VISUALIZING FACTOR ANALYSIS IN VARIABLE SPACE AND SUBJECT SPACE

Chong-Ho Yu, Sandra Andrews, David Winograd, Angel Jannasch-Pennell, &
Samuel A. DiGangi, Arizona State University
Chong-Ho Yu, Instruction and Research Support 0101, Arizona State University, Tempe AZ 85287

Key Words: Factor Analysis, Variable space, Subject space, Biplot

Teaching and learning factor analysis is challenging. As Pedhazur and Schelkin (1991) pointed out,

> The literature of factor analysis (FA) is vast and generally complex. Perusing even small segments of this literature in an effort to understand what FA is, how it is applied, and how the results are interpreted is bound to bewilder and frustrate most readers. This is due to a wide variety of contrasting and contradictory views on almost every aspect of FA, serious misconceptions about it, and lack of uniformity in terminology and notation. (p.590)

To trace the sources of misconceptions, a review of textbooks and websites dedicated to teaching factor analysis was conducted. Textbooks and websites were identified through a review of books in print, Web-based search engines (Alta Vista, Infoseek, and Yahoo), and discussions with faculty teaching quantitative methods courses as well as members of the Educational Statistics Listserv group (EDSTAT-L). A list of the textbooks can be viewed at http://is.asu.edu/research along with a discussion of the bases upon which they were chosen.

An assessment examining concepts of factor analysis was constructed and administered to students of differing levels of statistical literacy. The results of these investigations formed the basis for construction and implementation of a computer-based multimedia instructional program, centering on the perspective of "subject space." The impact of this instructional program was evaluated using "think aloud protocol" (Someren, Barnard, & Sandberg, 1994), a method of recording subjects' mental processes by having them verbalize their thinking as they navigate through the program.

Review of approaches of teaching factor analysis

Strategies of teaching factor analysis can be classified into the conceptual, mathematical, and geometrical approaches. These three approaches are used in varying combinations by the following authors.

Conceptual approach

Examples of the conceptual approach can be found in Thapalia (1998), Wulder (1998) and Ingram (1998). Although the purpose and the application of factor analysis are emphasized in this approach, questions regarding the underlying dimensions of the data and their relationship to the Cronbach Alpha coefficient are not mentioned. Before a factor analysis is performed, the Cronbach Alpha coefficient should be computed to check whether the data are unidimensional or multidimensional. Without an understanding of this premise, researchers often inadvisably conduct factor analysis regardless of the dimensionality of the data. Researchers commonly claim that they have extracted several subscales by factor analysis and that all subscales are strongly correlated to the total scales. This claim is invalid when all subscales are correlated to each other and indeed there is only one dimension of the data. Factor analysis is usually placed in textbooks under multivariate analysis; it is assumed that students understand how multivariate techniques are used to address the multi-dimensionality of the data. This assumption may not be applicable to all students, as Huberty (1994) saw and thus stressed that multivariate analysis techniques are analyses of data vectors for each individual observation under study, consisting of two or more scores.

In the conceptual approach, practical uses of factor analysis are emphasized. Technical terms such as "eigenvalue" and "orthogonality" are omitted, or mentioned only in passing. Serious misconceptions may arise when explanations of these terms are omitted. For example, Thapalia (1998), Wulder (1998), and Tabachnick and Fidell (1996) state that the researcher can "rotate" factors to gain a better interpretation of the data, possibly leading students to impose their own non-technical understanding on what appear to be words with everyday meanings. For example, vectors representing factors are rotated in subject space. Learners with no understanding of vectors and subject space may assume that this rotation implies spinning a plot to get a better perspective, or to use different variables at different times as if they were tires to be rotated.

Even common terms such as 'weight," "model," and "factor" that instructors assume will be understood by students may cause confusion. Students often confuse the meaning of the term "factor" in factorial analysis with that in factor analysis. In the former a factor is an observed variable with clearly identified levels while in the latter a factor is a latent and abstract mathematical construct. This major difference was not emphasized in the texts reviewed, which merely define a factor as a latent variable or a hypothetical construct (e.g. Harmon, 1976).

Some authors include these more difficult terms rather than avoiding them (e.g. Ingram, 1998). When technical terms are used to explain a common term such as "factor," students may be overwhelmed by what appears to be alien

language. Ingram (1998) defined a factor as "a vector which is weighted proportionally to the amount of the total variance which it represents. The factor loadings are the elements in the factor vector. The sum of the squares of these loadings should equal the eigenvalue." Understandably, students reading this may have difficulty with these definitions as they attempt to relate "factor" to "vector," "total variance," "loadings," and "eigenvalue."

The current mutlimedia program begins with an explanation of basic terms such as "space" and "variance" in order to ensure that readers do not impose their own non-technical understanding onto statistical terminology.

Mathematical approach

In the mathematical approach, factor analysis is taught within the context of the linear model (Harmon, 1973; Gorsuch, 1983; Joreskog & Sorbom, 1979; Basilevsky, 1994; Kim & Mueller, 1978). One difficulty with this approach is that while both regression and factor analysis are weighted linear combinations, the difference in mathematical terms used for the two procedures fail to help the learner integrate these procedures under the umbrella concept of the linear model. In regression, the weight of the linear combination is called "coefficient" or "beta weight" while in factor analysis this weight is called "loading." With the exception of Kim and Mueller (1978), the texts reviewed did not emphasize the relationships among the preceding terms, and students are unlikely to make the necessary connections themselves. The current multimedia program uses regression as a metaphor for factor analysis since the linear model subsumes both.

Eigenvector and eigenvalue are concepts central to the topic of factor analysis. In some introductory statistics texts, an overly mathematical discussion of these terms may be confusing for students or even researchers who do not have a strong mathematics background. For instance, Hagle (1995) has the following explanation: "X is called an eigenvector (characteristic vector; eigen is German for characteristic) if there exists a nonzero vector X_{n*1} such that $A_{n*n} X_{n*1} = \Lambda X_{n*1}$. This scalar Λ is called an eigenvalue of A_{n*n}." (p.89) One would be hard pressed to find an intermediate student that could make much sense of this equation. In an attempt to remove the conceptual block, the current program uses animated graphics to illustrate eigenvector and eigenvalue in the context of subject space.

Geometrical approach

Several topics such as "orthogonality" are spatially oriented. A text-based explanation would define "orthogonal" as "uncorrelated," but the learner may have difficulties understanding this statement. On the other hand, a spatial representation of two perpendicular vectors is clear (Gorsuch, 1983).

A number of reviewed texts (e.g. Basilevsky, 1994; Harmon, 1976; Comrey & Lee, 1992; Wulder, 1998) mentioned that factor analysis is sensitive to an illconditioned correlation matrix, which is a manifestation of multicollinearity. None of these texts utilize graphical representations to explain conditioning and multicollinearity. In a simplistic sense, multicollinearity is the oppositie of orthogonality. Perhaps the omission of graphical representation of the former is based upon the assumption that the student has learned the concept of orthogonality, but this is not necessarily the case. Multicollinearity is more comprehensible if orthogonality is understood. To fill this conceptual gap, the multimedia program series designed for this study contains a module addressing multicollinearity and employing graphical illustration.

The geometrical approach relies upon the concept of subject space as a means of visualization of spatial relationships. Most textbooks using the geometrical approach (e.g. Comrey & Lee, 1992; Pedhazur and Schelkin, 1991) begin with matrix algebra and then plot vectors in a coordinate system. In this context, it is difficult to convert the matrix algebra information to a representation in person space.

In addition, no textbook reviewed uses the terms "subject space" or "person space." Instead vectors are presented in "Euclidean space," (Joreskog & Sorbom, 1979) "Cartesian coordinate space," (Gorsuch, 1983), "factor space," (Cormrey & Lee, 1992; Reese & Lochm ler, 1998) and "n-dimensional space" (Krus, 1998). The first two terms do not adequately distinguish vector space from variable space. A scatterplot representing variable space is also an Euclidean space or a Cartesian coordinate space. The third is tautological. Stating that factors are in factor space may be compared to stating that Americans are in America. The term does not provide additional information. "N-dimensional space" is closer to the meaning of subject space since in subject space the number of dimensions is equal to the number of subjects. On the other hand, the notation "n" could mean either the number of subjects or just any number.

Without clearly distinguishing subject space from variable space, an explanation of vectors may be difficult to follow. The current multimedia project is based upon the belief that starting from variable space and then relating subject space to variable space is an easier path. Finally, the program shows both spaces at once using the biplot for illustration.

In summary, the three conventional approaches are adopted and enhanced. The conceptual approach is used with further explanations of some common terms such as "factor, " "space," "model," and "rotation." The mathematical approach is used to compare and contrast regression and factor analysis in the context of weighted linear combinations. Lastly, the geometrical approach is

applied to help learners transit easily from variable space into subject space.

Survey analysis

A survey was developed by a panel consisting of one statistician and two instructional designers, with content validity established by two experts in the field. Data were collected using a Web-enabled database server accessible at http://129.219.5.245/statsurvey. No time constraints were set. The survey contains five short-answer questions, one multiple-choice question, and three identify questions on concepts regarding factor analysis and linear models. Twenty-five graduate students responded to the survey. On the average, respondents have previously taken 4.94 undergraduate and graduate statistics courses. Respondents came from a wide variety of academic backgrounds that include a Bachelor's or Master's degree in education, mass communication, mathematics, engineering, psychology, sociology, economics, and others.

The survey confirmed the researchers' suspicion that most students confuse the definition of factor in factor analysis with that in factorial analysis. In factor analysis there are no dependent or independent variables, yet twenty-five percent of respondents referred to factors as predictors, independent variables or causes. Only eight percent of the participants could answer the question correctly while all others gave irrelevant answers.

The survey also verified the researchers' assertion that many students failed to conceptualize factor analysis under the premise of weighted linear combinations. Eighty-eight percent were not able to conceptualize the connections between weight, coefficient, and loading. Sixteen percent could not distinguish weights from variables.

The difference between Eigenvectors and regression lines is another area of major confusion. Twelve percent of the participants misidentified Eigenvectors as regression lines, thirty-two percent as "Regression vectors," and twelve percent as "Eigenlines," which do not exist.

Description of the program

Regression is a topic that most intermediate statistics students have studied. As the survey results indicate, this prior knowledge is a source of misconceptions since eigenvectors in subject space are often misperceived of as regression lines in variable space. Yet this misperception provides an opportunity to use regression as a basis of comparison in explaining the differences between variable space and subject space. Regression becomes a metaphor with which to illustrate factor analysis.

A multimedia program incorporating this approach was developed using Director (Macromedia, 1998) as a remedy for the problems discussed above.The multimedia

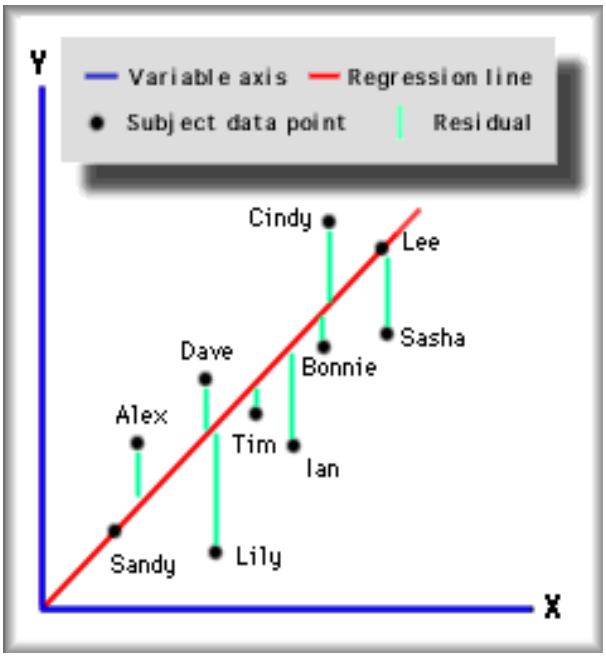

Figure 1(a). Regression represented in variable space

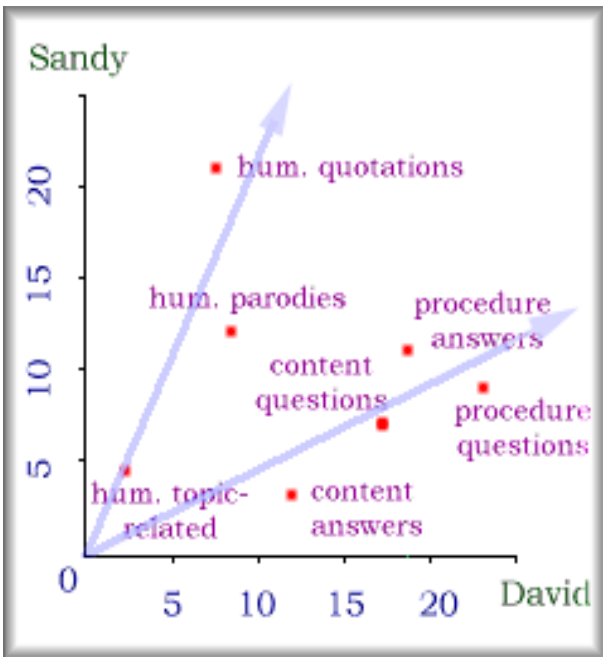

Figure 1(b). Factor analysis represented in subject space

program begins with a presentation of regression in variable space, then shows the user how information can be converted from variable space to subject space. Properties of regression are used to explain the properties of factor analysis as shown in Table 1 and Figure 1.

The meaning of a vector in subject space can be more easily understood if the learner can relate the vector to a person in variable space. Regression can thus be used as a metaphor to enhance understanding of the relationship

between regression and factor analysis.

The multimedia program uses graphics and animation to illustrate both subject and variable space. Two graphing techniques that combine these two types of space are Coneplot (Dawkinks, 1992), which is available in S-Plus, and Gabriel biplot (Gabriel, 1981), which is available in JMP (SAS Institute, 1997) and SAS/Insight (1998). Neither is directly related to factor analysis. Coneplot is primarily used for spotting multiple-dimensional outliers and discriminating between clusters; Gabriel biplot is intended for principal component analysis (PCA). Of the two, Gabriel biplot was chosen for three reasons. First, the basic objectives and principles of PCA and factor analysis are very similar except for the fact that the latter addresses communality. Second, Gabriel biplot can illustrate both regression lines and eigenvectors, which is in line with our instructional strategy. Finally, for beginning and intermediate statistics students, Coneplot may appear complicated and counterintuitive (see Figure 2) whereas Gabriel biplot is clear and self-explanatory (see Figure 3).

The biplot has two limitations associated with a graphical approach to data analysis: (a) A biplot uses only partial information from the singular value composition, which is a variance-maximizing transformation of the data matrix. In other words, it gives an approximation of the data rather than showing all data. (b) It is based on an assumption that the structure underlying the data is linear. If the data structure does not conform to linearity, a biplot will show a distorted view of the data (Jacoby, 1998).

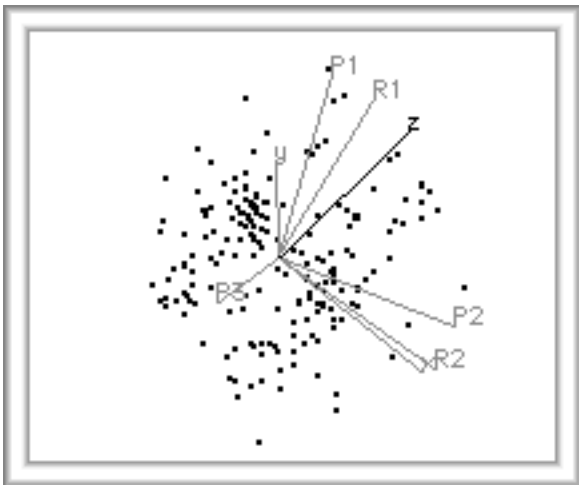

Figure 3. Example of a biplot

Realizing these shortcomings, the authors do not endorse the biplot as a data analysis tool in our tutorial, but use it only as a teaching tool.

Think aloud protocol analysis

Twelve subjects with differing levels of computer and statistical literacy were asked to perform think aloud protocol and were videotaped by the researcher as they navigated through the beta program. The recordings were analyzed and coded for common difficulties in user interace and statistical understanding. The program was then revised in accordance with user suggestions. Subjects found that the program helped them to clearly distinguish regression lines and egienvectors. Their comments also

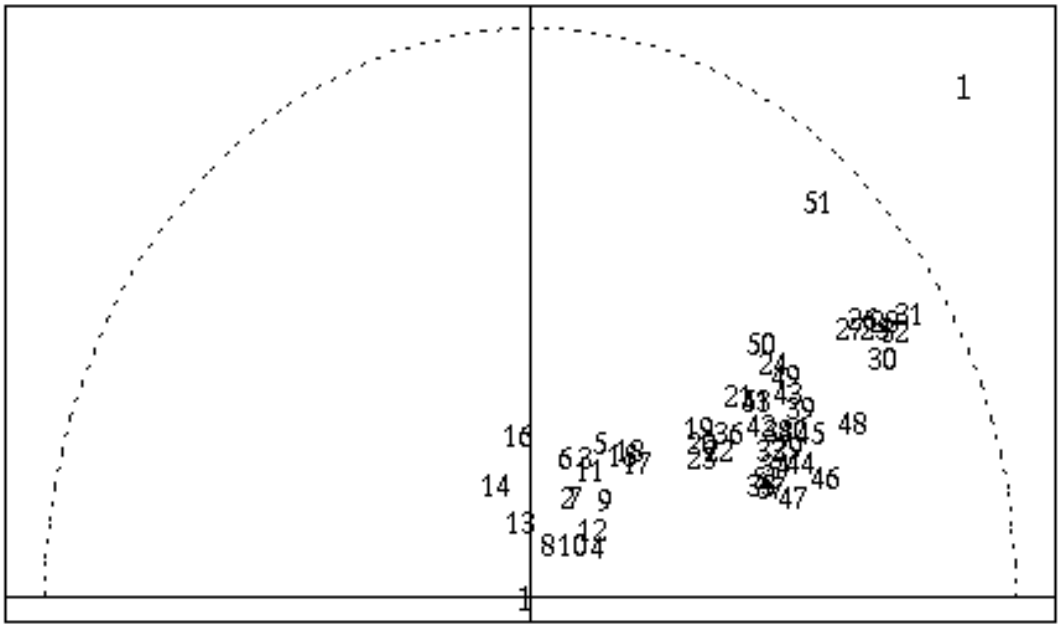

Figure 2. Example of a Coneplot

indicated that most subjects could easily follow the instruction, which was presented in a step-by-step manner.

Conclusion

Conventional pedagogical approaches were developed on the assumptions that certain terminology will be understood by learners, but this is not necessarily true. The computer industry began to realize the confusion caused by the command syntax and error messages during computer operation, and thus computer user interface have been redesigned to be more user-friendly. By the same token, statisticians should consider renaming certain terms or expanding on the explanations of those terms. In particular, they should further explain relationships among the terms and possible integration of these terms under the linear model. Further, conventional teaching methods are confined to limited computing resources. With the advance of high-power computers, visualizing eigenvectors in subject space is an easier path for students to conceptualize factor analysis. The computer-based multimedia program developed for this study is available at http://is.asu.edu/research/. The program reflects our pursuit of enhancing statistical education. Use of the application and dialogue on this topic are encouraged.

Acknowledgements

Special thanks to Mr. Eldon Norton and Robert Sookvong for reviewing this paper.

References

Basilevsky, A. (1994). Statistical factor analysis and related methods: Theory and applications. New York: John Wiley & Sons, Inc.

Comrey, A. L. & Lee, H. B. (1992). A first course in factor analysis (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Dawkins, B. P. (1992). Investigating the geometry of a p-dimensional data set. Wellington, New Zealand: The Institute of Statistics and Operations Research.

Gabriel, K. R. (1981). Biplot display of multivariate matrices for inspection of data and diagnose. In V. Barnett (Ed.) Interpreting multivariate data. London: John Wiley & Sons.

Gorsuch, R. L. (1983). Factor analysis (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Hagle, T. (1995). Basic math for social scientists. Thousand Oaks, CA: Sage Publication.

Harmon, H. H. (1976). Modern factor analysis (3rd ed.). Chicago, IL: The University of Chicago Press.

Huberty, C. (1994). Why multivariate analyses? Educational and Psychological Measurement, 54, 620-627.

Ingram, P. (1998). Multi-Variate Statistics. [On-line] Available URL: http://137.111.98.10/users/pingram/mmvar.html

Jacoby, W. G. (1998). Statistical graphics for visualizing multivariate data. Thousand Oaks: Sage Publications.

Joreskog, K. G. & Sorbom, D. (1979). Advances in factor analysis and structural equation models. Cambridge, MA: ABT Books.

Kim, J. O. & Mueller, C. W. (1978). Introduction to factor analysis: What is and how to do it? Newbury Park: Sage Publications.

Krus, D. (1998). Visual statistics with multimedia. Tempe, AZ: Cruise Scientific.

Macromedia, Inc. (1998). Macromedia 6. [On-line]. Available URL: http://www.macromedia.com.

Pedhazur, E. J. & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Reese, C. E. & Lochm ler, C. H. (1998). Introduction to factor analysis. [On-line] Available URL: http://www.chem.duke.edu/~reese/tutor1/factucmp.html

SAS Institute. (1997). JMP IN: Software for statistical visualization on the Apple Macintosh. Cary, NC: Author.

SAS Institute. (1998). SAS/INSIGHT User's Guide, Version 6 (3rd ed.). Cary, NC: Author.

Someren, M. W., Barnard, Y. F., & Sandberg, J.A.C. (1994). The think aloud method : A practical guide to modelling cognitive processes. London ; San Diego : Academic Press.

Tabachnick B. & Fidell, L. S. (1996). Using multivariate statistics (3rd ed.). New York: Harper Collins College Publishers.

Thapalia, F. (1998). Multivariate statistics: An introduction. [On-line] Available URL: http://trochim.human.cornell.edu/tutorial/flynn/multivar.htm

Wulder, M. (1998). Principal components and factor analysis. [On-line] Available URL: http://www.pfc.forestry.ca/landscape/inventory/wulder/mvstats/pca_fa.html

Appendix

Table 1
Mapping Between Variable Space and Subject Space

	Variable space	**Subject space**
Graphical representation	The axes are variables; the data points are people.	The axes are people; the data points are variables.
Reduction	The purpose of regression analysis is to reduce a large number of responses from people into a small manageable number of trends called "regression lines".	The purpose of factor analysis is to reduce a large number of variables into a small manageable number of factors, which are represented by eigenvectors.
Fit	The purpose of this reduction of responses is essentially to make the scattered data form a meaningful pattern. To find the pattern in variable space we "fit" the regression line to the responses. In statistical terms it is referred to as the "the best fit."	In subject space we look for the fit between the variables and the factors. One wants each variable to "load" into the factor most related to it. In statistical terms this is referred to as "factor loading."
Criterion	In regression the sum of the squares of residuals determines the best fit. In statistics it is referred to the least squares criterion, which is used to make the reduction and the fit.	In factor analysis one sums the squares of factor loadings to get the eigenvalue. The size of the eigenvalue determines how many factors are "extracted" from the variables.
Structure	In regression the goal is to have the regression line to pass through as many points as possible.	In factor analysis the eigenvalue is geometrically expressed in the eigenvector. One wants the eigenvector to pass through as many points as possible while each variable is clearly loaded into a factor. In statistical term this is referred to as "simple structure."
Equation	In regression the relationship between the outcome variable and the predictor variables can be expressed in a weighted linear combination such as $Y = a + b_1X_1 + b_2X_2 + e$. In regression the weight is called the coefficient.	In factor analysis the relationship between the latent variable (factor) and the observed variables can also be expressed in a weighted linear combination such as $Y = b_1X_1 + b_2X_2$ except that there is no intercept in the equation. In factor analysis the weight is called loading.