# On the prehistory of Correspondence Analysis

by J. DE LEEUW*

**Abstract** The early history of correspondence analysis is reviewed. It is shown that K. PEARSON came very close to discovering correspondence analysis in 1906. That he did not actually discover it may be because he was not familiar with the singular value decomposition, which is the basic existence result in correspondence analysis.

**Key Words:** history of statistics, correspondence analysis, optimal scaling.

## 1 Introduction

Correspondence analysis is usually credited to HIRSCHFELD (1935) or to FISHER (1940), compare also MAUNG (1941a) and (1941b). HIRSCHFELD showed that it was possible to choose systems to row-scores and column-scores that exactly linearize the regressions. FISHER and MAUNG find scores that maximize the correlation coefficient of the two variables, and give other interpretations of the scores in terms of analysis of variance, discriminant analysis, and canonical correlation analysis. At about the same time GUTTMAN (1941) discovered multiple correspondence analysis, which he described in almost the same terms as FISHER and MAUNG, and only much later GUTTMAN (1959) in terms of linearizing the regressions. In this short note we indicate that K. PEARSON was very close to discovering correspondence analysis in PEARSON (1904), (1906). If he had known it, he would have used it, because it is a relatively cheap and reliable method to compute the correlation coefficient in bivariate tables for which "the order of grouping" is not necessarily known.

## 2 PEARSON's contributions

In the famous 1904 paper on contingency PEARSON proved that the mean square contingency of the bivariate normal distribution is simply related to its correlation coefficient. In fact he proved that this makes it possible to estimate the correlation coefficient by the formula

$$r^2 = X^2/(X^2 + N),$$

where $X^2$ is the chi-square for departure from independence, and where $N$ is the total number of observations. PEARSON's motivation for studying this problem is also outlined very clearly. In the course of his empirical work on the inheritance of eye colour, coat colour of horses and greyhounds, and human hair colour, he continually ran into the problem of defining a scale in order to compute correlation. In his experiments with different scales he found, to his pleasant surprise, that in many cases the correlation coefficient did not change greatly if different scales or groupings were used. Since the coefficient of contingency is independent from any considerations of scale or order, it

* Department of Data Theory FSW/RUL, Middelstegracht 4, 2312 TW Leiden, The Netherlands.

seemed to PEARSON that his result explained the relative invariance of correlation in the case of a bivariate normal with sufficiently fine grouping.

On page 19 and 20 of the same memoir PEARSON derives another result, which we first give in words. "Hence we conclude that in any correlated system of variables, obeying the law of linear regression, we can, without sensibly modifying the correlation, interchange two adjacent $y$-arrays (e.g. two rows of the correlation table), provided the grouping be fine. But if we can change any two adjacent $y$-arrays, we can, by a repetition of such changes, interchange any two $y$-arrays whatever; and a precisely similar statement must be valid for any two $x$-arrays (e.g. two columns of the correlation table). Hence, given a sufficiently small system of grouping, we may state that in all cases of linear regression the actual order of the scales is immaterial as far as the determination of the correlation is concerned." (PEARSON (1904), p. 20).

This is a very sweeping statement, and the proof offered is quite disappointing. PEARSON shows that if the grouping is fine then interchanging two rows has a vanishing first order contribution to the correlation if the regression is linear. PEARSON was apparently criticized by YULE, who did not think it obvious that this comparatively simple mathematical result had such far-reaching practical implications. In PEARSON (1906) there is a defensive note. "I think this conclusion is quite sound, and deserves further consideration. Although in the statement of the proposition I have used the word 'small changes' in scale order (p. 19) and in the summary of my memoir (p. 35) stated what is to be understood by small, in this case, I think, as Mr. G. U. YULE points out to me, that the wording on p. 20 is too unguarded, if the reader has not been sufficiently impressed with the wording on p. 19, or reached the summary on p. 35." (PEARSON (1906), p. 176). In the 1906 note this is followed by a more complete proof, with even more far reaching practical implications.

We shall discuss the practical implications, and the connection with correspondence analysis, in a later section. First we give a generalized and modernized treatment of PEARSON's result. The main difference is that PEARSON treated a very restricted class of perturbations, the interchange of two adjacent arrays in a finely grouped but discrete bivariate probability distribution. We generalize to arbitrary random variables (with finite variances) and to arbitrary "small changes".

## 3 PEARSON's theorem

Suppose $x$ and $y$ are two random variables with a joint distribution and with finite variances. In fact we can suppose without loss of generality than they have expectation zero and variance one. Let $\varrho_{xy} = \mathsf{E}(xy)$. We now consider new random variables $\underline{x} = f(x)$ and $\underline{y} = g(y)$, with $f$ and $g$ measurable, with $\underline{x}$ and $\underline{y}$ having finite variance, and with $\varepsilon = \underline{x} - x$ and $\delta = \underline{y} - y$ in the role of "small changes". We compute perturbed moments. For the variance we find, using $\sigma^2$ for variances and $\varrho$ throughout for correlations,

$$\sigma_{\underline{x}}^2 = 1 + 2\sigma_\varepsilon \varrho_{x\varepsilon} + \sigma_\varepsilon^2. \tag{1a}$$

Similarly

$$\sigma_{\underline{y}}^2 = 1 + 2\sigma_\delta \varrho_{y\delta} + \sigma_\delta^2, \tag{1b}$$

while the covariance is given by

$$\sigma_{\underline{xy}} = \varrho_{xy} + \sigma_\delta \varrho_{x\delta} + \sigma_\varepsilon \varrho_{y\varepsilon} + \sigma_\varepsilon \sigma_\delta \varrho_{\varepsilon\delta}. \tag{2}$$

The perturbed correlation is

$$\varrho_{\underline{xy}} = \sigma_{\underline{xy}} \sigma_{\underline{x}}^{-1} \sigma_{\underline{y}}^{-1} = (\varrho_{xy} + \sigma_\delta \varrho_{x\delta} + \sigma_\varepsilon \varrho_{y\varepsilon} + \sigma_\varepsilon \sigma_\delta \varrho_{\varepsilon\delta})(1 + 2\sigma_\varepsilon \varrho_{x\varepsilon} + \sigma_\varepsilon^2)^{-1/2} \\ (1 + 2\sigma_\delta \varrho_{y\delta} + \sigma_\delta^2)^{-1/2} \tag{3}$$

If we expand this we find

$$\begin{aligned} \varrho_{\underline{xy}} = \varrho_{xy} &+ \sigma_\delta (\varrho_{x\delta} - \varrho_{xy}\varrho_{y\delta}) + \sigma_\varepsilon (\varrho_{y\varepsilon} - \varrho_{xy}\varrho_{x\varepsilon}) + \\ &- \tfrac{1}{2}\sigma_\varepsilon^2 \{\varrho_{xy}(1 - 3\varrho_{x\varepsilon}^2) + 2\varrho_{x\varepsilon}\varrho_{y\varepsilon}\} + \\ &- \tfrac{1}{2}\sigma_\delta^2 \{\varrho_{xy}(1 - 3\varrho_{y\delta}^2) + 2\varrho_{x\delta}\varrho_{y\delta}\} + \\ &+ \sigma_\varepsilon \sigma_\delta (\varrho_{xy}\varrho_{x\varepsilon}\varrho_{y\delta} - \varrho_{x\delta}\varrho_{x\varepsilon} - \varrho_{y\delta}\varrho_{y\varepsilon}) + \mathrm{o}\{\max(\sigma_\varepsilon^2, \sigma_\delta^2)\}. \end{aligned} \tag{4}$$

If both regressions are linear then

$$\varrho_{x\delta} = \varrho_{xy}\varrho_{y\delta}, \tag{5a}$$

$$\varrho_{y\varepsilon} = \varrho_{xy}\varrho_{x\varepsilon}. \tag{5b}$$

If we substitute this in the expansion, we find

$$\varrho_{\underline{xy}} = \varrho_{xy}\{1 - \tfrac{1}{2}\sigma_\varepsilon^2(1 - \varrho_{x\varepsilon}^2) - \tfrac{1}{2}\sigma_\delta^2(1 - \varrho_{y\delta}^2) - \sigma_\varepsilon\sigma_\delta\varrho_{x\varepsilon}\varrho_{y\delta}\} + \mathrm{o}\{\max(\sigma_\varepsilon^2, \sigma_\delta^2)\}. \tag{6}$$

Observe that this result remains true under the weaker condition that $x$ and $\delta$ are uncorrelated given $y$, and $y$ and $\varepsilon$ are uncorrelated given $x$. This is basically the result used by PEARSON, in a much generalized form.

## 4 PEARSON's conclusions

The first conclusion, corresponding to PEARSON ((1904), p. 19–20), follows from (4). If both regressions are linear, then

$$\varrho_{\underline{xy}} = \varrho_{xy} + \mathrm{O}\{\max(\sigma_\varepsilon^2, \sigma_\delta^2)\}. \tag{7}$$

Thus "small changes (i.e., such that the sum of their squares may be neglected as compared with the square of mean or standard deviation) may be made in the order of grouping without affecting the correlation coefficient." (PEARSON (1904), p. 35). To see the relation with correspondence analysis observe that this can also be formulated more suggestively as: if both regressions are linear then the correlation coefficient is stationary with respect to variations in $f$ and $g$. This is, at least partly, the key result of HIRSCHFELD (1935).

But PEARSON (1906) goes further than this. He concludes from (6) that if the regressions are linear and if $\varrho_{xy}$ is nonnegative, then actually $r_{xy}$ is an absolute maximum. "Or, we conclude that if there be one arrangement of the material for which the regression line is linear, then any interchanges, however extensive, will reduce the value of the correlation as calculated by the product moment method. This conception of the linear regression line as giving the arrangement with the maximum degree of correlation appears of considerable philosophical interest. It amounts practically to much the same thing as saying that if we have a fine classification, we shall get the maximum of correlation by arranging the arrays so that the means of the arrays fall as closely as possible on a line." (PEARSON (1906), p. 178). But this is exactly what correspondence analysis does. PEARSON just lacks an existence theorem which tells him that such an arrangement is always perfectly possible, and in many different ways. Or, to put it differently, he was not familiar with the singular value decomposition, although this had been discovered much earlier by BELTRAMI, SYLVESTER, and JORDAN.

We must emphasize that, of course, it does not follow from (6) that $\varrho_{xy}$ is an absolute maximum for variation of $f$ and $g$. Such a result cannot be proved with local methods. In general it is false, compare SARMANOV and BRATOEVA (1967), example 2. In PEARSON's context, in which we interchange adjacent arrays, it is easy to see that $\varrho_{x\varepsilon}$ and $\varrho_{y\delta}$ are both negative, which means that $\varrho_{xy}$ is at least a local maximum if the regressions are linear.

"While not desiring to encourage carelessness in observing or tabling or in the formation of scale orders without due consideration, still the results of this note seem to indicate that in many cases absolute unanimity of judgment in classifying or great stress on small details of scale grouping are not needful in order to reach sensibly identical values of the correlation." (PEARSON (1906), p. 178). We agree with the spirit of this remark, amount other things because of our experience with correspondence analysis as a method for optimal scaling. We also think that PEARSON's theorem indicates that it is very important to investigate linearity of the regressions for the scoring system one intends to use. One way to do this is by using correspondence analysis.

## References


FISHER, R. A. (1940), The precision of discriminant functions, *Annals of Eugenics 10,* 422-429.

GUTTMAN, L. (1941), The quantification of a class of attributes: a theory and method for scale construction, in: P. HORST (ed.) *The prediction of personal adjustment,* New York, Social Science Research Council.

GUTTMAN, L. (1959), Metricizing rank-ordered or unordered data for a linear factor analysis, *Sankhya 21,* 257-268.

HIRSCHFELD, H. O. (1935), A connection between correlation and contingency, *Proceedings Cambridge Philosophical Society 31,* 520-524.

MAUNG, K. (1941a), Discriminant analysis of TOCHER's eye colour data for Scottish school children, *Annals of Eugenics 11,* 64-76.

MAUNG, K. (1941b), Measurement of association in a contingency table with special reference to the pigmentation of hair and eye colours of Scottish school children, *Annals of Eugenics 11,* 189-223.

PEARSON, K. (1904), Mathematical contributions to the theory of evolution XIII. On the theory of contingency and its relation to association and normal correlation, *Drapers Company Research Memoirs, Biometric Series 1.*

PEARSON, K. (1906), On certain points connected with scale order in the case of a correlation of two characters which for some arrangement give a linear regression line, *Biometrika 5,* 176-178.

SARMANOV, O. V. and Z. N. BRATOEVA (1967), Probabilistic properties of bilinear expansions of Hermite polynomials, *Theory of Probability and its applications 12,* 470-481.



*Received March 1983.*