Introduction to Principal Components and FactorAnalysis

Multivariate Analysis often starts out with data involving a substantial number of correlated variables.

Principal Component Analysis (PCA) is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set.

- Principal component analysis (PCA) is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called *principal components*.

- The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.

- Principal components analysis is similar to another multivariate procedure called Factor Analysis. They are often confused and many scientists do not understand the difference between the two methods or what types of analyses they are each best suited.

- Traditionally, principal component analysis is performed on a square symmetric matrix.

- It can be a SSCP matrix (pure sums of squares and cross products), Covariance matrix (scaled sums of squares and cross products), or Correlation matrix (sums of squares and cross products from standardized data).

- The analysis results for objects of type SSCP and Covariance do not differ, since these objects only differ in a global scaling factor.

- A correlation matrix is used if the variances of individual variates differ much, or if the units of measurement of the individual variates differ.

Geometry of Principal Components

Campbell and Atchley

Syst. Zool., 30(3), 1981, pp. 268–280

A principal component analysis can be considered as a rotation of the axes of the original variable coordinate system to new orthogonal axes, called principal axes, such that the new axes coincide with directions of maximum variation of the original observations. Consider the line or axis passing through the ends of the elliptical cluster of points in Figure 1. Project the original data points onto this axis. The point y_{1m} is the projection of the point (x_{1m}, x_{2m}) onto the axis defined by the direction Y_1. This axis has the property that the variance of the projected points y_{1m}, $m = 1, \ldots, n$, is greater than the variance of the points when projected onto any other line or axis passing through $(\bar{x}_1, \bar{x}_2)$. Any line parallel to Y_1 also has the property of maximum variance of the projected points. It is however convenient geometrically to use the first representation.

The property of maximum variation of the projected points defines the first principal axis; it is the line or direction with maximum variation of the projected values of the original data points. The projected values corresponding to this direction of maximum variation are the *principal component scores*. The first principal axis is often called the line of best fit since the sum of squares (SSQ) of the perpendicular deviations of the original data points from the line is a minimum. Successive principal axes are determined with the property that they are orthogonal to the previous principal axes and that they maximize the variation of the projected points subject to these constraints.

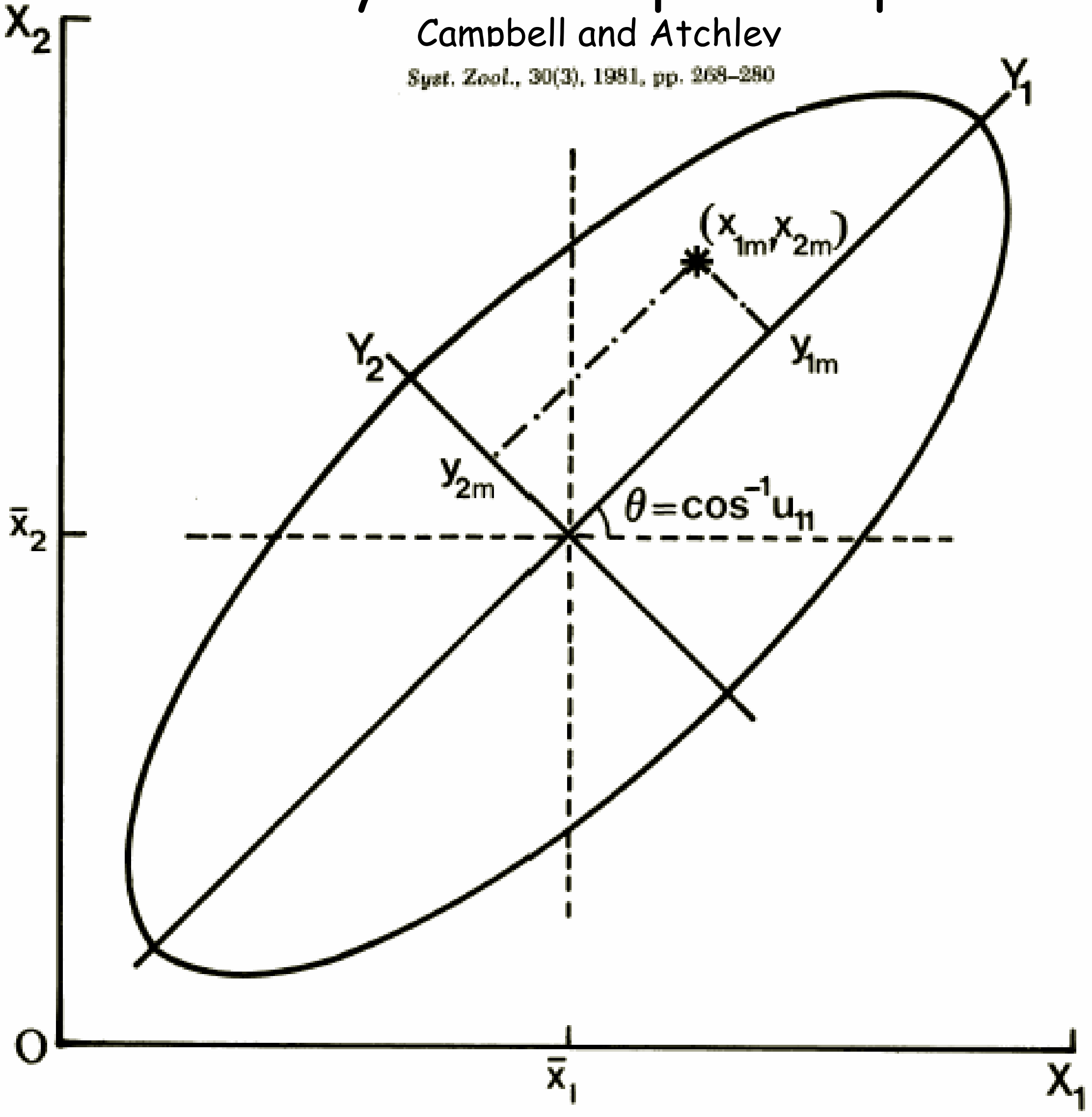


FIG. 1.—Idealized representation of scatter diagram for two variables, showing the mean for each variable ($\bar{x}_1$ and $\bar{x}_2$), 95% concentration ellipse, and principal axes Y_1 and Y_2. The points y_{1m} and y_{2m} give the principal component scores for the observation $\mathbf{x}_1 = (x_{1m}, x_{2m})^T$. The cosine of the angle θ between Y_1 and X_1 gives the first component u_{11} of the eigenvector corresponding to Y_1.

Objectives of principal component analysis

- PCA reduces attribute space from a larger number of variables to a smaller number of factors and as such is a "non-dependent" procedure (that is, it does not assume a dependent variable is specified).

- PCA is a dimensionality reduction or data compression method. The goal is dimension reduction and there is no guarantee that the dimensions are interpretable (a fact often not appreciated by (amateur) statisticians).

- To select a subset of variables from a larger set, based on which original variables have the highest correlations with the principal component.

Factor Analysis & Principal Components
Definitions for Beginners

Principal component analysis: Factor model in which the factors are based on summarizing the total variance. With PCA, unities are used in the diagonal of the correlation matrix computationally implying that all the variance is common or shared. Algorithm lacking underlying model.

Common factor analysis: Factor model explores a reduced correlation matrix. That is, communalities (r^2) are inserted on the diagonal of the correlation matrix, and the extracted factors are based only on the common variance, with specific and error variances excluded**. Explores underlying "latent" structure of data. Model assumes variability partitionable into common and unique components**

Common variance: Variance shared with other variables in the factor analysis.

Specific or unique variance: Variance of each variable unique to that variable and not explained or associated with other variables in the factor analysis**.**

Communality: Total amount of variance an original variable shares with all other variables included in the analysis.

Eigenvalue: Column sum of squared loadings for a factor, i.e., the latent root. It conceptually represents that amount of variance accounted for by a factor.

FAfB – continued - 2

Sphericity test: Statistical test for the overall significance of all correlations within a correlation matrix

Factor: Linear combination (variate) of the original variables. Factors also represent the underlying dimensions (constructs) that summarize or account for the original set of observed variables.

Factor loadings: Correlation between the original variables and the factors, and the key to understanding the underlying nature of a particular factor. Squared factor loadings indicate what percentage of the variance in an original variable is explained by a factor.

Factor matrix: Table displaying the factor loadings of all variables on each factor.

Factor score: Composite measure created for each observation on each factor extracted in the factor analysis. The factor weights are used in conjunction with the original variable values to calculate each observation's score. The factor scores are standardized to reflect a z-score. Factor scores place each variable in a plane of multivariate variability.

Principal components analysis (PCA): PCA seeks a linear combination of variables such that the maximum variance is extracted from the variables.

It then removes this variance and seeks a second linear combination which explains the maximum proportion of the remaining variance, and so on. This is called the principal axis method and results in orthogonal (uncorrelated) factors. PCA analyzes total (common and unique) variance.

Eigenvectors: *Principal components* (from PCA - principal components analysis) reflect both common and unique variance of the variables and may be seen as a variance-focused approach seeking to reproduce both the total variable variance with all components and to reproduce the correlations.

PCA is far more common than PFA, however, and it is common to use "factors" interchangeably with "components."

The principal components are linear combinations of the original variables weighted by their contribution to explaining the variance in a particular orthogonal dimension

Eigenvalues: Also called *characteristic roots*. The eigenvalue for a given factor measures the variance in all the variables which is accounted for by that factor.

The ratio of eigenvalues is the ratio of explanatory importance of the factors with respect to the variables. If a factor has a low eigenvalue, then it is contributing little to the explanation of variances in the variables and may be ignored as redundant with more important factors.

Eigenvalues measure the amount of variation in the total sample accounted for by each factor.

A factor's eigenvalue may be computed as the sum of its squared factor loadings for all the variables.

Note that the eigenvalues associated with the unrotated and rotated solution will differ, though their total will be the same.

Factor loadings (factor or component coefficients) : The factor loadings, also called component loadings in PCA, are the correlation coefficients between the variables (rows) and factors (columns).

Analogous to Pearson's r, the squared factor loading is the percent of variance in that variable explained by the factor.

To get the percent of variance in all the variables accounted for by each factor, add the sum of the squared factor loadings for that factor (column) and divide by the number of variables. (Note the number of variables equals the sum of their variances as the variance of a standardized variable is 1.) This is the same as dividing the factor's eigenvalue by the number of variables.

PC scores: Also called *component scores* in PCA, these scores are the scores of each case (row) on each factor (column). To compute the factor score for a given case for a given factor, one takes the case's standardized score on each variable, multiplies by the corresponding factor loading of the variable for the given factor, and sums these products.

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	19.2196613	7.497092	0.3559	0.3559
2	11.7225694	3.7876295	0.2171	0.573
3	7.9349399	3.1810705	0.1469	0.7199
4	4.7538694	2.5056124	0.088	0.808
5	2.248257	0.2362884	0.0416	0.8496
6	2.0119686	0.498058	0.0373	0.8869
7	1.5139106	0.3555246	0.028	0.9149

Portion of PCA Analysis of 54 Amino Acid Physio-Chemical Attributes

Variable	PCA1	PCA2	PCA3	PCA4
MDS3	0.130	-0.007	0.243	-0.093
MDS16	0.171	-0.110	0.174	-0.005
MDS17	-0.104	0.228	0.105	-0.001
MDS19	0.146	0.196	0.059	0.121
MDS24	-0.188	-0.003	0.089	0.106
MDS25	0.203	-0.073	0.114	-0.037
MDS29	0.105	-0.064	0.266	-0.103
MDS30	-0.073	0.239	0.097	0.092
MDS31	0.084	-0.137	-0.178	0.233
MDS32	0.097	-0.020	-0.172	-0.031
MDS36	-0.066	0.239	0.128	-0.076
MDS43	0.117	0.193	0.054	0.145
MDS44	0.061	0.174	0.018	0.161
MDS45	0.089	0.038	-0.156	-0.242

PCA Coefficient

Introduction to Factor Analysis

Factor analysis is a statistical procedure to identify interrelationships that exist among a large number of variables, i.e., to identify how suites of variables are related.

Factor analysis can be used for exploratory or confirmatory purposes.

As an exploratory procedure, factor analysis is used to search for a possible underlying structure in the variables. In confirmatory research, the researcher evaluates how similar the actual structure of the data, as indicated by factor analysis, is to the expected structure.

The major difference between exploratory and confirmatory factor analysis is that researcher has formulated hypotheses about the underlying structure of the variables when using factor analysis for confirmatory purposes.

As an exploratory tool, factor analysis doesn't have many statistical assumptions. The only real assumption is presence of relatedness between the variables as represented by the correlation coefficient. If there are no correlations, then there is no underlying structure.

Steps in conducting a factor analysis

There are five basic factor analysis steps:

- data collection and generation of the correlation matrix
- partition of variance into common and unique components (unique may include random error variability)
- extraction of initial factor solution
- rotation and interpretation
- construction of scales or factor scores to use in further analyses

Factor Analysis & Principal Components
Definitions for Beginners

Principal component analysis: Factor model in which the factors are based on the total variance. With PCA, unities are used in the diagonal of the correlation matrix computationally implying that all the variance is common or shared.

Common factor analysis: Factor model explores a reduced correlation matrix. That is, communalities (r^2) are inserted on the diagonal of the correlation matrix, and the extracted factors are based only on the common variance, with specific and error variances excluded**.**

Common variance: Variance shared with other variables in the factor analysis.

Specific or unique variance: Variance of each variable unique to that variable and not explained or associated with other variables in the factor analysis**.**

Communality: Total amount of variance an original variable shares with all other variables included in the analysis.

Eigenvalue: Column sum of squared loadings for a factor; = the latent root. It conceptually represents that amount of variance accounted for by a factor.

FFB – continued - 2

Sphericity test: Statistical test for the overall significance of all correlations within a correlation matrix

Factor: Linear combination (variate) of the original variables. Factors also represent the underlying dimensions (constructs) that summarize or account for the original set of observed variables.

Factor loadings: Correlation between the original variables and the factors, and the key to understanding the nature of a particular factor. Squared factor loadings indicate what percentage of the variance in an original variable is explained by a factor.

Factor matrix: Table displaying the factor loadings of all variables on each factor.

Factor score: Composite measure created for each observation on each factor extracted in the factor analysis. The factor weights are used in conjunction with the original variable values to calculate each observation's score. The factor scores are standardized to according to a z-score.

FFB – continued - 3

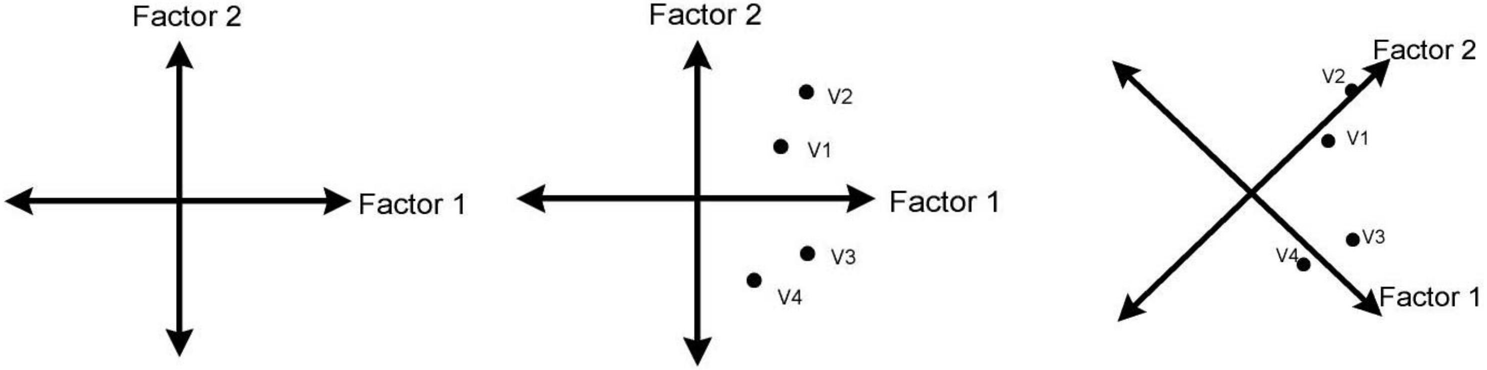
Factor rotation: Process of manipulation or adjusting the factor axes to achieve a simpler and pragmatically more meaningful factor solution.

Oblique factor rotation: Factor rotation computed so that the extracted factors are correlated. Rather than arbitrarily constraining the factor rotation to an orthogonal (90 degree angle) solution, the oblique solution identifies the extent to which each of the factors are correlated.

Orthogonal factor rotation: Factor rotation in which the factors are extracted so that their axes are maintained at 90 degrees. Each factor is independent of, or orthogonal to, all other factors. The correlation between teh factors is determined to be zero.

VARIMAX: One of the most popular orthogonal factor rotation methods.

Factor Rotation



Unrotated Axes

Rotated Axes

Each variable lies somewhere in the plane formed by these two factors. The factor loadings, which represent the correlation between the factor and the variable, can also be thought of as the variable's coordinates on this plane.

In unrotated factor solution the Factor "axes" may not line up very well with the pattern of variables and the loadings may show no clear pattern. Factor axes can be rotated to more closely correspond to the variables and therefore become more meaningful. ***Relative relationships between variables are preserved***.

The rotation can be either orthogonal or oblique

Rotation of Factors to “Simple Structure”

Variable	Factor 1	Factor 2
1	.31	.19
2	.44	.39
3	.40	-.21
4	.18	-.30

Variable	Factor 1	Factor 2
1	.04	.32
2	.02	.54
3	.47	.12
4	.32	-.06

Factors subjected to Varimax orthogonal rotation to simple structure.

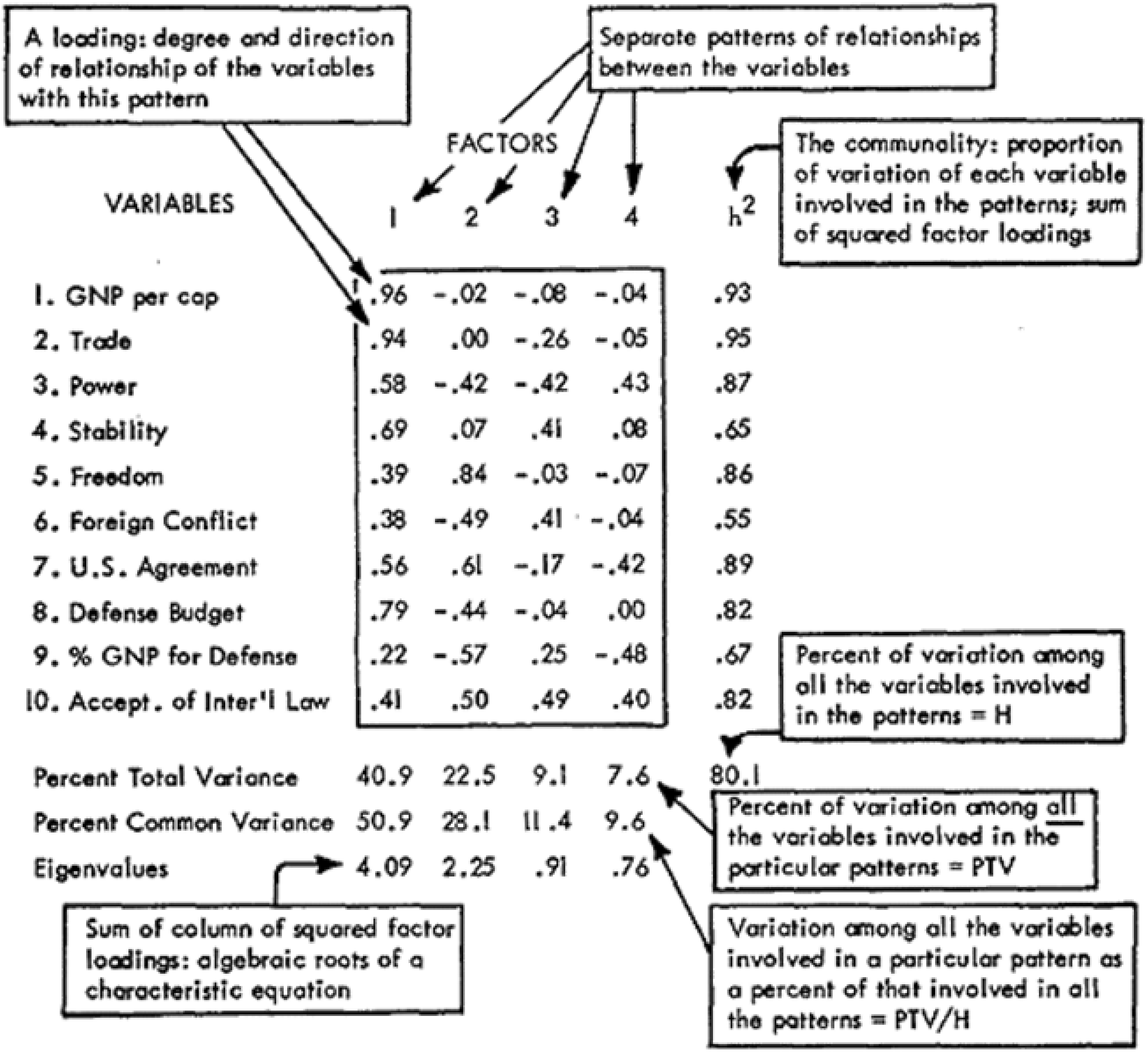
Simple structure attempts to clarify the relationships among variables by producing factors with either very high or very low coefficients and variables with high coefficients on only one variable.

Rotation to simple structure generally simplifies the relationships among the variables and clarifies the interpretation of the factors.

TABLE 2
Correlation Matrix,* Selected Sample Data

Characteristic	1	2	3	4	5	6	7	8	9	10
(1) GNP per capita	.97									
(2) Trade	.93	.97								
(3) Power	.55	.66	.89							
(4) Stability	.62	.55	.25	.63						
(5) Freedom of opposition	.31	.40	-.10	.32	.91					
(6) Foreign conflict	.36	.30	.25	.46	-.32	.61				
(7) U.S. agreement	.58	.59	-.07	.36	.74	-.11	.89			
(8) Defense budget	.79	.71	.66	.49	-.07	.38	.18	.90		
(9) % GNP for defense	.17	.17	.06	.15	-.28	.44	-.11	.47	.73	
(10) International law acceptan	.34	.22	-.02	.56	.57	.04	.24	.14	-.24	.82

* These are product moment correlation coefficients. The data for these characteristics are given in Table 6-1. Elements in the principal diagonal are the squared multiple correlation coefficient of the variable with all the others. Excluding the diagonal, the signs on the columns of correlations in column 6 and 7 had been mistakenly reversed in this table in the original article, and in Table 6-2 (p. 136) of Applied Factor Analysis, and what was posted on this web site up to June 6 2001. The resulting factor analyses are correct, however, since they are based on the correlations calculated with the computer program employed.



A loading: degree and direction of relationship of the variables with this pattern

Separate patterns of relationships between the variables

The communality: proportion of variation of each variable involved in the patterns; sum of squared factor loadings

	FACTORS				
VARIABLES	1	2	3	4	h^2
1. GNP per cap	.96	-.02	-.08	-.04	.93
2. Trade	.94	.00	-.26	-.05	.95
3. Power	.58	-.42	-.42	.43	.87
4. Stability	.69	.07	.41	.08	.65
5. Freedom	.39	.84	-.03	-.07	.86
6. Foreign Conflict	.38	-.49	.41	-.04	.55
7. U.S. Agreement	.56	.61	-.17	-.42	.89
8. Defense Budget	.79	-.44	-.04	.00	.82
9. % GNP for Defense	.22	-.57	.25	-.48	.67
10. Accept. of Inter'l Law	.41	.50	.49	.40	.82
Percent Total Variance	40.9	22.5	9.1	7.6	80.1
Percent Common Variance	50.9	28.1	11.4	9.6	
Eigenvalues	4.09	2.25	.91	.76	

Percent of variation among all the variables involved in the patterns = H

Percent of variation among all the variables involved in the particular patterns = PTV

Sum of column of squared factor loadings: algebraic roots of a characteristic equation

Variation among all the variables involved in a particular pattern as a percent of that involved in all the patterns = PTV/H

FIG. 6. Unrotated factor matrix (diagrammed) from data in Table 1. (Principal axes technique. Factoring stopped at eigenvalues less than .50.)

TABLE 3
FACTOR MATRICES,[*] SELECTED SAMPLE DATA

										Primary oblique factors							
	Unrotated factors[a]					Orthogonally rotated factors[b]				Pattern factors[c]				Structure factors[c]			
Variables	1	2	3	4	h^2	1	2	3	4	1	2	3	4	1	2	3	4
1. GNP per capita	(.96)	−.02	−.08	−.04	.93	(.73)	.47	.29	.30	(.64)	.46	.21	.19	(.78)	(.57)	.42	.46
2. Trade	(.94)	.00	−.26	−.05	.95	(.79)	(.51)	.19	.16	(.73)	(.52)	.07	.04	(.81)	(.60)	.33	.34
3. Power	(.58)	−.42	−.42	.43	.87	(.92)	−.17	−.03	−.01	(.98)	−.15	−.17	−.07	(.90)	−.08	.14	.08
4. Stability	(.69)	.07	.41	.08	.65	.32	.25	.34	(.63)	.17	.16	.31	(.60)	.40	.35	.39	(.69)
5. Freedom	.39	(.84)	−.03	−.07	.86	−.02	(.77)	−.34	.40	−.07	(.73)	−.34	.32	−.04	(.81)	−.34	.49
6. Foreign conflict	.38	−.49	.41	−.04	.55	.25	−.19	(.64)	.23	.12	−.23	(.62)	.26	.35	−.14	(.67)	.25
7. US agreement	(.56)	(.61)	−.17	−.42	.89	.13	(.93)	−.03	.11	.05	(.94)	−.03	−.01	.12	(.94)	−.01	.27
8. Defense budget	(.79)	−.44	−.04	.00	.82	(.75)	.10	.48	.12	(.67)	.09	.39	.06	(.82)	.17	(.61)	.24
9. % GNP for defense	.22	(−.57)	.25	−.48	.67	.07	−.03	(.77)	−.17	−.05	.00	(.82)	−.18	.17	−.05	(.79)	−.15
10. Accept. IR law	.41	.50	.49	.40	.82	.03	.18	−.13	(.87)	−.06	.06	−.15	(.89)	.08	.30	−.13	(.89)
Percent total variance	40.9	22.5	9.1	7.6	80.1	27.6	21.0	16.2	15.3								
Percent common variance	50.9	28.1	11.4	9.6		34.7	26.5	20.4	19.4								
									Sum of squares	2.41	2.01	1.52	1.39				

[*] Loadings greater than an absolute value of .50 shown in parentheses.

[a] From Figure 6.

[b] Varimax rotation.

[c] Biquartimin rotation at 12 major cycles and 712 iterations.

TABLE 5
Factor Correlations, Selected Sample Data

	Factors			
Factors	1	2	3	4
1. Power	1.00			
2. US Agreement	.09	1.00		
3. Foreign Conflict	.31	.00	1.00	
4. Accept. of Inter'l Law	.20	.28	.04	1.00

TABLE 6

Selected Sample Factor Scores*

	Orthogonally rotated factors			
	1	2	3	4
Nations	(Power)	(Agree US)	(For. conflict)	(Inter'l law)
Brazil	– .389	1.053	–1.227	–1.070
Burma	– .584	.010	– .097	– .955
China	.325	–1.601	– .083	– .641
Cuba	– .662	.859	– .325	–1.183
Egypt	– .716	– .761	.448	1.331
India	.182	– .639	–1.807	.909
Indonesia	.027	– .480	– .712	– .757
Israel	–1.275	.518	.097	1.897
Jordan	–1.577	.426	2.018	– .719
Netherlands	–0.315	.570	– .638	1.292
Poland	.410	–1.382	– .296	– .287
USSR	1.129	–1.336	1.726	– .304
UK	1.081	1.178	– .024	– .404
US	2.365	1.586	.919	.906

* These are standardized regression estimates.

Summary of Factor (not PCA) Analyses

Correlation Matrix

Amino acid attributes

$$\begin{array}{llll} c_1 & & & \\ r_{12} & c_2 & & \\ r_{13} & r_{23} & c_3 & \\ r_{1n} & r_{2n} & r_{3n} & c_n \end{array}$$

Factor matrix

$$\begin{array}{llll} & \lambda_I & \lambda_{II} & \lambda_{III} \\ \mathit{Ala} & X_{1\,I} & X_{1\,II} & X_{1\,III} \\ \mathit{Arg} & X_{2\,I} & X_{2\,II} & X_{2\,III} \\ \mathit{Asn} & X_{3\,I} & X_{3\,II} & X_{3\,III} \\ \mathit{Val} & X_{n\,I} & X_{n\,II} & X_{n\,III} \end{array}$$

Factor scores

Ala	-0.158	-1.385	0.205
Arg	1.096	0.726	2.467
Asn	1.383	-0.340	-1.065

Transform sequences

C-MYC **ALRDQIPELE**
L-MYC **ALRDQVPTLA**
N-MYC **TLRDHVPELV**

1. Decompose variances and covariances
2. Manova and discriminant analysis
3. Model amino acid and protein behaviors

Portion of factor pattern matrix of 54 amino acid attributes.

Amino Acid Attributes	F1	F2	F3	F4	F5	Comm
Average non-bonded energy per atom	1.028	0.074	0.152	0.047	-0.079	0.982
Percentage of exposed residues	1.024	0.016	0.194	0.095	0.025	0.965
Average accessible surface area	1.005	-0.034	0.159	0.059	0.153	0.994
Residue accessible surface area	0.950	0.098	0.178	0.039	0.237	0.961
Number of hydrogen bond donors	0.809	0.021	0.122	0.021	0.357	0.808
Polarity	0.790	-0.044	-0.388	0.027	-0.092	0.956
Hydrophilicity value	0.779	-0.153	-0.333	0.213	0.023	0.862
Polar requirement	0.775	-0.128	-0.335	-0.020	-0.245	0.939
Long range non-bonded energy	0.725	-0.024	-0.394	0.189	-0.104	0.905
Negative charge	0.451	-0.218	-0.024	-0.052	-0.714	0.737
Positive charge	0.442	-0.246	-0.225	-0.085	0.708	0.730
Size	0.440	-0.112	0.811	-0.144	0.108	0.915
Normalized relative frequency of bend	0.435	0.674	-0.225	0.082	-0.118	0.912
Normalized frequency of beta-turn	0.416	0.648	-0.346	-0.019	-0.079	0.969
Molecular weight	0.363	-0.091	0.657	-0.504	-0.047	0.923
Relative mutability	0.337	-0.172	-0.183	0.297	-0.296	0.416
Normalized frequency of coil	0.271	0.863	0.028	0.123	0.073	0.860
Average volume of buried residue	0.269	-0.153	0.766	-0.340	0.016	0.928
Conformational parameter of beta-turn	0.243	0.693	-0.185	-0.439	0.078	0.837
Residue volume	0.225	-0.172	0.794	-0.292	0.036	0.946
Isoelectric point	0.224	-0.060	-0.049	0.163	0.967	0.955
Propensity to form reverse turn	0.224	-0.005	-0.433	0.319	-0.194	0.563
Chou-Fasman coil conformation	0.201	0.780	-0.338	-0.052	0.048	0.948

Factors describe latent structure of amino acid attributes