



AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
<http://ageconsearch.umn.edu>
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

ERASMUS

Rept 8644/A

ECONOMETRIC INSTITUTE

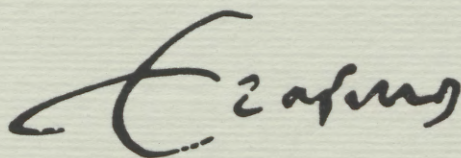
GIANNINI FOUNDATION OF
AGRICULTURAL ECONOMICS
LIBRARY

NOV 17 1987

PRINCIPAL COMPONENTS

T. KLOEK

REPORT 8644/A



ERASMUS UNIVERSITY ROTTERDAM, - P.O. BOX 1738 - 3000 DR ROTTERDAM - THE NETHERLANDS

PRINCIPAL COMPONENTS

by

Teun Kloek

A brief survey, prepared for The New Palgrave:
a dictionary of economic theory and doctrine

Abstract

The theory and practice of principal components are considered both from the point of view of statistical theory and from that of descriptive statistics. Some well known applications are briefly discussed.

PRINCIPAL COMPONENTS

The principal components of a set of m variables are m artificially constructed variables with the following properties. The first component 'explains' as much as possible of the total variance of the original variables. The second has the same property under the additional condition that it is uncorrelated with the first, and so on. It often happens that a few principal components account for a large part of the total variance of the original variables. In such a case one may omit the remaining components. The effect is a substantial reduction of the dimension of the problem. The method is used to explore the relations present in a set of data or to combat the problems created by multicollinearity.

As in linear regression, several approaches are possible. One may view the principal components as the solution to a simple mathematical plane fitting problem, or one may assume a statistical model with an unknown covariance matrix, which is to be estimated. A normality assumption may (but need not) be added, with the consequence that the method of maximum likelihood is available.

If we have a statistical model with an m -vector of random variables ξ with covariance matrix Σ , the k -th principal component can be defined as $\pi_k = \xi' a_k$ where a_k is the eigenvector (characteristic vector) of Σ that corresponds to the k -th eigenvalue (characteristic root, latent root), the eigenvalues λ_k being arranged in descending order

$$(1) \quad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m.$$

If Σ is estimated by S the same operations of taking eigenvectors and eigenvalues are carried out with respect to S . The mathematics of this approach is explained in almost every book on multivariate statistical analysis. A classic in this field is Anderson (1958). In the econometrics literature a detailed account is given in Dhrymes (1970).

A descriptive approach starts with an $n \times m$ matrix X consisting of n observations on each of m variables. Then the principal components P are the columns of $P = XA$ where A is the matrix of eigenvectors of $X'X$. As in (1) the

eigenvectors are always arranged according to the descending order of the eigenvalues. The first principal component p_1 may also be obtained as the result of minimizing the sum of the squares of the residuals E defined as

$$(2) \quad E = X - pa'$$

where p is an n -vector and a an m -vector. This approach to the subject is described in detail by Theil (1971, 1983).

Since both p and a are unknown we need an additional constraint in order to obtain unique results. Most authors choose $a'a = 1$, some $p'p = 1$. The choice is arbitrary and a matter of convenience. Here, it is henceforth assumed that $a'a = 1$, and more generally that $A'A = I$. Another consequence of the fact that both p and a are unknown is that our problem does not have the simple linear structure of least squares regression. Hence the resulting A and P depend (in a non-trivial way) on the origin and scale of the original variables. In the statistical approach the variables are usually measured from their means, in the descriptive approach this is not always the case. If all variables are measured in the same units, there is a natural solution of the problem of the units of measurement. If this is not the case one often chooses the solution to take correlations rather than covariances. (This holds for Σ and S in the statistical approach but it may also be applied to $X'X$ in the descriptive approach.)

Geometrically, the principal components transformation is equivalent to rotating the scatter (in the descriptive approach) or the density (in the statistical approach). Consider the case $m=2$ and suppose that the scatter has the form of an ellipse. Then the principal components transformation is equivalent to rotating the ellipse in such a way that the principal axes of the ellipse coincide with the axes of the coordinate system. Equivalently, one might rotate the coordinate axes in such a way that they coincide with the principal axes of the ellipse. More details on the geometry of principal components are given by Fomby, Hill and Johnson (1984, pp. 287-293).

The main purpose of applying principal components is reduction of the dimension of a data set. The idea originated with Hotelling (1936) and in the present author's opinion it can be interpreted as a mathematician's reaction on Thurstone's (1931) paper on factor analysis. Indeed, Hotelling applies his approach to psychological test scores. Precisely for this type of data the

psychologists developed factor analysis. The main difference between factor analysis and principal components can be given as follows. In factor analysis it is assumed that Σ can be decomposed as:

$$\Sigma = CC' + D$$

where C is an $m \times h$ matrix and D a diagonal matrix of order $m \times m$. If

$$h < \frac{1}{2}[2m + 1 - \sqrt{(8m+1)}]$$

this assumption implies restrictions on the elements of Σ , while the principal components approach does not impose any restrictions on Σ .

A well-known economic example of dimension reduction was given by Stone (1947), who took 17 time-series from the US national accounts in the period 1922-1938. They describe several income and expenditure aggregates relating to consumers, producers and the government. It appeared that in this period the first three principal components accounted for more than 97 per cent of the total variance of these 17 series (the first 80.8 per cent, the second 10.6 per cent, the third 6.1 per cent). The first principal component appears to be highly correlated with total income, the second with the annual change in income, the third with time. It should be emphasized that usually such simple interpretations are not available. More details are given by Stone (1947); also by Theil (1971).

Dimension reduction may also be desirable in the so-called undersized sample problem. Consider a (linear) simultaneous equation model. Suppose one wants to estimate the parameters of a simple structural equation by means of two-stage least-squares or a similar method. Then the first step requires the regressions of the current endogenous variables at the right hand side of the equation on the total set of predetermined variables. This is impossible if $n < m$ (the number of predetermined variables), but it may also have undesirable properties if $n < 2.5 m$, say. In large models, but even in models of medium size, these rules may be violated. Kloek and Mennes (1960) proposed to tackle this problem by replacing the m predetermined variables by a limited number of principal components. For a further discussion and modifications, see Amemiya (1966). The limitations of this approach were discussed by Fisher (1965).

Dimension reduction may also be desirable in more general regressions

with multicollinear explanatory variables. The principal components of these variables can play a very useful role in clarifying the consequences of multicollinearity for the estimates of the regression parameters and their estimated covariance matrix. The case where one eigenvalue (λ_m) is relatively very small is particularly simple. Consider the linear regression model

$$y = X\beta + \epsilon$$

where y is an n -vector containing the observations on the variable to be explained, X an $n \times m$ matrix, as before, containing n observations on each of m explanatory variables, β a vector of unknown parameters to be estimated and ϵ a vector of disturbances, with zero means and covariance matrix $\sigma^2 I$. Let A denote the matrix of eigenvectors of $X'X$ and Λ the diagonal matrix containing the corresponding eigenvalues, then we have $X'X = A\Lambda A'$ with $A'A = I$. Then the inverse satisfies $X'X = A\Lambda^{-1}A'$ and v_{ii} , the i -th diagonal element of the covariance matrix $V = \sigma^2(X'X)^{-1}$, can be written as

$$v_{ii} = \sum_j a_{ij}^2 (1/\lambda_j)$$

where a_{ij} is the typical element of A . So v_{ii} is small if the a_{ij}^2 that correspond to small values of λ_j are small and large in the opposite case. This knowledge is helpful in understanding the problem of multicollinearity. Fomby, Hill and Johnson (1984) give a more extensive treatment and more references.

The next question is whether the relationship between principal components and multicollinearity and principal components can be exploited in order to solve the problems created by multicollinearity. It has been suggested that one might delete a number of principal components. Since the possibility exists that some of the principal components with small variances have a strong influence on the variable to be explained, it cannot be guaranteed that deleting these is a good choice. This may be decided by means of a preliminary test.

When applying the principal components method we transform a set of variables into linear combinations that are uncorrelated. Theil (1976) extends this approach in the context of the Rotterdam model of consumer demand. He constructs linear combinations of commodities that are preference independent and, hence, have a diagonal matrix of price coefficients in his demand system.

In an example (p.287) he transforms beef, pork and chicken into artificial preference independent commodities called inexpensive meat, beef/pork contrast and antichicken. He also gives an example containing clothing, footwear and other goods. His discussion on p. 311 and 312 is illustrative for the interpretation problems that may arise.

In general one may say that principal components have elegant mathematical properties, but that their interpretation in applications is often far from simple.

T. Kloek

References

- Amemiya, T. (1966), On the use of principal components of independent variables in two-stage least squares estimation. Int.Econ.Rev., September.
- Anderson, T.W. (1958). An Introduction to Multivariate Statistical Analysis, New York, Wiley.
- Dhrymes, P.J. (1970). Econometrics, New York, Harper and Row.
- Fisher, F.M. (1965). The choice of instrumental variables in the estimation of economy-wide models. Int.Econ.Rev., September.
- Fomby, T.B., R.C. Hill and S.R. Johnson (1984). Advanced Econometric Methods, New York, Springer.
- Hotelling, H. (1936). Analysis of a Complex of Statistical Variables into Principal Components, J. Educat. Psychol., July.
- Kloek, T. and L.B.M. Mennes (1960). Simultaneous equations estimation based on principal components of predetermined variables, Econometrica, January.
- Stone, J.R.N. (1947). On the interdependence of blocks of transactions. J. Roy. Statist. Soc. B, Supplement.
- Theil, H. (1970). Principles of Econometrics, Amsterdam, North-Holland and New York, Wiley.
- Theil, H. (1976). Theory and Measurement of Consumer Demand, Vol.2, Amsterdam, North-Holland.
- Theil, H. (1983). Mathematical and statistical methods in econometrics, in Z. Griliches and M.D. Intriligator (eds.), Handbook of Econometrics, Vol.1, Amsterdam, North-Holland.
- Thurstone, L.L. (1931), Multiple Factor Analysis. Psychol. Rev. 38, 406-427.

LIST OF REPORTS 1986

- 8600 "Publications of the Econometric Institute Second Half 1985: List of Reprints 415-442, Abstracts of Reports".
- 8601/A T. Kloek, "How can we get rid of dogmatic prior information?", 23 pages.
- 8602/A E.G. Coffman jr, G.S. Lueker and A.H.G Rinnooy Kan, "An introduction to the probabilistic analysis of sequencing and packing heuristics", 66 pages.
- 8603/A A.P.J. Abrahamse, "On the sampling behaviour of the covariability coefficient ζ ", 12 pages.
- 8604/C A.W.J. Kolen, "Interactieve routeplanning van bulktransport: Een praktijktoepassing", 13 pages.
- 8605/A A.H.G. Rinnooy Kan, J.R. de Wit and R.Th. Wijmenga, "Nonorthogonal two-dimensional cutting patterns", 20 pages.
- 8606/A J. Csirik, J.B.G. Frenk, A. Frieze, G. Galambos and A.H.G. Rinnooy Kan, "A probabilistic analysis of the next fit decreasing bin packing heuristic", 9 pages.
- 8607/B R.J. Stroeker and N. Tzanakis, "On certain norm form equations associated with a totally real biquadratic field", 38 pages.
- 8608/A B. Bode and J. Koerts, "The technology of retailing: a further analysis for furnishing firms (II)", 12 pages.
- 8609/A J.B.G. Frenk, M. van Houweninge and A.H.G. Rinnooy Kan, "Order statistics and the linear assignment problem", 16 pages.
- 8610/B J.F. Kaashoek, "A stochastic formulation of one dimensional pattern formation models", 17 pages.
- 8611/B A.G.Z. Kemna and A.C.F. Vorst, "The value of an option based on an average security value", 14 pages.
- 8612/A A.H.G. Rinnooy Kan and G.T. Timmer, "Global optimization", 47 pages.
- 8613/A A.P.J. Abrahamse and J.Th. Geilenkirchen, "Finite-sample behaviour of logit probability estimators in a real data set", 25 pages.
- 8614/A L. de Haan and S. Resnick, "On regular variation of probability densities", 17 pages
- 8615 "Publications of the Econometric Institute First Half 1986: List of Reprints 443-457, Abstracts of Reports".

- 8616/A W.H.M. van der Hoeven and A.R. Thurik, "Pricing in the hotel and catering sector", 22 pages.
- 8617/A B. Nootboom, A.J.M. Kleijweg and A.R. Thurik, "Normal costs and demand effects in price setting", 17 pages.
- 8618/A B. Nootboom, "A behavioral model of diffusion in relation to firm size", 36 pages.
- 8619/A J. Bouman, "Testing nonnested linear hypotheses II: Some invariant exact tests", 185 pages.
- 8620/A A.H.G. Rinnooy Kan, "The future of operations research is bright", 11 pages.
- 8621/A B.S. van der Laan and J. Koerts, "A logit model for the probability of having non-zero expenses for medical services during a year", 22 pages
- 8622/A A.W.J. Kolen, "A polynomial algorithm for the linear ordering problem with weights in product form", 4 pages.
- 8623/A A.H.G. Rinnooy Kan, "An introduction to the analysis of approximation algorithms", 14 pages.
- 8624/A S.R. Wunderink-van Veen and J. van Daal, "The consumption of durable goods in a complete demand system", 34 pages.
- 8625/A H.K. van Dijk, J.P. Hop and A.S. Louter, "An algorithm for the computation of posterior moments and densities using simple importance sampling", 59 pages.
- 8626/A N.L. van der Sar, B.M.S. van Praag and S. Dubnoff, "Evaluation questions and income utility", 19 pages.
- 8627/C G. Renes, A.J.M. Hagenars and B.M.S. van Praag, "Perceptie en realiteit op de arbeidsmarkt", 18 pages.
- 8628/C B.M.S. van Praag and M.E. Homan, "Lange en korte termijn inkomens-elasticiteiten", 16 pages.
- 8629/A R.C.J.A. van Vliet and B.M.S. van Praag, "Health status estimation on the basis of mimic health care models", 32 pages.
- 8630/A K.M. van Hee, B. Huitink and D.K. Leegwater, "Portplan, a decision support system for port terminals", 25 pages.
- 8631/A D.K. Leegwater, "Economical effects of delay and acceleration of (un)loading multipurpose ships for stevedore firms", 18 pages.

- 8632/A **A.M. Wesselman and B.M.S. van Praag**, "Elliptical regression operationalized", 10 pages.
- 8633/C **R.C.J.A. van Vliet and E.K.A. van Doorslaer**, "De relatie tussen ziekenhuiscapaciteit en -gebruik: een analyse van de gevolgen van aggregatie", 81 pages.
- 8634/B **J. Brinkhuis**, "Normal integral bases and complex conjugation", 19 pages.
- 8635/A **B.S. van der Laan, J. Koerts and J. Reichardt**, "A statistical model for the expenses for medical services during a year", 39 pages.
- 8636/A **B.M.S. van Praag and A.M. Wesselman**, "Elliptical multivariate analysis", 17 pages.
- 8637/A **R.H. Byrd, C.L. Dert, A.H.G. Rinnooy Kan and R.B. Schnabel**, "Concurrent stochastic methods for global optimization", 40 pages.
- 8638/A **B.S. van der Laan**, "An econometric model for the costs of claims of passenger car traffic accidents in the Netherlands", 24 pages.
- 8639/A **P.M.C. de Boer**, "An algorithm for maximum likelihood estimation of a new covariance matrix specification for sum-constrained models", 32 pages.
- 8640/A **H.W.J.M. Trienekens**, "Parallel branch and bound on an MIMD system", 26 pages.
- 8641/A **N.L. van der Sar, B.M.S. van Praag and S. Dubnoff**, "Evaluation questions and income utility", 18 pages.
- 8642/A **B.M.S. van Praag and N.L. van der Sar**, "Household cost functions and equivalence scales", 29 pages. (a revised version of report 8527/A).
- 8643/A **B.M.S. van Praag and N.L. van der Sar**, "Social distance on the income dimension", 25 pages.
- 8644/A **T. Kloek**, "Principal components", 6 pages.

A. Economics, Econometrics and Operations Research

B. Mathematics

C. Miscellaneous

