp.278-9).

(x) $AA' = I$

(xi) $X = ZA' = Z_1 A'_1 + U$

where Z_1 is subset of Z and the model is called factor analysis model.

It is obvious from the above results that in actual practice one faces the problem of choosing a z_k from amongst K components defined in (2.4) to obtain the composite indices. If one opts for the component corresponding to maximal λ, one is in fact opting for maximum p_j defined in (2.7, VII) and at the same time opts for maximum sum of squared deviations of the observed points from the fitted plane. Thus, choice of any one of the principal components is linked with two conflicting objectives. Besides sum of squared correlations of all the components with a chosen x variable is unity in view of (2.4, VIII). This indicates that none of the principal components are related completely with any one of the x variables. In other words, none of the principal components provide fully valid composite indices of the constituent variables. Further, the method of principal components does not enable us to easily test as to which of the x variables should be considered on the basis of significant statistical evidence.

3. Full Information Based Composite Index

Let us consider the problem of fitting a plane $x'\alpha$ to any observed scatter of n observations on each one of the variables contained in $K \times 1$ vector x. Taking deviations around the sample means, the plane can be made to pass through the origin though it is not necessary. Deviation of any observed point x_i from the plane $x'\alpha$ is given by $x_i'\alpha$. Therefore, sum of squared deviations is given by

$$(3.1) \qquad \sum_{i=1}^{n} (x'_i\alpha)^2 = \alpha'X'X\alpha = z'z$$

in view of the notations in the preceding section. Since minimisation of $z'z$ is afforded by any arbitrary choice of α, we propose to constraint the elements of as

$$(3.2) \qquad \alpha'\iota = 1$$

where ι represents a K x 1 column vector each element of which is unity and the constraint (3.2) implies that sum of weights is equal to one. In fact, the assumption that sum of the weights is equal to one is more natural than the assumption that sum of squares of the weights equals to one.

This is a constrained optimisation problem. We may use a Lagrangian multiplier λ and specify a function as

$$(3.3) \qquad \phi = \alpha'X'X\alpha - \lambda(\alpha'\iota - 1)$$

First order conditions of optimisation are given by

$$(3.4) \qquad \begin{aligned} 2X'X\alpha &= \lambda\iota \\ \iota'\alpha &= 1 \end{aligned}$$

These relations give the following solution for α and λ

$$(3.5) \qquad \begin{aligned} \alpha &= (X'X)^{-1}\iota\left[\iota'(X'X)^{-1}\iota\right]^{-1} \\ \lambda &= 2\left[\iota' X'X)^{-1}\iota\right]^{-1} \end{aligned}$$

It can be easily verified that

$$(3.6) \qquad \begin{aligned} \iota'\alpha &= \iota'(X'X)^{-1}\iota\left[\iota'(X'X)^{-1}\iota\right]^{-1} \\ &= 1 \end{aligned}$$

The Hessian of ϕ in case of constrained optimisation can be written as

$$(3.7) \qquad \begin{aligned} H &= \begin{vmatrix} 2X'X & \iota \\ \iota' & 0 \end{vmatrix} \\ &= - \left|X'X\right| \iota'(X'X)^{-1}\iota \end{aligned}$$

in view of relevant result of the determinant of partitioned matrices.[1] It follows from (3.7) that determinants of all the principal minors of the Hessian are negative and therefore the solutions in (3.5) imply minimisation of the residual sum of squares in ~~(3.1) subject to~~ constraint (3.2). Thus, the solution of weights of the composite index z_i in (2.1) is unique and provides full information based solution to constrained minimisation.

4. Properties

Properties of the estimated weights and the underlying model in Section 3 can be analysed in different contexts. We propose to do it under the following headings:

Polynomial Functions

The estimation procedure proposed in Section 3 holds good when the relation (2.1) contains a constant term or terms of higher degrees of x variables as in polynomial specification. Accommodation of a constant term would require specification of one of the variables, say x_1, to be such that it assumes the value unity for all the sample units. Polynomial specification would require addition of more variables by defining a new variable corresponding to each one of the added terms of higher than one degree. Thus, if K_1 such terms are added, the specification (2.1) will contain $K + K_1$ weights to be estimated. The number of units of observations n will have to be made large enough so that $n-K-K_1$ is positive to ensure existence of $(X'X)^{-1}$. Taking deviations of variables in (2.1) around their respective sample means would eliminate the problem of separate inclusion of a constant term. In case of polynomial specification, the means have to be

[1] The result is given as

$$\begin{aligned} |A| &= \begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix} \\ &= |A_{22}| \quad |A_{11} - A_{12}A_{22}^{-1} A_{21}| \\ &= |A_{11}| \quad |A_{22} - A_{21}A_{11}^{-1} A_{12}| \end{aligned}$$

of each one of the new variables defined instead of the means of the basic variables whose higher powers provide the additional variables.

Dimensional Changes

Let us consider two alternative specifications of (2.1), namely,

$$(4.1) \quad z_{1i} = \alpha_0 + \alpha_1 x_{1i} + \dots + \alpha_K x_{Ki}$$

and

$$(4.2) \quad z_{2i} = \alpha_0 + \alpha_1 d_1 x_{1i} + \dots + \alpha_K d_K x_{Ki}$$

where α_0 represents the constant term and $d_1, \dots d_K$ represent the dimensional constants of the scale type associated with variables $x_1, \dots, x_K$ respectively. Estimation of coefficients in specification (2.1), (4.1) and (4.2) in accordance with the method proposed in Section 3 would require the following constraints to be satisfied, respectively.

$$(4.3) \quad \alpha_1 + \dots + \alpha_K = 1$$
$$(4.4) \quad \alpha_0 + \alpha_1 + \dots + \alpha_K = 1$$
$$(4.5) \quad \alpha_0 + \alpha_1 d_1 + \dots + \alpha_K d_K = 1$$

It is obvious from these that the estimates of weights will be different in the three situations. Thus estimates of $\alpha_1, \dots, \alpha_K$ from specification (4.1) is not similar to estimates of $\alpha_1, \dots, \alpha_K$ when α_0 has been eliminated after taking deviations around the sample means. Similarly, use of different scaling factors of the dimension of the variables will affect the estimated weights. The weights are thus affected by shifts in coordinates of linear as well as scalar types. The choice of the specification has to be made from a priori conceptual considerations and the composite indices are affected by these in each case. It is natural to expect the indices to behave like this because it has to represent the variables in a joint fashion and therefore depends upon the way these are specified.

Orthogonal Variables

Let us consider the situation where x variables are standardised in the sense that they are measured around their sample means and divided by their standard deviations and the weights are estimated in accordance with specification (2.1). Let us assume further that the standardised variables are orthogonal so that their cross products are zero for the given sample size. In that case

(4.6) $X'X = I$

where I is the identity matrix of size K x K. Using (4.6), we find that the estimates of weights as in (3.5) can be obtained as

(4.7) $\alpha = \iota\,(\iota'\iota)^{-1}$

so that each element of α is equal to K^{-1}. In this case, therefore, the composite index is simply the arithmatic average of the constituent variables. The weights as in (3.5) are therefore natural estension of the centroid concept in the presence of correlations in the constituent variables.

Total Variation of Indices

Given the value of estimated weights, the composite indices of all the n units can be expressed in n x 1 vector form

(4.8) $z = X\alpha$

Sum of squares of these composite indices can be expressed as

(4.9)
$$\begin{aligned} z'z &= \alpha'X'X\alpha \\ &= \left[\iota'(X'X)^{-1}\iota\right]^{-1} \\ &= 2^{-1}\lambda \end{aligned}$$

where use has been made of (3.5). This shows that total variation in composite indices is based upon all the elements of X'X which includes total variation in the x variables as well as covariation in them.

Correlation of Indices with Constituents

Let us consider the situation when x variables are measured around their sample means. In that case

$$(4.10)\qquad \sum_{i=1}^{n} z_i = \sum_{i=1}^{n} \sum_{k} \alpha_k x_{ki}$$
$$= 0$$

since

$$(4.11)\qquad \sum_{k} x_{ki} = 0, \quad k = 1, \ldots\ldots, K$$

Now correlation between z and x_k can be obtained as

$$(4.12)\qquad r_k^2 = (z'z)^{-1} (x'_k x_k)^{-1} (x'_k z)^2$$
$$= z'z (x'_k x_k)^{-1}$$

Because

$$(4.13)\qquad x'_k z = a'_k X'X (X'X)^{-1} \iota \left[\iota'(X'X)^{-1} \iota \right]^{-1}$$
$$= \left[\iota'(X'X)^{-1} \iota \right]^{-1}$$
$$= z'z$$

where vector a_k is defined to consist zero and unity such that

$$(4.14)\qquad x_k = Xa_k$$

and use has been made of (4.9).

Single Common Factor

Let us express each one of the constituent variables in regression form as

(4.15) $$x_k = \beta z + u_k \qquad k = 1, \ldots\ldots, K$$

where least squares estimator of coefficient of z is unity in view of (4.13). Using (4.13), we observe that

(4.16) $$z' \hat{u}_k = 0$$

where $\hat{u}_k$ is least squares estimated residual. The relations (4.15) and (4.16) together show that the variable x_k is broken into two orthogonal components. Further, we can write sum of squares of observations on x_k as

(4.17) $$x_k\, x_k = z'z + \hat{u}_k'\, \hat{u}_k$$

in view of (4.16). This shows that proportion of variation explained by z of x_k is given by r_k^2 defined in (4.12). Relation (4.15) also indicates that z is least square estimate of the systematic part in each one of the constituent variables. In fact, the model (4.15) represents a factor analytic model in the same sense as in (2.7, XI) with a single common factor z for all the constituent variables and factor loading is unity in each case. We may also write sum of residual squares as

(4.18) $$\hat{u}_k'\ \hat{u}_k = (x_k - z)'\ (x_k - z)$$

which shows that z is that common point for all the x_k variables about which $u_k'u_k$ is minimum for each $k = 1, \ldots\ldots, K$. Thus, the point z is centroid in the multivariate sense.

Distarce Measure of Indices

Use of the indices may be made to form homogenous groups of n units. One way of doing this is to rank the n units in ascending or descending order in respect of magnitude of the composite indices. Any such reordering will not disturb the estimates of the α weights. These ranks together with reasonable judgement in respect of separating points may be used to group the sample units into homogenous groups.

Alternatively, one may obtain distance measure of elements of z as

$$(4.19) \quad d_{ij}^2 = (z_i - z_j)^2 = (x_i - x_j)' \alpha\alpha' (x_i - x_j)$$

where

$$(4.20) \quad z_i = x_i'\alpha$$

is i-th (i=1, ..., n) element of z and x'_i is i-th row of matrix X. The distance measure as in (4.19) preserves the distance in terms of the original data because $\alpha\alpha'$ remains invariant over i, j = 1, ...,n. The distances in (4.19) may be used to group individual units into appropriate number of groups. In most cases, ranking and distance methods would lead to similar groups.

Testing of Significance

In order to test as to which one of the constituents of z make statistically significant contribution, we may make use of the regression model implied by (4.15). This is equivalent to two variable regression model and we may define a t statistic with (n-2) degrees of freedom as follows:

$$(4.21) \qquad t = r_k(1-r_k^{\;2})^{-\frac{1}{2}} (n-2)^{\frac{1}{2}}$$
$$k = 1, \ldots, K$$

The factor z can be said to be correlated significantly with x_k if computed t as in (4.21) is greater than the corresponding tabular value with n-2 degrees of freedom.

This test procedure may be used to ascertain those constituents that are found to be related with the composite indices at a specified level of significance. This provides a convenient criterion of deciding as to which of the several competing variables have appreciable statistical evidence to be included in the set of constituents of a composite index. In other words, the test procedure can be used to determine statistically valid composite concepts such as development, personality and resource regions, etc.

REFERENCES

Andersen, T.W., An Introduction to Multivariate Statistical Analysis, Wiley Eastern, New Delhi, 1972.

Harman, H.H., Modern Factor Analysis (2nd ed.), University of Chicago Press, 1967.

Harman, H.H., "How Factor Analysis can be used in classification", RB-71-65, Educational Testing Service, Princeton, N.J., 1971.

Johnston, J. Econometric Methods, (2nd ed.), McGraw Hill, International Student Edition, New Delhi, 1972.

McDonald, Roderick, P., "A Unified Treatment of the Weighting Problem," Psychometrica, Vol. 33, No. 3, 1968

Pearson, K., "On Lines and Plane of closest Fit to Systems of Points in Space," Phil. Mag, 2(sixth series), 559-572, 1901.