# On the Applicability of Truncated Component Analysis Based on Correlation Coefficients for Nominal Scales

**Svante Janson and Jan Vegelius**
**University of Uppsala**

The possibility of using component analysis for nominal data is discussed. Particularly, two nominal scale correlation coefficients are applicable, namely, Tschuprow's coefficient and the J index. The reason is that they are $E$-correlation coefficients; that is, they satisfy the requirements of a scalar product between normalized vectors in a Euclidean space. Some characteristics of these coefficients are described. The contingency coefficient and Cramér's $V$ are shown not to be applicable in a component analysis. An example of a truncated component analysis on artifical nominal data is included with both the J index and Tschuprow's coefficient.

One frequently used vector-algebra-based method in psychological research is truncated component analysis (see Gorsuch, 1974). By using some kind of similarity measure between test items as the scalar product between the normalized vectors corresponding to the items, a Euclidean space is created. Certain choices of base vectors of a linear subspace will then be interpreted as the main factors of the item structure. Not all similarity measures are applicable. In order to clarify this problem, Vegelius (1973, 1976a, in press) introduced the concept *E-(correlation) coefficient*, where $E$ stands for Euclidean.

An $E$-coefficient is a similarity measure which satisfies the requirements of a scalar product between normalized vectors in a Euclidean space. From the definition of a scalar product, it is possible to deduce that the following conditions must be fulfilled for all variables $u$, $v$, and $w$ in order for a similarity measure $r$ to be an $E$-coefficient (Shilov, 1961):

1. $r(u,v) = r(v,u)$
2. $|r(u,v)| \leqslant 1$
3. $r(u,u) = 1$
4. If $r(v,w) = -1$, then $r(u,v) = -r(u,w)$
5. If $r(v,w) = 1$, then $r(u,v) = r(u,w)$
6. If $r(u,v) = 1$ and $r(u,w) = 1$, then $r(v,w)$ must also be equal to 1 (transitive relation)
7. A correlation matrix based on $r$ must be non-negative definite

For interval, ordinal, and dichotomized scales there exist many types of coefficients that are $E$-coefficients, e.g., the product-moment correlation coefficient, Kendall's $\tau$, and the G index.

*APPLIED PSYCHOLOGICAL MEASUREMENT*
*Vol. 2, No. 1 Winter 1978 pp. 135–145*

## Traditional Nominal Scale Correlation Coefficients

The most common correlation measure for nominal scales is the contingency coefficient. It was first presented by Pearson (1904), who called it the first coefficient of contingency. It is defined by

$$r_c(v_k,v_p) = \sqrt{\frac{x^2}{m + x^2}} \qquad [1]$$

where $v_k$ and $v_p$ represent the two variables,
$m$ = number of units (individuals) and
$\chi^2$ = the $\chi^2$ value, computed from the contingency table.

Pearson assumed that the variable pair in question has a bivariate normal distribution. Both variables are then divided into a finite number of parts, so that the entire variable plane is split into a finite number of rectangles (of finite or infinite area). If a sample of the variable pair population is drawn, it will be split into a contingency table by this division. Pearson showed that if the rectangles are made infinitely small, the limit of the expected value of the contingency coefficient will be equal to the absolute value of the product-moment correlation coefficient between the original continuous variables.

The contingency coefficient is not an $E$-correlation coefficient. One reason is that $r_c(v_k,v_p)$ can never be equal to 1. In Appendix 2 it is further shown that a correlation matrix based on a contingency table is not necessarily non-negative definite. Therefore, the contingency coefficient should not be used in a component analysis.

Tschuprow (1925, 1939) introduced another alternative based on the $\chi^2$ value of the contingency table. It is defined by

$$r_T(v_k,v_p) = \frac{x^2}{m \cdot \sqrt{(c_k-1)\cdot(c_p-1)}} \qquad [2]$$

where $c_k$ = number of categories of variable $v_k$ and
$c_p$ = number of categories of variable $v_p$.

Janson and Vegelius (1977) proved that this coefficient is an $E$-correlation coefficient.

As Tschuprow's coefficient can be 1 only for a square contingency table, Cramér (1946) suggested a modification called $V$, which is defined as

$$r_V(v_k,v_p) = \frac{x^2}{m \cdot \min(c_k-1, c_p-1)} \qquad [3]$$

$V$ can be maximum (=1) also for a non-square table.

However, Cramér's $V$ is not an $E$-coefficient, as perfect correlation is not a transitive relation (shown in Appendix 3). Therefore, Tschuprow's coefficient will be considered as a superior version in this connection.

Other coefficients with more limited areas of application have also been proposed, e.g., kappa (Cohen, 1960) and weighted kappa (Cohen, 1968; Fleiss & Cohen, 1973). As they cannot be used for an arbitrary pair of variables measured on a nominal scale, they will not be further discussed here.

## J index: Definition and Characteristics

Recently a new correlation coefficient for nominal scales was introduced under the designation the J index (Janson & Vegelius, 1977; Vegelius, 1976b). It was defined as a special case of Kendall's general definition of a correlation coefficient (Kendall, 1948). The following denotations were used:

$m$ = number of individuals
$y_{ik}$ = original value of individual $i$ on variable $k$ as obtained by measurement
$c_k$ = number of levels of variable $k$ ( it is assumed that $c_k$ is at least 2)
$x_{ijk}$ = value of individual pair $(i,j)$ on variable $k$, somehow defined as a function of the $y$-values (and similarly for variable $p$).

Kendall's coefficient between the variables $k$ and $p$ is defined as

$$r_{Kg}(v_k,v_p) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} x_{ijk}\cdot x_{ijp}}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m} x_{ijk}^2}\cdot\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m} x_{ijp}^2}} \qquad [4]$$

Different definitions of the $x$-values will yield different correlation coefficients.

The J index is now defined as the special case of Kendall's coefficient, where

$$x_{ijk} = \begin{cases} c_k - 1, & \text{if } y_{ik} = y_k \\ -1, & \text{if } y_{ik} \neq y_{jk} \end{cases} \qquad [5]$$

and similarly for variable $p$.

Janson and Vegelius (1977) proved some theorems about the J index. The main conclusions are listed below:

1. The J index is an $E$-coefficient. Consequently it satisfies all the requirements mentioned in the introduction.
2. The range of variation of the J index is the closed interval [0,1].
3. The J index will be equal to its maximal value 1 if, and only if, one of the following cases is true:
   a. all values are equal in each of the two variables (see Table 7 in Appendix 1);
   b. the table is square, and in each row and each column there is at most one frequency which is not zero (see Table 8 in Appendix 1).
4. The J index will be equal to its minimal value 0 if, and only if, the number of units in each cell is a sum of a fixed row index and a fixed column index (see Table 9 in Appendix 1).
5. Removal of an empty category will affect the J index (see Tables 10 and 8 in Appendix 1).
6. For dichotomized variables the J index equals the square of the G index (Holley & Guilford, 1964).

## Tschuprow's Coefficient: Definition and Characteristics

Janson and Vegelius (1977) have shown that Tschuprow's coefficient is a special case of Kendall's general correlation coefficient (Kendall, 1948). If we let $m_{kr}$ = number of individuals who have the

value $r$ in variable $k$, and use the denotation that was defined in the previous section, Equation 4 can be applied with the $x$ values defined as

$$x_{ijk} = \begin{cases} \frac{m}{m_{kr}} - 1, & y_{ik} = y_{jk} = k_r \\ -1, & \text{if } y_{ik} \neq y_{jk} \end{cases} \qquad [6]$$

Then the Kendall coefficient will equal Tschuprow's coefficient.

The following characteristics can be deduced for Tschuprow's coefficient (cf. Tschuprow, 1925; Yule & Kendall, 1953; Guilford, 1956; Siegel, 1956).

1. Tschuprow's coefficient is an $E$-coefficient.
2. The range of variation for Tschuprow's coefficient is the closed interval [0,1].
3. Tschuprow's coefficient is equal to its maximum value 1 if, and only if, the contingency table is square and in each row and each column exactly one frequency is positive (see Table 8 in Appendix 1).
4. Tschuprow's coefficient is equal to its minimal value 0 if, and only if, the frequency in each cell is the product of a fixed positive row index and a fixed positive column index (see Table 12 in Appendix 1).
5. If one category of one variable is empty, it must be omitted in order for Tschuprow's coefficient to be defined (see Table 10 in Appendix 1).
6. For dichotomized variables, Tschuprow's coefficient is equal to the square of the phi-coefficient.

## Component Analysis for Nominal Scale Variables

The J index and Tschuprow's coefficient are $E$-correlation coefficients. From a mathematical point of view, there is consequently no objection to using them in component analysis. As their range of variation only includes the closed interval [0,1], and no negative values, the entire $n$-dimensional Euclidean space will not be used.

Some researchers may find it strange to base a component analysis on nominal data, since the Euclidean space has metric properties. However, it must be pointed out that the existence of distances (and angles) in a Euclidean space depends on the scalar product chosen. If a scalar product is chosen that is relevant for nominal scale, metric properties will be obtained in the Euclidean space.

## Example

The following is an example of a truncated components analysis based on the J index (and on Tschuprow's coefficient), using artificial data:

Twenty Swedish persons were assumed to have answered various preference questions with four al-

ternatives on a nominal scale. The three main correlation measures for nominal scales were applied to the data by the CONTIN program.[1] As the contingency coefficient is not an $E$-coefficient, it was not further analyzed. The matrices of the other two types were then used in a components analysis. The linear subspace generated by the two major components was then rotated by the method of simple loadings (Gorsuch, 1974). The program used in the component analysis and the rotation was BMDX72 (Sampson & Jennrich, 1970).

The preference questions were as follows:

1. Which one of the following slogans will you give most support to?
   a. We must preserve what is good from the past.
   b. Every man should have a right to have a job.
   c. Every man should have a right to create his own life.
   d. Eliminate the pollution of the environment.
2. Which one of the following parties do you prefer?
   a. The socialists
   b. The liberals
   c. The center party
   d. The conservatives
3. Which job would you prefer of the following ones?
   a. Mathematics teacher
   b. Policeman
   c. Woodcutter
   d. Zoologist
4. Which leisure activity would you choose?
   a. Weight-lifting
   b. Orienteering
   c. Chess
   d. Cross-country skiing

*Raw Data Matrix:*

Table 1.

| | Person | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | 1 | 2 | 3 | 4 | 5 | . | . | . | .10 | . | . | . | .15 | . | . | . | .20 | | | |
| 1 | a | b | b | c | c | d | a | b | d | b | c | a | c | c | a | a | b | d | d | d |
| 2 | d | a | a | b | b | c | d | a | c | a | b | d | d | b | d | d | a | c | c | b |
| 3 | d | d | b | a | c | d | a | a | a | d | d | b | b | a | a | d | d | a | d | d |
| 4 | b | b | a | c | d | b | c | c | c | b | b | a | a | b | c | b | b | b | b | b |

[1] A Fortran program CONTIN which is available from the authors is capable of computing the contingency coefficient, the J index, Tschuprow's coefficient, and Cramér's $V$ between each pair of variables read by the program.

*Contingency Coefficients:*

Table 2.

| | Item | | | |
|---|---|---|---|---|
| Item | 1 | 2 | 3 | 4 |
| 1 | .866 | .837 | .469 | .468 |
| 2 | .837 | .866 | .506 | .506 |
| 3 | .469 | .506 | .866 | .849 |
| 4 | .468 | .506 | .849 | .866 |

*Tschuprow's Coefficients:*

Table 3.

| | Item | | | |
|---|---|---|---|---|
| Item | 1 | 2 | 3 | 4 |
| 1 | 1.000 | .782 | .094 | .093 |
| 2 | .782 | 1.000 | .115 | .115 |
| 3 | .094 | .115 | 1.000 | .861 |
| 4 | .093 | .115 | .861 | 1.000 |
| Eigenvalues | 2.034 | 1.609 | .218 | .139 |

*J Indices:*

Table 4.

| | Item | | | |
|---|---|---|---|---|
| Item | 1 | 2 | 3 | 4 |
| 1 | 1.000 | .775 | .059 | .057 |
| 2 | .775 | 1.000 | .053 | .051 |
| 3 | .059 | .053 | 1.000 | .718 |
| 4 | .057 | .051 | .718 | 1.000 |
| Eigenvalues | 1.860 | 1.633 | .282 | .225 |

*Rotated Components (based on Tschuprow's coefficient):*

Table 5.

| | Item | | | |
|---|---|---|---|---|
| Component | 1 | 2 | 3 | 4 |
| 1 | -.013 | .013 | .965 | .965 |
| 2 | .946 | .942 | .0003 | -.0003 |

Note: The component intercorrelation = .114

*Rotated Components (based on J indices):*

Table 6.

| | Item | | | |
|---|---|---|---|---|
| Component | 1 | 2 | 3 | 4 |
| 1 | .004 | -.004 | .927 | .927 |
| 2 | .942 | .942 | .001 | -.001 |

Note: The component intercorrelation = .063

Both component analyses gave essentially the same result with two main "factors." The first "factor" had strong loadings in variables 3 and 4 and is thus an "interest factor." The second factor had similarly strong loadings in variables 1 and 2 and may therefore be interpreted as a "factor" of political attitude.

In both analyses the rotated component 1 can be interpreted as a component of interest and component 2 as one of ideological political attitude. The difference between the structures based on the J index and Tschuprow's coefficients was very small in this example. If, however, some alternative, e.g., alternative *c* in question 3, had not been chosen, Tschuprow's coefficient would have been undefined unless this category was deleted, which would be a doubtful convention.

## References

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 1960, *20,* 27–46.

Cohen, J. Weighted kappa: nominal scale agreement with provision for scaled disagreement of partial credit. *Psychological Bulletin,* 1968, *70,* 213–230.

Cramér, H. *Mathematical methods of statistics.* Princeton: Princeton University Press, 1946.

Fleiss, J. L., & Cohen, J. Equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, nominal scale. *Educational and Psychological Measurement,* 1973, *33,* 613–619.

Gorsuch, R. L. *Factor Analysis.* Philadelphia: W. B. Saunders Company, 1974.

Guilford, J. P. *Fundamental statistics in psychology and education.* New York: McGraw-Hill, 1956.

Holley, J. W., & Guilford, J. P. A note on the G-index of agreement. *Educational and Psychological Measurement,* 1964, *24,* 749–753.

Janson, S., & Vegelius, J. *Correlation coefficients for nominal scales* (Report 77–1). Uppsala: Department of Statistics, 1977.

Kendall, M. G. *Rank correlation methods.* London: Charles Griffin, 1948.

Pearson, K. Mathematical contributions to the theory of evolution. XIII. On the theory of contingency and its relation to association and normal correlation. *Draper's Company Research Memoirs, Biometric Series,* London: Cambridge University Press, 1904, 1.

Sampson, P. F., & Jennrich, R. I. BMDX72 factor analysis. In W. J. Dixon (Ed.), *University of California publication in automatic computation (X-series supplement).* Berkeley: University of California Press, 1970.

Shilov, G. E. *An introduction to the thory of linear spaces.* Englewood Cliffs, NJ: Prentice Hall, 1961.

Siegel, S. *Non-parametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.

Tschuprow, A. A. *Grundbegriffe und grundprobleme der korrelationstheorie.* Leipzig: Teubner, 1925.

Tschuprow, A. A. *Principles of the mathematical theory of correlation.* New York: William Hodge, 1939.

Vegelius, J. *Correlation coefficients as scalar products in Euclidean spaces* (Report 145). Uppsala: Department of Psychology, 1973.

Vegelius, J. On generalizations of the G index. *Educational and Psychological Measurement*, 1976, *36*, 595–600. (a)

Vegelius, J. On various G index generalizations and their applicability within the clinical domain. Studia Psychologica Upsaliensia 4, Uppsala: *Acta Universitatis Upsaliensis*, 1976. (b)

Vegelius, J. On the utility of the *E*-correlation coefficient concept in psychological research. *Educational and Psychological Measurement*, in press.

Yule, G. U., & Kendall, M. G. *An introduction to the theory of statistics* (14th ed.). London: Charles Griffin, 1953.

## Appendix 1
## Tables

In order to show the magnitude of the four nominal scale correlation coefficients, some contingency tables are given below with the values computed.

Table 7.

| | | | | | | |
|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 18 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |

J index = 1.000
Tschuprow's coefficient undefined
Contingency coefficient undefined
Cramér's V undefined

Table 8.

| | | |
|---|---|---|
| 2 | 0 | 0 |
| 0 | 3 | 0 |
| 0 | 0 | 5 |

J index = 1.000
Tschuprow's coefficient = 1.000
Contingency coefficient = .816
Cramér's V = 1.000

Table 9.

| | | | |
|---|---|---|---|
| 1 | 2 | 3 | 4 |
| 2 | 3 | 4 | 5 |
| 3 | 4 | 5 | 6 |

J index = .0
Tschuprow's coefficient = .003
Contingency coefficient = .085
Cramér's V = .004

Table 10.

| | | | |
|---|---|---|---|
| 2 | 0 | 0 | 0 |
| 0 | 3 | 0 | 0 |
| 0 | 0 | 5 | 0 |

J index = .986
Tschuprow's coefficient undefined
Contingency coefficient undefined
Cramér's V undefined

Table 11.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 |
| 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 |
| 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 |
| 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 |
| 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 |

J index = .0
Tschuprow's coefficent = .0
Contingency coefficent = .0
Cramér's V = .0

Table 12.

| | | |
|---|---|---|
| 1 | 2 | 3 |
| 3 | 6 | 9 |

J index = .028
Tschuprow's coefficient = .000
Contingency coefficient = .000
Cramér's V = .000

Table 13.

| | | | |
|---|---|---|---|
| 2 | 0 | 0 | 0 |
| 0 | 3 | 0 | 0 |
| 0 | 0 | 5 | 1 |

J index = .840
Tschuprow's coefficient = .816
Contingency coefficient = .816
Cramér's V = 1.000

Table 14.

| | | | |
|---|---|---|---|
| 2 | 0 | 0 | 0 |
| 0 | 3 | 0 | 0 |
| 0 | 0 | 5 | 3 |

J index = .673
Tschuprow's coefficient = .816
Contingency coefficient = .816
Cramér's V = 1.000

## Appendix 2

*Theorem:* A correlation matrix based on contingency coefficients is not necessarily non-negative definite.

*Proof:* Eight persons have obtained the following scores on four dichotomized items:

Table 15.

| | Person | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Variable | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | + | + | + | + | − | − | − | − |
| 2 | + | + | − | + | − | − | + | − |
| 3 | + | + | + | − | − | − | − | + |
| 4 | + | + | − | − | − | − | + | + |

From these scores the following correlation matrix with contingency coefficients will be obtained.

Table 16.

$$\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{5}} \\ 0 & \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

This matrix has the following eigenvalues

$$\left( \frac{1}{\sqrt{2}} + \frac{2}{\sqrt{5}}, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} - \frac{2}{\sqrt{5}} \right)$$

As the last one of them is negative, the matrix is a negative definite matrix.

Q.E.D.

## Appendix 3

*Theorem:* Cramér's $V$ is not an $E$-coefficient.

*Proof:* Variables 1 and 3 have four possible categories, while variable 2 has only three categories. Eight persons have been scored with the following results:

Table 17.

| | Person | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Variable | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| 2 | 1 | 1 | 2 | 2 | 3 | 3 | 3 | 3 |
| 3 | 1 | 1 | 2 | 2 | 3 | 4 | 3 | 4 |

If Cramér's $V$ is applied to the variable pairs, the following result is obtained:

$$V(v_1,v_2) = V(v_2,v_3) = 1$$
$$V(v_1,v_3) = \frac{2}{3}$$

Consequently, although $v_2$ should be maximally similar both to $v_1$ and $v_3$, $v_1$ and $v_3$ are not maximally similar.

Thus Cramér's $V$ is not on $E$-coefficient; It should not be applied in a component analysis.

**Acknowledgements**

*This research was supported by the Swedish Council of Social Research.*

**Author's Address**

Jan Vegelius, Dept. of Statistics, University of Uppsala, Box 513, S-751 20 Uppsala, Sweden.