References

Morgan Guaranty Trust, 1983, Global debt: Assessment and long-term strategy, in: World financial markets (Morgan Guaranty Trust, New York) 1–14.
Cline, William R., 1983, Developing country debt under alternative global conditions: 1983–1986 (Institute for International Economics, Washington, DC).

Journal of Development Economics 21 (1986) 319–325. North-Holland

A NOTE ON VARIABLES AND OBSERVATIONS IN FACTOR ANALYSIS

Peter MURRELL

University of Maryland, College Park, MD 20742, USA

Received December 1984, final version received August 1985

Bumb (1982) has claimed that factor analysis estimates are spurious when the number of observations is less than the number of variables. It is shown that the distinction between observation and variables is not one that carries any theoretical force in factor analysis. One can treat variables and observations symmetrically, leading to two economic interpretations of any model used in factor analysis. Therefore, the relation between numbers of variables and numbers of observations is not relevant to the question of whether a factor analysis model can be estimated.

1. Introduction

If judiciously used, factor analysis can provide valuable information in situations where the standard regression methods of econometrics are not applicable. However, because the assumptions of factor analysis are very different from those usually employed in econometrics, researchers must be alert to the possibility of error. In particular, one must be especially clear about the domain of applicability of this technique. It is the objective of this paper to clarify the limits of this domain. Happily, this clarification shows that factor analysis can be used in a wider variety of situations than was previously thought appropriate by some authors [Conklin and Hadden (1974), Bumb (1982a, b)].

In their seminal application of factor analysis to the economics of development, Adelman and Morris (1967) used, at some stages, data sets that contained fewer countries than social and political indicators. If a country is viewed as the unit of observation and an indicator as a variable, then one has more variables than observations – perhaps a troubling fact to economists steeped in standard econometrics. Bumb has based his criticism of Adelman–Morris on just this fact: 'In my methodological note on factor analysis, I have shown that when the number of variables ... exceeds the number of observations ... the sample correlation matrix is singular and positive definite, and therefore, the factor loadings derived from such [a] correlation matrix are spurious. Hence, in the empirical estimation of factor

loadings, the researcher should not include more variables than observations.'[1]

Adelman and Morris (1982), in replying to Bumb, do not concede his general point, but they do appear to give up much ground.[2] They resort to the argument that their estimates have been shown to be robust. While the robustness in indeed impressive, Adelman and Morris did not need to adopt this line of defense. In this brief note, I show that Bumb's criticism is unjustified from a theoretical perspective.

The central argument of this paper is that the distinction between 'variables' and 'observations', as it is conventionally employed in econometrics, is irrelevant in factor analysis. In section 2, I show that there are two ways of viewing any set of equations to which factor analysis is applicable. The equations lead to two models, with the roles of 'variables' and 'observations' reversed. If the existence of one model with variables greater in number than observations necessarily implies the existence of a companion model in which the reverse is true, one must doubt whether the distinction between variables and observations is important, or even valid. Section 3 establishes that this doubt is justified. In that section, I show that estimates obtained from the two models are identical. Since one model satisfies the condition that 'variables' are fewer than 'observations', while the other does not, it must be incorrect to argue that satisfaction of this condition is necessary for the use of factor analysis.[3] Finally, in section 4, I move the argument from an abstract level to a practical one. I show that the Adelman–Morris model does lend itself to two interpretations. By examining these interpretations, one can clearly see that the distinction between 'observations' and 'variables' is purely an aid to conceptual understanding. This distinction is not one that has any force in the process of deciding which statistical techniques to apply.

2. The two models implicit in factor analysis

The basic set of equations used in factor analysis is

$$y_{ij}=\sum_{k=1}^{K} a_{ik}f_{kj}+e_{ij}, \qquad i=1,\ldots,I, \quad j=1,\ldots,J, \tag{1}$$

[1]Bumb [1982b, p. 125]. This quote summarizes Bumb (1982a). The ellipses solely replace Bumb's mathematical symbols.

[2]This is my interpretation. I am not sure Adelman–Morris would agree. Their paper contains the following sentence: 'The reduced-variable analyses, with plenty of degrees of freedom, confirmed the all-variable analyses.' I take this sentence to mean that Adelman–Morris view the relation between variables and observations as indicating the number of degrees of freedom. I will show that this is incorrect.

[3]The distinction may be helpful in understanding the economic content of the theory, however. The example given, I believe, clearly shows this point.

where y_{ij} is the jth observation on the ith variable, the a_{ik} are coefficients, the f_{kj} are 'factors', and the e_{ij} are error terms. (I use the terms 'observation' and 'variable' as they are conventionally employed in factor analysis. The reader should be warned, however, that the aim of the present paper is to show that these terms are interchangeable.) The y_{ij}'s are observable dependent variables. The factors are unobservable independent variables.

The first interpretation of (1) is developed by placing the equations in matrix form. Throughout the following, a letter with a dot subscript will indicate the column vector formed by listing all the elements obtained by varying the subscript that the dot has replaced [e.g., $y'_{\cdot j}=(y_{1j},\ldots,y_{Ij})$]. One can then rewrite (1) as

$$y_{\cdot j}=Af_{\cdot j}+e_{\cdot j}, \qquad j=1,\ldots,J, \tag{2}$$

where A is the matrix of elements formed from the a_{ij} (the ith row of A is $a'_{i\cdot}$).

In the interpretation following from (2), $y_{\cdot j}$ is a single observation on a vector-valued variable, A is a matrix of coefficients, and $f_{\cdot j}$ is one realization of a vector-valued variable. Rewriting (2) in matrix form, where Y is the matrix with $y_{\cdot j}$ as its columns and F and E are similarly defined, one obtains

$$Y=AF+E. \tag{3}$$

There is an alternative manner, however, in which one could view (1). Nothing in the equations above, nor, as will be shown, in the statistical procedures or the underlying economic theory, dictates that observations must be treated differently from variables. Thus, if instead one views y_{ij} as the ith observation on the jth variable, the a_{ik} can be interpreted as factors and the f_{kj} as coefficients. Then, instead of (2), one could have written

$$y_{i\cdot}=F'a_{i\cdot}+e_{i\cdot}, \qquad i=1,\ldots,I, \tag{4}$$

in which the interpretation is now that $y_{i\cdot}$ is a single observation on a vector valued variable. Proceeding as before, one can obtain, by writing (4) in matrix notation,

$$Y'=F'A'+E'. \tag{5}$$

Obviously, (3) and (5) are identical equations. However, in constructing (3) and (5) the roles of variables and observations have been reversed, as have the interpretations A and F. At this stage, it is important to emphasize that there is nothing in the assumptions necessary to make factor analysis applicable that invalidates either interpretation.

The above has merely shown that the development of a factor analysis model does not require an immutable definition of which elements in the model are variables and which observations. The use of factor analysis implies the existence of two models, one of which must have at least as many variables as observations. Hence, one must begin to doubt the validity of a condition of estimation expressed in terms of the relation between numbers of variables and observations. To show that this doubt is justified, one should examine the procedures for estimation of factor analysis models.

3. Estimates of the two models

As is well known [Lawley and Maxwell (1971, p. 7)], eqs. (1) cannot be estimated without placing further restrictions on the model. Here I adopt the simplest and most commonly used set of restrictions. My goal is simply to show that 'variables fewer than observations' is not a *general* condition necessary for factor analysis. This goal can be most efficiently accomplished by confining the exposition to the set of estimating techniques with which the largest number of readers is likely to be familiar.

The assumptions used follow those of Dhrymes (1970, pp. 77–84). They are applied to the first model of section 2 [eqs. (2)]. First, it is assumed that the factors are uncorrelated and have unit variance:

$$E(f_{\cdot j} f'_{\cdot j}) = I, \tag{6}$$

where I is an identity matrix of appropriate order. Second, the error terms have zero mean, are uncorrelated, and have a common variance:

$$E(e_{\cdot j}) = 0, \qquad E(e_{\cdot j} e'_{\cdot j}) = \sigma^2 I, \tag{7}$$

where σ^2 is a scalar.

Under (6) and (7), the estimation of the factor analysis model reduces to the problem of principal components. The estimate of A is the matrix $\hat{A}$ whose columns are the K eigenvectors of YY' corresponding to the K largest eigenvalues. Thus, $\hat{A}$ is found from the equation

$$(YY')\hat{A} = \hat{A}L, \tag{8}$$

where L is a diagonal matrix whose elements are the K largest eigenvalues of YY'.[4]

[4]Some authors have suggested that YY' must be non-singular in order to find $\hat{A}$ from (8). This may be true for certain computational routines, but it is certainly not correct in a theoretical sense. Modern iterative computational routines do not rely on non-singularity.

It is often a goal of factor analysis to estimate the values of the underlying factors, $\hat{F}$. Noting (2) and (7), these values are provided by $Y = \hat{A}\hat{F}$. Since $\hat{A}$ is a matrix of eigenvectors, $\hat{A}'\hat{A} = I$ [Lawley and Maxwell (1971, pp. 128–129)]. Thus, $\hat{F}' = Y'\hat{A}$. Premultuplying by $Y'Y$, one obtains

$$\begin{aligned} Y'Y\hat{F}' &= Y'YY'\hat{A} \\ &= Y'\hat{A}L, \end{aligned}$$

from (8). Therefore,

$$(Y'Y)\hat{F}' = \hat{F}'L. \tag{9}$$

Since the K largest eigenvalues of YY' are identical to those of $Y'Y$,[5] $\hat{F}'$ is the matrix whose columns are the eigenvectors corresponding to the K largest eigenvalues of $Y'Y$.

The essential conclusion of the above analysis derives from observing that (8) and (9) are analogous equations. The formula for the estimation of A, (8), was derived by applying assumptions (6) and (7) to the first model [eq. (2)]. From (8), however, it is easy to see that starting from the second model [eq. (4)] and adopting assumptions analogous to (6) and (7) by applying them to $a_{i\cdot}$ and $e_{i\cdot}$, one would have obtained (9). The objective then would have been to find $\hat{F}'$ directly by applying standard factor analysis methods to model 2. Yet, here (9) was derived directly from the first model. The estimate of F' [and consequently A, noting the symmetric role of A and F in (2) and (4)] is exactly the same, whichever interpretation of the factor analysis equations is used.[6]

The above has shown that 'variables fewer than or equal to the number of observations' cannot be a necessary condition for deriving meaningful estimates of a factor analysis model. If the first model [eqs. (2)] does not satisfy this condition, then the second model would. Since the estimates of the two models are the same and are unique, the ability to estimate one model implies the ability to estimate its companion model, even though the condition must be violated in one case. Bumb's (1982a) claim, that 'spurious' estimates must be obtained from a model with fewer variables than observations, is incorrect.[7]

In order to forestall misinterpretation of the above analysis, it is appropriate at this point to emphasize that there is one important restriction that must be satisfied by the model. Meaningful estimates can be obtained only if the number of factors (K) is less than both the number of variables and the

[5]Eqs. (9) and (8) show that the K largest eigenvalues of YY' and of $Y'Y$ are identical.

[6]Here, it should be noted that eigenvectors are unique up to multiplication by a scalar.

[7]In fact, the relationship between the number of variables and the number of observations determines the relative degree of reliability of the estimates of the A and F matrices.

number of observations. Moreover, to gain some confidence in the estimates, K should be well below at least one of the latter numbers. This restriction is, of course, perfectly consistent with a fundamental objective of factor analysis – the reduction of dimensionality.

4. An example

Up to the present, the analysis has proceeded in a somewhat abstract manner. In the remainder of the paper, I will make that analysis concrete by casting it within a specific example. Presentation of the example will, I hope, finally convince any still sceptical readers that the terms 'variables' and 'observations' are interchangeable in factor analysis. I have chosen to present the Adelman–Morris model as the example both because it is the seminal application of factor analysis in development economics and because the estimates of this model were the focus of Bumb's criticisms.

Adelman and Morris [(1967); see, for example, pp. 149–172] used as their data the values for *countries* of various political and social *indicators*, such as urbanization or literacy. They saw these values as being determined by a small number of underlying *characteristics* (e.g., degree of rationalization of social behavior, values and institutions).[8] Implicit in the relationship between indicators and characteristics is the existence of a set of *parameters* summarizing that relationship.

The terms italicized in the previous paragraph indicate the four basic elements of the Adelman–Morris study. In applying factor analysis, one can associate each of those terms with one of the following general concepts: variables, observations, factors, and coefficients. The way in which Adelman–Morris (and Bumb) made this association is shown in table 1. The first column of the table lists the four general concepts. The second column lists the elements that Adelman–Morris associated with these concepts.

The aim in presenting the table is to delineate the two alternative interpretations of the Adelman–Morris analysis. Thus, in column three, the companion model to the Adelman–Morris interpretation is presented. For the sake of brevity, I will not elaborate on the description given in the table. The important point to note is that the roles of variables and observations are reversed in moving from Model 1 to Model 2, as are the roles of coefficients and factors. Hence, the table establishes that the two models, inherent in any factor analysis, can be given a real interpretation in the Adelman–Morris context.

The above analysis should not be taken to imply that both interpretations of the Adelman–Morris analysis should be given equal weight. Rather, the intention has been solely to show that each interpretation is logically

[8]The reader should be warned not to assume immediately that these characteristics are properties of countries. This assumption is only true of Model 1. In Model 2, the characteristics must be viewed as properties of indicators.

Table 1

The two models implicit in the Adelman–Morris study.

General theoretical concept	Model 1	Model 2
One observation on all variables	All social indicators for one country (e.g., urbanization, literacy, etc. in Nigeria)	All countries' values for one social indicator (e.g., urbanization in Nigeria, Egypt, etc.)
A coefficient	The degree to which an increase in a particular social characteristic in any country increases the level of a particular indicator in the country (for example, the effect of rationalization on urbanization)	The degree to which an increase in any indicator's embodiment of a particular social characteristic increases a country score for that indicator (e.g., the amount of rationalization in Nigeria)
A factor score	The level of a social characteristic in a country (e.g., the level of rationalization in Nigeria). Viewed as a property of a country	The extent to which a social indicator embodies a particular social characteristic (e.g., urbanization's degree of embodiment of rationalization). Viewed as a property of an indicator
A vector of observations of one variable	All countries' values for one social indicator	The values of all social indicators for one country

possible. Certainly, one interpretation (Model 1) is more intuitive than the other, as the reader can easily verify by noting the semantic difficulties involved in describing Model 2. Use of the interpretation embodied in Model 1 is surely compelling for expositional reasons alone. The present paper has shown that one can freely choose either interpretation even if, as a result of this choice, the number of variables is larger than the number of observations.[9]

References

Adelman, I. and C.T. Morris, 1967, Society, politics, and economic development (Johns Hopkins University Press, Baltimore, MD).
Adelman, I. and C.T. Morris, 1982, Factor analysis and development: A reply, Journal of Development Economics 11, no. 1.
Bumb, B., 1982a, Factor analysis and development: A note, Journal of Development Economics 11, no. 1.
Bumb, B., 1982b, Factor analysis and development: A rejoinder, Journal of Development Economics 11, no. 1.
Conklin, G.H. and S.G. Hadden, 1974, Response, Journal of Asian Studies, Nov.
Dhrymes, P., 1970, Econometrics (Harper and Row, New York).
Lawley, D.N. and A.G. Maxwell, 1971, Factor analysis as a statistical method (Elsevier, New York).

[9]Nothing in the present paper should be taken to imply that researchers need not be concerned about small numbers of observations or variables.