VISUAL DISPLAY OF INTERACTION IN MULTIWAY CONTINGENCY TABLES BY USE OF HOMOGENEITY ANALYSIS: THE 2 x 2 x 2 x 2 CASE

Jacqueline J. Meulman
Willem J. Heiser

Department of Data Theory
Leiden University

RR-96-05

Visual Display of Interaction in Multiway Contingency Tables by Use of Homogeneity Analysis: the $2 \times 2 \times 2 \times 2$ Case

Abstract

Multiway contingency tables contain the product-multinomial distribution formed by all possible cross-classifications from a set of categorical variables. Two techniques for the analysis of such tables are loglinear analysis and homogeneity analysis (also called multiple correspondence analysis or dual scaling), where the latter should be applied to the associated profile frequency table. Homogeneity analysis gives a spatial representation of the latter table by mapping the rows and columns into points in a low-dimensional (often, but not necessarily, a two-dimensional) space. It is usually thought that homogeneity analysis relies on the assumption that the total structure in the data can be sufficiently captured in the joint bivariate (first order) interactions. This assumption can be considered equivalent to assuming a loglinear model that includes the pairwise associations only. If, however, these are not sufficient to produce a decent fit, the conclusion would be that homogeneity analysis should be discarded in favor of a loglinear analysis that includes the higher-order interactions. It will be shown in the present paper that in contrast to this belief, homogeneity analysis represents the higher-order interaction too. The presence of these higher-order interactions can be checked by examining the odds that are obtained as distance ratios derived from particular category quantifications.

1. Introduction

Multiway contingency tables express the relationships between the categories of several categorical variables *A, B, C,* ... at several levels of complexity. In the body of the multiway table, all these relationships are confounded. By adding over all variables except *A* and *B*, we obtain a *bivariate marginal* table, showing the bivariate relationship between *A* and *B*; by adding over all variables except *A, B* and *C*, we obtain a *trivariate marginal* table, showing the trivariate relationship between *A, B* and *C*, and so on. Of course, it is of interest to know whether these relationships, once separated, perhaps are still more simple than they appear; in particular, to ask whether or not the higher-order ones are simple combinations of the lower-order ones. From this

question, several natural forms of dependence and independence arise.

Suppose for the moment that we restrict ourselves to the case of three categorical variables, with n_A, n_B and n_C categories. Lack of independence implies the presence of *interaction* between the categories. Let π_{ijk} denote the probability that an individual unit of observation (in the sequel denoted by the neutral term object) falls in category i of variable A, in category j of variable B, and in category k of variable C. We consider π_{ijk} as a probability defined over all cells of the three-way contingency table, which implies that the total sum $\Sigma_i\Sigma_j\Sigma_k\, \pi_{ijk} = 1$. As will be demonstrated in section 2, we can discuss several concepts of independence structure without referring, for example, to *loglinear modelling*, which is perhaps the most common approach to analyze data of the present type, or to any other form of modelling. The reason is simple: models involve assumptions to relate concepts of structure to observed counts, but the concepts exist regardless of the additional assumptions. Modelling is a way to *smooth* empirical frequencies, that is, to replace them by frequencies satisfying certain regularities. Loglinear modelling just uses various types of independence as a set of possible structures for the expected values μ_{ijk} of a multinomial sampling process with $n_A \times n_B \times n_C$ categories. Then, by a famous result of Birch (1963), the maximum likelihood fitted values $\hat{\mu}_{ijk}$ are smoothed versions of the counts in the observed multiway contingency table that match them in specified marginal distributions, but have higher-order interactions that satisfy the chosen independence patterns. For choosing between submodels with a different independence structure, the likelihood ratio statistic G^2 or Pearson's chi-squared statistic χ^2 are used.

Once we know the most likely (in)dependence structure among the variables, how do we interpret the interactions? In loglinear modelling, interactions correspond to groups of model parameters. To interpret the model parameters of a loglinear model, we have to express them in terms of odds and odds ratios (also called *cross-product ratios*, see Fienberg, 1980), which are ratios of (smoothed) frequencies or probabilities. This reformulation is not easy; it involves taking the exponential of a model parameter and describing the corresponding odds verbally. A verbal description of a three-way interaction can become incomprehensible rather quickly, because it consists of a nesting of conditional statements. The main thesis of this paper is that the

(in)dependence structure can *also* be represented in a spatial model, in which categories are mapped as points, and variables as groups of points. It will be shown that in this spatial representation odds are ratios of distances, a property that offers the possibility of visual display of interaction.

The spatial representation will be obtained through the use of homogeneity analysis (Gifi, 1990, Chapter 3 and section 8.6; also called multiple correspondence analysis, Benzécri, 1973; Greenacre, 1984, or dual scaling, Nishisato, 1994). In the present context, the technique will be regarded as a method that maps the rows of a *profile frequency table* into points in a low-dimensional space (often, but not necessarily, a two-dimensional space). The profile frequency table is the multiway contingency table turned inside-out: it codes the cells by listing, in some predetermined order, which category of each variable is involved (forming the profile), and attaches to each profile the cell frequency. Points representing the profiles are called *profile points*, and *category points* are obtained as *centers of gravity* of certain subsets of profile points. This construction will described in greater detail in section 3; an example of a profile frequency table will be given in section 4.

Homogeneity analysis was developed with a focus on bivariate marginal tables. If χ^2_{AB} is the usual chi-squared statistic for lack of independence in the cross tabulation of variables A and B, and λ_s^2 the usual summary statistic (called *eigenvalue*) for dimension s from a homogeneity analysis, then from an early result by Guttman (1941) it follows that

$$\Sigma_A \Sigma_B \chi^2_{AB} = N \Sigma_s (m \lambda_s^2 - 1)^2 ,$$

where N is the number of objects and m the number of variables (Gifi, 1990, section 3.10). Thus, if all variables are mutually independent, all eigenvalues will be equal to $1/m$, and an analysis result in which the first p eigenvalues deviate substantially from $1/m$ implies that the first p dimensions account for all two-way interactions. It is asserted in Gifi (1990) that homogeneity analysis relies on the assumption that in most cases the total structure in the data can be sufficiently captured in the joint bivariate (first order) interactions. This assumption can be considered equivalent to

assuming a loglinear model that includes the pairwise associations only. If, however, these are not sufficient to produce a decent fit, the conclusion would be that homogeneity analysis should be discarded in favor of a loglinear analysis that includes the higher-order interactions. The major purpose of the present paper is to show that - in contrast to this widespread idea - homogeneity analysis includes the representation of the higher-order interactions as well.

An important idea in the sequel of this paper is the balanced use of cross-classification variables; for two variables A and B, the cross-classification variable AB is composed of the $n_A \times n_B$ combinations of the original categories of A and B. The paper is then organized as follows. First, we will review briefly the important concepts of independence structure and odds structure. Next, we will show that odds are ratios of distances between category points in a homogeneity analysis, and how independence between variables leads to additivity of their category quantifications. Having performed such an extended homogeneity analysis, additional discrimination measures will be available for cross-classification variables expressing first and higher-order interactions. It will be shown that discrimination measures of two variables are additive when they are independent, and that the presence of higher-order interaction can be checked through distance ratios based on category coordinates (quantifications) of a higher-order cross-classification variable and category coordinates associated with lower-order interactions. From a practical point of view, a remarkable phenomenon should be mentioned, which is that the results of the extended homogeneity analysis, i.e. the object scores, the discrimination measures and the category quantifications, are very similar to those in a simple homogeneity analysis, i.e., when only the original variables would be included.

To conclude the introduction we briefly describe the data that will be used throughout for empirical illustration. The data pertain to a sample of men and women who had petitioned for divorce; a similar number of married people were asked the following questions:

(*a*) "Before you married your (former) husband/wife, had you ever made love with anyone else?";

(*b*) "During your (former) marriage, (did you have) have you had any affairs or brief sexual

encounters with another man/woman?".

The variables in the $2 \times 2 \times 2 \times 2$ cross tabulation (with total sample size $N = 1036$) are *Gender (G), Premarital Sex (P), Extramarital Sex (E), and Marital Status (M).* The associated profile frequency matrix will be given in section 4. The multiway contingency table is analysed in Agresti (1990, section 7.2.4); the original British study was reported by Thornes and Collard (1979), and described by Gilbert (1981).

2. Independence Structures and Odds Structures

The following cases of simplification of a three-way contingency table are commonly distinguished (e.g., see Agresti, 1990). Three variables *A, B,* and *C* are called *mutually independent* if

$$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k} , \tag{1}$$

for all categories i of A, j of B, and k of C, where, as usual, π_{i++} indicates that we have summed over j and k, giving the *univariate marginals* for variable A. Under mutual independence, there is no association whatsoever, and all cells of the three-way table can be constructed by the simple product of the univariate marginals. Variable A is called *jointly independent* of B and C when, again for all categories,

$$\pi_{ijk} = \pi_{i++}\pi_{+jk} . \tag{2}$$

This decomposition corresponds to ordinary two-way independence for A and a new variable, called the *cross-classification* variable BC, which is composed of the $n_B \times n_C$ combinations of the categories of B and C. Under joint independence, the association between two variables is distributed proportionally over the levels of the third variable to obtain the three-way probability.

Next, consider the relationship between A and B, controlling for the contribution of C. Here the concept of control implies that we study the *conditional probability* that two categories, say i of A, and j of B, are present in the same object, given the fact that we know the object is in category k of

C. The usual notation for this event is $\pi_{ij|k}$, defined as $\pi_{ij|k} = \pi_{ijk} / \pi_{++k}$, where the division by the univariate marginal π_{++k} ensures that the $\pi_{ij|k}$'s form a proper set of probabilities summing to one within the subtable indexed by k. From this definition it follows that the cell probabilities can be expressed in terms of conditional probabilities as:

$$\pi_{ijk} = \pi_{++k}\pi_{ij|k} \ .$$

Now if, for all k, the conditional probabilities $\pi_{ij|k}$ are independent, we must have the marginal decomposition

$$\pi_{ij|k} = (\pi_{i+k} / \pi_{++k}) \ (\pi_{+jk} / \pi_{++k}),$$

and therefore we obtain, combining the last two equations, *conditional independence of A and B given C* when

$$\pi_{ijk} = (\pi_{i+k}\pi_{+jk}) / \pi_{++k} \ . \qquad (3)$$

Under conditional independence of A and B, each of this pair of variables is associated with C; these associations, together with the univariate marginal, completely account for the apparent association between A and B in the original table and in the bivariate marginal π_{ij+}.

Note that cases (3), (2), and (1) are fundamentally different only in terms of the number of two-way cross-classification variables that are needed to account for the cell probabilities. For conditional independence we need two cross-classification variables, for joint independence we need one such variable, and for mutual independence we need none. Conversely, it is also useful to think of the situation in terms of *conditionally dependent* variables. The strongest case is (1), in which there are no conditionally dependent variables. In case (2), B and C are conditionally dependent given A, implying that the conditional probability $\pi_{jk|i} = \pi_{+jk}$ does not simplify, while $\pi_{ij|k} = \pi_{i++}\pi_{j|k}$ and $\pi_{ik|j} = \pi_{i++}\pi_{k|j}$ are independent. In case (3), only the conditional probability $\pi_{ij|k} = \pi_{i|k}\pi_{j|k}$ can be decomposed into the product of its marginals, while $\pi_{jk|i}$ and $\pi_{ik|j}$ depend on π_{+jk} and π_{i+k}, respectively; so both B and C and A and C are conditionally dependent.

It is natural also to consider the case where π_{ijk} depends on *three* double subscripted quantities,

$$\pi_{ijk} = \alpha_{ij}\beta_{ik}\gamma_{jk}, \tag{4}$$

a case for which no closed-form expression in terms of marginal probabilities exists. Here, none of the pairs of variables is conditionally independent, yet there still is a typical form of simplification: odds ratios between two variables are identical for each (given) category of the third variable. The *odds ratio* is a classic way of measuring association (Yule, 1912) that compares two ratios of probabilities (odds) by forming a ratio again. Thus, the odds of being in category 1 of A rather than in category 2 of A, are compared for those who are in category 1 of B against those who are in category 2 of B. In a 2×2 table, the odds ratio θ is defined as $\theta = (\pi_{11}\pi_{22}) / (\pi_{12}\pi_{21})$. For three variables, with π_{ijk} satisfying the stated condition, we find $\theta_k = (\pi_{11k}\pi_{22k}) / (\pi_{12k}\pi_{21k}) = (\alpha_{11}\alpha_{22}) / (\alpha_{12}\alpha_{21})$, showing that θ_k is independent of the chosen category k (by symmetry, the effect is the same if the categories of the other variables are kept fixed). After Bartlett (1935), this case is usually called "no three-way interaction".

In summary, all cases of independence have a typical *odds structure*, which is shown in Table 1, displaying the result of inserting (1) – (4) into the definition of the odds ratio.

Table 1

Odds structures in a three-way table under different forms of independence

		$\theta_i = \frac{\pi_{i11}\pi_{i22}}{\pi_{i12}\pi_{i21}}$	$\theta_j = \frac{\pi_{1j1}\pi_{2j2}}{\pi_{1j2}\pi_{2k1}}$	$\theta_k = \frac{\pi_{11k}\pi_{22k}}{\pi_{12k}\pi_{21k}}$
I	mutual independence	1	1	1
II	joint independence	$\frac{\pi_{+11}\pi_{+22}}{\pi_{+12}\pi_{+21}}$	1	1
III	conditional independence	$\frac{\pi_{+11}\pi_{+22}}{\pi_{+12}\pi_{+21}}$	$\frac{\pi_{1+1}\pi_{2+2}}{\pi_{1+2}\pi_{2+1}}$	1
IV	no three-way interaction	$\frac{\gamma_{11}\gamma_{22}}{\gamma_{12}\gamma_{21}}$	$\frac{\beta_{11}\beta_{22}}{\beta_{12}\beta_{21}}$	$\frac{\alpha_{11}\alpha_{22}}{\alpha_{12}\alpha_{21}}$

Under mutual independence, all odds ratios are equal to one. Under joint independence (of A with respect to B and C), there is one set of odds ratios that does not become equal to one: all θ_i become equal to the marginal odds ratio, that is, the two-way tables conditioned on category i are equal to the marginal table. Under conditional independence (of A and B upon C), both the θ_i's and the θ_j's become equal to the marginal odds ratio, while $\theta_k = 1$ for all k. Under lack of three-way interaction, all odds ratios for different categories of the same variable are equal, but unequal to one. Finally, three-way interaction implies that all odds ratios are different, both within and across variables. As we shall see shortly, these various odds structures each have a distinctive spatial pattern.

3. Odds as Distance Ratios

In this section it will be shown that odds are distance ratios between category points, and how this leads to additivity of category quantifications. Homogeneity analysis finds the spatial representation of the profile frequency table by *projection.* Projection is a linear transformation which intuitively involves dropping points onto a line (or plane), along a direction perpendicular to the line (or plane). We will first describe what is projected (a high-dimensional representation of the table), then show some of the properties that hold in this high-dimensional space, and next indicate which properties remain (approximately) preserved under projection. For more technical details on projection, the reader is referred to Gifi (1990) or Van de Geer (1993).

In the high-dimensional representation of the profile frequency table, all objects with the same profile coincide in one point, called the *profile point* $\mathbf{z}_{ijk\ldots}$. We associate with each profile point a *mass* (also called *weight*), equal to the cell frequency $\pi_{ijk\ldots}$ of that profile. Note that our starting point is the cells of the multiway contingency table itself, not any of its marginal tables. We also assume that the number of variables is much smaller than the number of objects, $N >> m$, a condition similar to what is required for a loglinear analysis, and that all frequencies are strictly greater than zero (although this is not necessary for the spatial method). If $n_A, n_B, \ldots, n_m$ are the

number of categories of the m variables, this construction generates $n_A \times n_B \times \ldots \times n_m$ profile points. In the following discussion, we limit ourselves to the $2 \times 2 \times 2$ case.

In the binary case, the 2^m profile points are the vertices of an m-dimensional (hyper)cube associated with some probability mass. Thus, three variables are represented as 8 profile points on a cube in three dimensions. Focusing on the edges between the two faces of the cube that correspond to the categories 1 and 2 of variable A, we may locate on each edge between the vertices $\mathbf{z}_{1jk}$ and $\mathbf{z}_{2jk}$ the point $\mathbf{z}_{*jk}$, defined as

$$\mathbf{z}_{*jk} = \frac{\pi_{1jk}}{\pi_{1jk} + \pi_{2jk}} \mathbf{z}_{1jk} + \frac{\pi_{2jk}}{\pi_{1jk} + \pi_{2jk}} \mathbf{z}_{2jk} , \tag{5}$$

which is the *center of gravity* (or *centroid*) of all objects in category j of B and k of C, of which there are π_{1jk} in 1 of A and π_{2jk} in 2 of A. Because we know that the points $\mathbf{z}_{1jk}$, $\mathbf{z}_{*jk}$ and $\mathbf{z}_{2jk}$ are located on a line, in that order (because (5) is a convex combination), we may write $d(\mathbf{z}_{1jk}, \mathbf{z}_{2jk}) = d(\mathbf{z}_{1jk}, \mathbf{z}_{*jk}) + d(\mathbf{z}_{2jk}, \mathbf{z}_{*jk})$, where the notation $d(\mathbf{x}, \mathbf{y})$ is used for the ordinary Euclidean distance between two points $\mathbf{x}$ and $\mathbf{y}$. So the edge between two profile points is divided by the centre of gravity into two parts. Using (5) and the *order* along the line, the lengths of these two segments are

$$d(\mathbf{z}_{1jk}, \mathbf{z}_{*jk}) = \frac{\pi_{1jk}}{\pi_{1jk} + \pi_{2jk}} \mathbf{z}_{1jk} + \frac{\pi_{2jk}}{\pi_{1jk} + \pi_{2jk}} \mathbf{z}_{2jk} - \mathbf{z}_{1jk} = \frac{\pi_{2jk}}{\pi_{1jk} + \pi_{2jk}} d(\mathbf{z}_{1jk}, \mathbf{z}_{2jk}) , \tag{6}$$

$$d(\mathbf{z}_{2jk}, \mathbf{z}_{*jk}) = \mathbf{z}_{2jk} - \frac{\pi_{1jk}}{\pi_{1jk} + \pi_{2jk}} \mathbf{z}_{1jk} - \frac{\pi_{2jk}}{\pi_{1jk} + \pi_{2jk}} \mathbf{z}_{2jk} = \frac{\pi_{1jk}}{\pi_{1jk} + \pi_{2jk}} d(\mathbf{z}_{1jk}, \mathbf{z}_{2jk}) . \tag{7}$$

From (6) and (7) it follows that the odds of being in category 2 of A against being in category 1, given the fact that the object is in j of B and k of C is equal to:

$$\frac{\pi_{2jk}}{\pi_{1jk}} = \frac{d(\mathbf{z}_{1jk}, \mathbf{z}_{*jk})}{d(\mathbf{z}_{2jk}, \mathbf{z}_{*jk})} , \tag{8}$$

that is, the odds are displayed in the spatial representation of the profile frequency table as a *reverse distance ratio* (larger probabilities corresponding to smaller distances between $\mathbf{z}_{*jk}$ and the

profile point). From (8) we can now derive novel expressions for the odds ratios in subtables of a three-way table; for example, for the association between A and C given category j of B we obtain, by putting $k = 1$ and $k = 2$ and dividing the odds,

$$\theta_j = \frac{\pi_{1j1}\pi_{2j2}}{\pi_{1j2}\pi_{2j1}} = \frac{d(\mathbf{z}_{1j2}, \mathbf{z}_{*j2})\, d(\mathbf{z}_{2j1}, \mathbf{z}_{*j1})}{d(\mathbf{z}_{1j1}, \mathbf{z}_{*j1})\, d(\mathbf{z}_{2j2}, \mathbf{z}_{*j2})} . \tag{9}$$

Thus, the odds ratio θ_j is a multiplicative combination of four distances, defined between four profile points and two centroids. It is well-known that the odds ratio is invariant under permutation of the rows and columns of the four-fold table. This property implies here that θ_j may *also* be derived from the ratio of π_{ij2} and π_{ij1}, which leads to an alternative expression for (9) in terms of the centroids $\mathbf{z}_{ij*}$, defined analogously to $\mathbf{z}_{*jk}$ in (5). We shall have a closer look at this duplication when we illustrate the spatial relationships with an example in section 5.

What is the spatial representation of independence? The reader is advised to draw a square with vertices $\mathbf{z}_{1j1}$, $\mathbf{z}_{1j2}$, $\mathbf{z}_{2j2}$ and $\mathbf{z}_{2j1}$; when $\mathbf{z}_{*jk}$ and $\mathbf{z}_{ij*}$ are added to this figure, the following relationships are verified easily. If variable A is jointly independent of B and C, we know (see Table 2) that $\theta_j = 1$, so from (9) we derive $d(\mathbf{z}_{1j2}, \mathbf{z}_{*j2})\, d(\mathbf{z}_{2j1}, \mathbf{z}_{*j1}) = d(\mathbf{z}_{1j1}, \mathbf{z}_{*j1})\, d(\mathbf{z}_{2j2}, \mathbf{z}_{*j2})$. But we also know, by the construction of the spatial representation, that the interprofile distances are equal, that is, $d(\mathbf{z}_{1j1}, \mathbf{z}_{2j1}) = d(\mathbf{z}_{1j2}, \mathbf{z}_{2j2})$, from which we derive $d(\mathbf{z}_{1j1}, \mathbf{z}_{*j1}) + d(\mathbf{z}_{2j1}, \mathbf{z}_{*j1}) = d(\mathbf{z}_{1j2}, \mathbf{z}_{*j2}) + d(\mathbf{z}_{2j2}, \mathbf{z}_{*j2})$. Taken together, and after some algebraic manipulation, these two equalities imply that *the four distances are equal in opposite pairs*: $d(\mathbf{z}_{1j1}, \mathbf{z}_{*j1}) = d(\mathbf{z}_{1j2}, \mathbf{z}_{*j2})$ and $d(\mathbf{z}_{2j1}, \mathbf{z}_{*j1}) = d(\mathbf{z}_{2j2}, \mathbf{z}_{*j2})$. If we consider the *intersection line* connecting $\mathbf{z}_{*j1}$ with $\mathbf{z}_{*j2}$, it must be parallel to the edges $(\mathbf{z}_{1j1}, \mathbf{z}_{1j2})$ and $(\mathbf{z}_{2j1}, \mathbf{z}_{2j2})$. A similar relation holds for the intersection line between $\mathbf{z}_{1j*}$ and $\mathbf{z}_{2j*}$ with its corresponding edges. So we conclude that independence is a necessary condition for the intersection lines to be parallel to the edges.

It is natural to assign to each centroid $\mathbf{z}_{*jk}$ a mass, π_{+jk}, indicating the proportion of objects that has a profile with j of B and k of C. Similarly, the marginal proportion π_{ij+} will be assigned to $\mathbf{z}_{ij*}$, that is, the sum of the masses of which it is the balancing point. The two intersection lines

($\mathbf{z}_{*j1}$, $\mathbf{z}_{*j2}$) and ($\mathbf{z}_{1j*}$, $\mathbf{z}_{2j*}$) themselves intersect in a point $\mathbf{z}_{*j*}$, called the *category point* (the coordinates of which are called *category quantifications*) of category j, which is easily shown *also* to be a center of gravity (of all objects in category j, calculated in any of a number of different ways), with mass π_{+j+}. Continuing in this way, the intersection lines connecting the category points, ($\mathbf{z}_{1**}$, $\mathbf{z}_{2**}$), ($\mathbf{z}_{*1*}$, $\mathbf{z}_{*2*}$) and ($\mathbf{z}_{**1}$, $\mathbf{z}_{**2}$), intersect in $\mathbf{z}_{***}$, the centroid of all objects, with mass 1.

Our high-dimensional spatial representation of the profile frequency table is now complete. For the $2 \times 2 \times 2$ case, it contains the 8 original profile points, 3×4 added one-asterisk centroids, 3 added two-asterisks centroids, and the overall, three-asterisks centroid. The masses of these points correspond exactly to all the cell probabilities and the complete set of marginal probabilities of the three-way contingency table. Just as the centroids of the form $\mathbf{z}_{*j*}$ are called category points of B, centroids of the form $\mathbf{z}_{*jk}$ are called the category points (quantifications) of the cross-classification variable BC (similarly we have category points for AB and AC). So all cells of the bivariate marginal tables can be viewed as categories of some cross-classification variable, which is quantified by centroids located on the edges of the cube of profile points.

As we have seen, lack of interaction implies parallel intersection lines, and this has important further implications for the relationship between the category quantifications of the bivariate marginals with the profile points on the one hand, and with the univariate marginals on the other hand. We suppose that the origin of the space is chosen as $\mathbf{z}_{***}$. Considering vectors in the face of the cube corresponding to category j of B, which are obtained from the original ones by translation with an amount $-\mathbf{z}_{*j*}$, parallelism implies additivity:

$$(\mathbf{z}_{ijk} - \mathbf{z}_{*j*}) = (\mathbf{z}_{ij*} - \mathbf{z}_{*j*}) + (\mathbf{z}_{*jk} - \mathbf{z}_{*j*}) \,, \tag{10}$$

which follows from the definition of vector addition in terms of the parallelogram formed by the points $\mathbf{z}_{*j*}$, $\mathbf{z}_{*jk}$, $\mathbf{z}_{ijk}$, and $\mathbf{z}_{ij*}$ (this is in fact a rectangle, but we only want to use the parallelism, not any properties of the angles). Thus, conditional independence must manifest itself by the fact that one of the three possible pairs of cross-classification variables has additive quantifications

when viewed with respect to the univariate centroid, as in (10). Under joint independence, we must have two pairs of cross-classification variables with additive quantifications with respect to their joint univariate centroid. Similarly, it can be shown that, when variables B and C are independent, we have a marginal odds ratio $(\pi_{+11}\pi_{+22}) / (\pi_{+12}\pi_{+21}) = 1$, which implies

$$\mathbf{z}_{*jk} = \mathbf{z}_{*j*} + \mathbf{z}_{**k} , \tag{11}$$

that is, the quantifications of the cross-classification variable are equal to the sum of the quantifications of the categories of the original variables. Combining (10) and (11), we obtain the spatial representation of mutual independence:

$$\mathbf{z}_{ijk} = \mathbf{z}_{i**} + \mathbf{z}_{*j*} + \mathbf{z}_{**k} . \tag{12}$$

In this case, the category points of all three cross-classification variables form a parallelogram. So there is a clear one-to-one correspondence between odds structures in the three-way table and additivity structures in the spatial model.

All relationships described so far are exact in the original cube, and we may wonder how well they remain intact in the projected configuration that constitutes the usual result of a homogeneity analysis. Angles and distances are not preserved under projection: squares and rectangles become parallelograms. Projection does preserve parallelism, so (10), (11), and (12) remain *completely valid* in a low-dimensional representation of the profile frequency table.

4. Some Special Properties of Discrimination Measures

We will now propose and illustrate a general procedure to study interaction in higher-way contingency tables that allows us to distinguish the various additivity structures in a low-dimensional representation of the profile frequency table. Our procedure simply amounts to a homogeneity analysis of a profile frequency matrix including all cross-classification variables that can be formed from the original variables in a completely balanced way. If there is reason to expect

a three-way interaction (for example, as indicated by a model search in a preliminary loglinear analysis), we include all bivariate and trivariate cross-classification variables. It is essential for our procedure to introduce the additional variables in blocks, and not to make some selection among them. (We shall return to this point later on.) First, we will focus on the so-called *discrimination measures* (Gifi, 1990, section 3.8.4), which are quantities that show how well a variable is represented as a group of category points in low-dimensional space.

Let $\mathbf{P}$ be the projection matrix that defines the optimal projection; the sth row of $\mathbf{P}$, which produces the projection on *component* (or *dimension*) s, is denoted by $\mathbf{p}_s$. We introduce a different, but consistent notation for the projected points to distinguish them from the high-dimensional ones. The projected profile points $\mathbf{x}_{ijk}$ are defined by $\mathbf{x}_{ijk} = \mathbf{P}\mathbf{z}_{ijk}$; the projected centroids are defined by $\mathbf{y}_{i**} = \mathbf{P}\mathbf{z}_{i**}$, $\mathbf{y}_{ij*} = \mathbf{P}\mathbf{z}_{ij*}$, and so on. The scalar value $y_{ij*}(\mathbf{p}_s) = \mathbf{p}_s'\mathbf{z}_{ij*}$ is the coordinate of the projection of the centroid for category ij of cross-classification variable AB on the component defined by $\mathbf{p}_s$. Discrimination measures then measure the dispersion of the projected category points as

$$\eta_A^2(\mathbf{p}_s) = \sum_i \pi_{i++} (y_{i**}(\mathbf{p}_s) - y_{***}(\mathbf{p}_s))^2, \tag{13}$$

$$\eta_{AB}^2(\mathbf{p}_s) = \sum_i \sum_j \pi_{ij+} (y_{ij*}(\mathbf{p}_s) - y_{***}(\mathbf{p}_s))^2, \tag{14}$$

for the original variables and cross-classification variables, respectively, where it will be clear how to continue for the higher-order interactions. Thus, $\eta_A^2(\mathbf{p}_s)$ is a weighted sum of squares of the category quantifications of variable A with respect to the overall center of gravity along component s. Since the weights (being probabilites) sum to one, it is the *variance* of the quantified categories in dimension $\mathbf{p}_s$ of the spatial model. The average discrimination measure across all variables on component s is denoted by λ_s^2, and is identical to the *eigenvalue* introduced in section 1. We are now ready to look at the results for our example.

The profile frequency matrix including all cross-classification variables is given in Table 2, where the observed count has been supplemented with the expected count under the hypothesis of

mutual independence. Five different homogeneity analyses were performed, always in two dimensions. The first analysis pertains to the original set of variables (*G, P, E, M*); the second uses the two-way cross-classification variables (*GP, GE, GM, PE, PM, EM*) only. The third analysis includes both the main effect variables and the two-way interactions (*G, P, E, M, GP, GE, GM, PE, PM, EM*), the fourth analysis adds the four three-factor cross-classification variables and the four-factor interaction to the second analysis, and finally the whole set (*G, P, E, M, GP, GE, GM, PE, PM, EM, GPE, GPM, GEM, PEM, GPEM*) was analyzed. The resulting discrimination measures are given in Table 3, along with the associated eigenvalues.

Table 2.

Profile Frequency Table for Marital Status Data with All Possible Cross-Classification Variables, Observed Count and Expected Count Under the Hypothesis of Independence.
G=Gender (1=female, 2=male), P=Premarital Sex (PMS: 1=yes, 2=no), E=Extramarital Sex (EMS: 1=yes, 2=no), and M=Marital Status (1=divorced, 2=married). GP=Two-Way Cross-Classification Gender × Premarital Sex (1=Female/Premarital Sex, 2=Female/No Premarital Sex, 3=Male/Premarital Sex, 4=Male/No Premarital Sex), etc. GPE=Three-Way Cross-Classification Gender × Premarital Sex × Extramarital Sex (1=Female/Premarital Sex/Extramarital Sex, 2=Female/Premarital Sex/No Extramarital Sex, 3=Female/No Premarital Sex/Extramarital Sex, 4=Female/No Premarital Sex/No Extramarital Sex, 5=Male/Premarital Sex/Extramarital Sex, 6=Male/Premarital Sex/No Extramarital Sex, 7=Male/No Premarital Sex/Extramarital Sex, 8=Male/No Premarital Sex/No Extramarital Sex), etc.

G	P	E	M	GP	GE	GM	PE	PM	EM	GPE	GPM	GEM	PEM	GPEM	obs	exp
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	17	8.76
1	1	1	2	1	1	2	1	2	2	1	2	2	2	2	4	9.61
1	1	2	1	1	2	1	2	1	3	2	1	3	3	3	54	66.23
1	1	2	2	1	2	2	2	2	4	2	2	4	4	4	25	72.66
1	2	1	1	2	1	1	3	3	1	3	3	1	5	5	36	28.89
1	2	1	2	2	1	2	3	4	2	3	4	2	6	6	4	31.70
1	2	2	1	2	2	1	4	3	3	4	3	3	7	7	214	218.47
1	2	2	2	2	2	2	4	4	4	4	4	4	8	8	322	239.69
2	1	1	1	3	3	3	1	1	1	5	5	5	1	9	28	4.66
2	1	1	2	3	3	4	1	2	2	5	6	6	2	10	11	5.12
2	1	2	1	3	4	3	2	1	3	6	5	7	3	11	60	35.27
2	1	2	2	3	4	4	2	2	4	6	6	8	4	12	42	38.70
2	2	1	1	4	3	3	3	3	1	7	7	5	5	13	17	15.39
2	2	1	2	4	3	4	3	4	2	7	8	6	6	14	4	16.88
2	2	2	1	4	4	3	4	3	3	8	7	7	7	15	68	116.34
2	2	2	2	4	4	4	4	4	4	8	8	8	8	16	130	127.65

In the first panel of Table 3, we see that all variables contribute to the first component, *Premarital Sex* and *Extramarital Sex* being the most important, while the second component is predominantly determined by *Gender* and *Marital Status*. The second panel reports the analysis with the bivariate cross-classification variables only, while the third panel reports the combined analysis. Comparing the third with the first and second panels, we see a remarkable similarity between the solutions: the first four discrimination measures of the combined analysis are about equal to those of the analysis with the original variables only, while the last six discrimination measures are about equal to those of the analysis with the bivariate cross-classification variables only. To be completely clear, we stress that we do not obtain perfectly identical results, but very similar ones.

Table 3.

Discrimination Measures, Eigenvalues, and Average Discrimination Measures for Partitions from Five Different Homogeneity Analyses with Increasing Number of Cross-Classified Variables

	Analysis 1		Analysis 2		Analysis 3		Analysis 4		Analysis 5	
	dim 1	dim 2	dim 1	dim 2	dim 1	dim 2	dim 1	dim 2	dim 1	dim 2
G	0.259	0.525			0.254	0.510			0.252	0.501
P	0.538	0.055			0.553	0.059			0.558	0.061
E	0.432	0.092			0.428	0.094			0.425	0.095
M	0.319	0.365			0.314	0.373			0.312	0.377
GP			0.656	0.502	0.650	0.512	0.660	0.494	0.654	0.505
GE			0.610	0.655	0.615	0.664	0.605	0.648	0.611	0.657
GM			0.560	0.883	0.566	0.887	0.558	0.878	0.563	0.883
PE			0.808	0.213	0.804	0.206	0.810	0.220	0.806	0.213
PM			0.733	0.522	0.729	0.513	0.738	0.530	0.733	0.520
EM			0.602	0.415	0.606	0.407	0.599	0.420	0.603	0.412
GPE							0.888	0.680	0.886	0.685
GPM							0.845	0.916	0.843	0.918
GEM							0.789	0.945	0.795	0.948
PEM							0.910	0.593	0.908	0.581
GPEM							1.000	1.000	1.000	1.000
λ^2	0.387	0.259	0.662	0.532	0.552	0.422	0.764	0.666	0.663	0.557
$1/4\ \Sigma_1^4 \eta^2$	0.387	0.259			0.387	0.259			0.387	0.258
$1/6\ \Sigma_5^{10} \eta^2$			0.662	0.532	0.662	0.532	0.662	0.532	0.662	0.532
$1/5\ \Sigma_{11}^{15} \eta^2$							0.886	0.827	0.886	0.826
$\Sigma\ \eta^2$	1.549	1.037	3.972	3.192	5.520	4.220	8.402	7.323	9.948	8.356

Since eigenvalues are averages of discrimination measures, the eigenvalues of the combined analysis are about equal to 0.4 times the eigenvalues of the first panel plus 0.6 times the eigenvalues of the second panel, or, equivalently, the sum over all discrimination measures per dimension in the first and second analyses together is about equal to the sum over all discrimination measures in the third analysis, and so on. The overall similarity between the results for cross-classification variables included or excluded, is only obtained if cross-classification variables are included in a completely balanced way. Otherwise, the similarity would be lost.

How do we recognize independence from these tables? We discuss this question in two steps. First, for a precise judgment we need a standard of comparison, because the expected level of a discrimination measure depends on the number of categories. As our standard, we choose the *expected value* of a discrimination measure under the hypothesis of mutual independence (alternatively, the quantities that we call expected value could also be interpreted as the mean discrimination measure across all components). When we consider all higher-order cross-classification variables, including the highest one corresponding to a saturated model, starting with m original variables we will have $2^m - 1$ analysis variables. If the total number of categories of the original variables is denoted by $q = n_A + n_B + n_C + ...$, then there are $q - m$ non-trivial components to consider. Under the hypothesis of mutual independence, these components will have equal eigenvalues. We derive the expected discrimination measure $\eta^2_A(*)$ for one of the original variables as $(n_A - 1)/(q - m)$, the expected discrimination measure $\eta^2_{AB}(*)$ for one of the two-way analysis variables as $(n_A + n_B - 2)/(q - m)$, the expected discrimination measure $\eta^2_{ABC}(*)$ for one of the three-way analysis variables as $(n_A + n_B + n_C - 3)/(q - m)$, and so on. In our example, where the variables G, P, E and M have two categories each, under mutual independence $\eta^2_G(\mathbf{p}_S)...\eta^2_M(\mathbf{p}_S)$ will be equal to 0.25; $\eta^2_{GP}(\mathbf{p}_S)...\eta^2_{EM}(\mathbf{p}_S)$ will be equal to 0.50; and $\eta^2_{GPE}(\mathbf{p}_S)...\eta^2_{PEM}(\mathbf{p}_S)$ will be equal to 0.75. The discrimination measure $\eta^2_{GPEM}(\mathbf{p}_S)$ will be equal to 1.00, representing perfect fit, which corresponds to the saturated model in a loglinear analysis.

Secondly, in the previous section we have seen that independence implies additivity of category quantifications. We will now show that under two-way independence the discrimination measures

are additive, too. For instance, the fact that *Gender* and *Marital Status* are independent (χ^2_{GM} = 0.031) is reflected in the discrimination measures $\eta^2_{GM}(\mathbf{p}_S)$ = (0.563, 0.883) being approximately equal to $\eta^2_G(\mathbf{p}_S)$ = (0.252, 0.501) plus $\eta^2_M(\mathbf{p}_S)$ = (0.312, 0.377). Geometrically, departure from independence can be depicted as a distance between two vectors that represent the discrimination measures in two-space (see Figure 1), the first vector (with coordinates 0.563, 0.883) displaying the observed discrimination measure and the second vector (0.565 0.878) displaying the expected discrimination measure when G and M are independent. For GM, this distance is 0.005; for the other two-way interactions, these distances are 0.167 (GP), 0.090 (GE), 0.186 (PE), 0.161 (PM), and 0.146 (EM) respectively. This pattern shows a very close resemblance with the results from the loglinear analysis with two-way interactions only, as can be seen from Agresti (1990, Table 7.4).

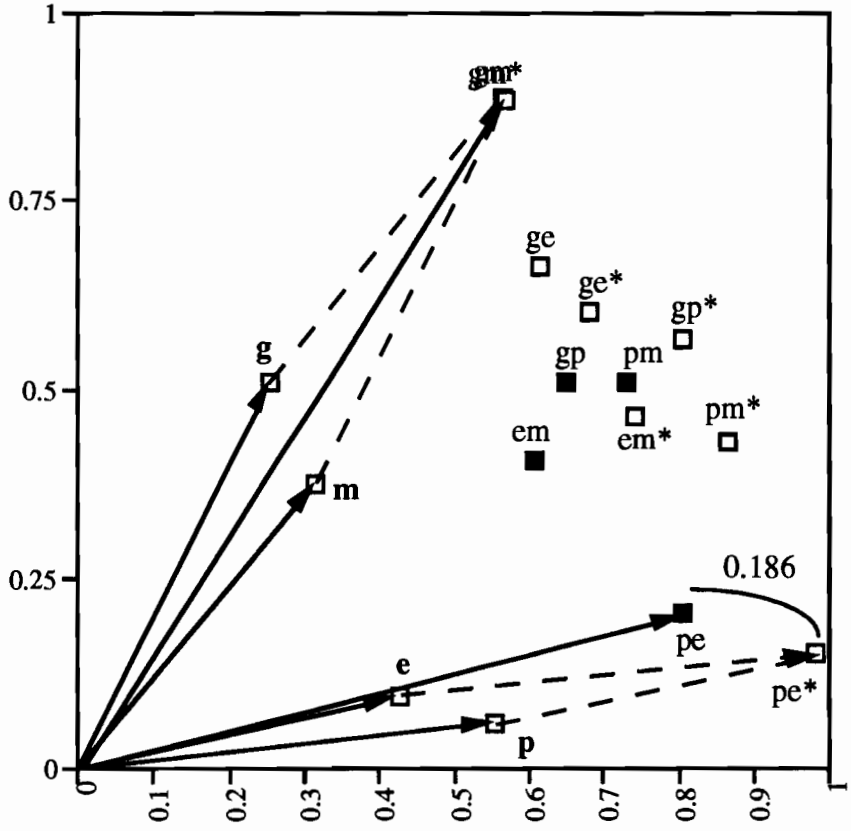

Figure 1. Display of (in)dependence through discrimination measures: distances between vectors for observed and expected (*) two-way discrimination measures. Interaction effects as found by Agresti are indicated with black squares.

To show that additivity must hold in any component, we first note that additivity in high-dimensional space (e.g., $\mathbf{z}_{ij*} = \mathbf{z}_{i**} + \mathbf{z}_{*j*}$) carries through to low-dimensional space by virtue of the distributive character of projection: $\mathbf{p}_s'\mathbf{z}_{ij*} = \mathbf{p}_s'\mathbf{z}_{i**} + \mathbf{p}_s'\mathbf{z}_{*j*}$. Looking at marginal independence, then by substituting $\pi_{ij+} = \pi_{i++}\pi_{+j+}$ and $y_{ij*}(\mathbf{p}_s) = y_{i**}(\mathbf{p}_s) + y_{*j*}(\mathbf{p}_s)$ into (14), we obtain

$$\begin{aligned}\eta^2_{AB}(\mathbf{p}_s) &= \textstyle\sum_i\sum_j \pi_{i++}\pi_{+j+}\,[(y_{i**}(\mathbf{p}_s) - y_{***}(\mathbf{p}_s)) + (y_{*j*}(\mathbf{p}_s) - y_{***}(\mathbf{p}_s))]^2 \\ &= \textstyle\sum_i \pi_{i++}\,(y_{i**}(\mathbf{p}_s) - y_{***}(\mathbf{p}_s))^2 + \sum_j \pi_{+j+}\,(y_{*j*}(\mathbf{p}_s) - y_{***}(\mathbf{p}_s))^2 \\ &= \eta^2_A(\mathbf{p}_s) + \eta^2_B(\mathbf{p}_s)\,, \end{aligned} \tag{15}$$

where the cross-product vanishes because $\sum_i \pi_{i++}(\mathbf{z}_{i**} - \mathbf{z}_{***}) = \sum_j \pi_{+j+}(\mathbf{z}_{*j*} - \mathbf{z}_{***}) = \mathbf{0}$ by definition of $\mathbf{z}_{***}$, and therefore any projected value must be 0. In a similar way we obtain, for the case in which variable A is jointly independent of B and C,

$$\eta^2_{ABC}(\mathbf{p}_s) = \eta^2_A(\mathbf{p}_s) + \eta^2_{BC}(\mathbf{p}_s)\,, \tag{16}$$

and when A and B are conditionally independent given C

$$\eta^2_{ABC}(\mathbf{p}_s) = \eta^2_{AC}(\mathbf{p}_s) + \eta^2_{BC}(\mathbf{p}_s) - \eta^2_C(\mathbf{p}_s)\,. \tag{17}$$

While these relationships are exact when the stipulated type of independence is exactly fulfilled, for the "no three-way interaction" case, we must do something different. One possible idea would be to settle for an approximation. For instance, using Darroch's (1962) condition of a "perfect" table (which does not exhibit paradoxes), "no three-way interaction" implies that $\pi_{ijk} = (\pi_{ij+}\pi_{i+k}\pi_{+jk})/(\pi_{i++}\pi_{+j+}\pi_{++k})$, and from this condition we may derive the approximate relationship

$$\eta^2_{ABC}(\mathbf{p}_s) = \eta^2_{AB}(\mathbf{p}_s) + \eta^2_{AC}(\mathbf{p}_s) + \eta^2_{BC}(\mathbf{p}_s) - \eta^2_A(\mathbf{p}_s) - \eta^2_B(\mathbf{p}_s) - \eta^2_C(\mathbf{p}_s)\,. \tag{18}$$

At this point, some experimentation indicated that this is not the way to go; instead, it seems that to test the "no three-way interaction" case, we should rely on a higher-order statistic as well, in

contrast to the discrimination measure, which we could argue is a two-way statistic. The suggested diagnostics would then be the category quantifications, and in the next section these will be used to demonstrate that they indeed display three-way interactions.

5. Visual Display of Odds Ratios as Distance Ratios Between Category Points

In section 3, we have seen how odds are represented as ratios of distances, so that an odds ratio of one corresponds to parallel lines. Two-way association leads to non-parallel lines, three-way association leads to different non-parallel lines, conditional upon one fixed variable. At this point we will display the results from the extended homogeneity analysis in two dimensions, including the first and second-order cross classification variables. The category quantifications will be labeled with their level, for instance, g_1 and g_2 denote the female and male categories, respectively. Similarly, the label p_1e_1 denotes respondents who reported both premarital and extramarital sex. In the second-order interactions, $g_1p_1e_2$ denotes the category of female respondents who did report premarital sex but no extramarital sex, and $p_2e_1m_1$ respondents who did not report premarital sex, did report extramarital sex, and are divorced. This notation will also be used in the equations given below that give the distance ratios between selected category points to study two particular higher-way interactions, i.e., the three-way interaction between Premarital Sex, Extramarital Sex, and Marital Status (*PEM*), and between Gender, Premarital Sex, and Extramarital Sex (*GPE*). We already know that *G* and *M* are independent, so all higher-order interactions that include *GM* are not very interesting.

According to Agresti (1990, p. 221), the *PEM* interaction seems vital to explaining relationships in the data. To describe this PEM interaction, Agresti uses the estimated odds ratios for the loglinear model (*GP*, *GM*, *GE*, *PEM*), and concludes: "Given gender, for those who reported premarital sex, the odds of a divorce are estimated to be 1.82 times higher for those who reported extra-marital sex than for those who did not; for those who did not report premarital sex, the odds of a divorce are estimated to be 10.94 times higher for those who reported extramarital sex than for those who did not".

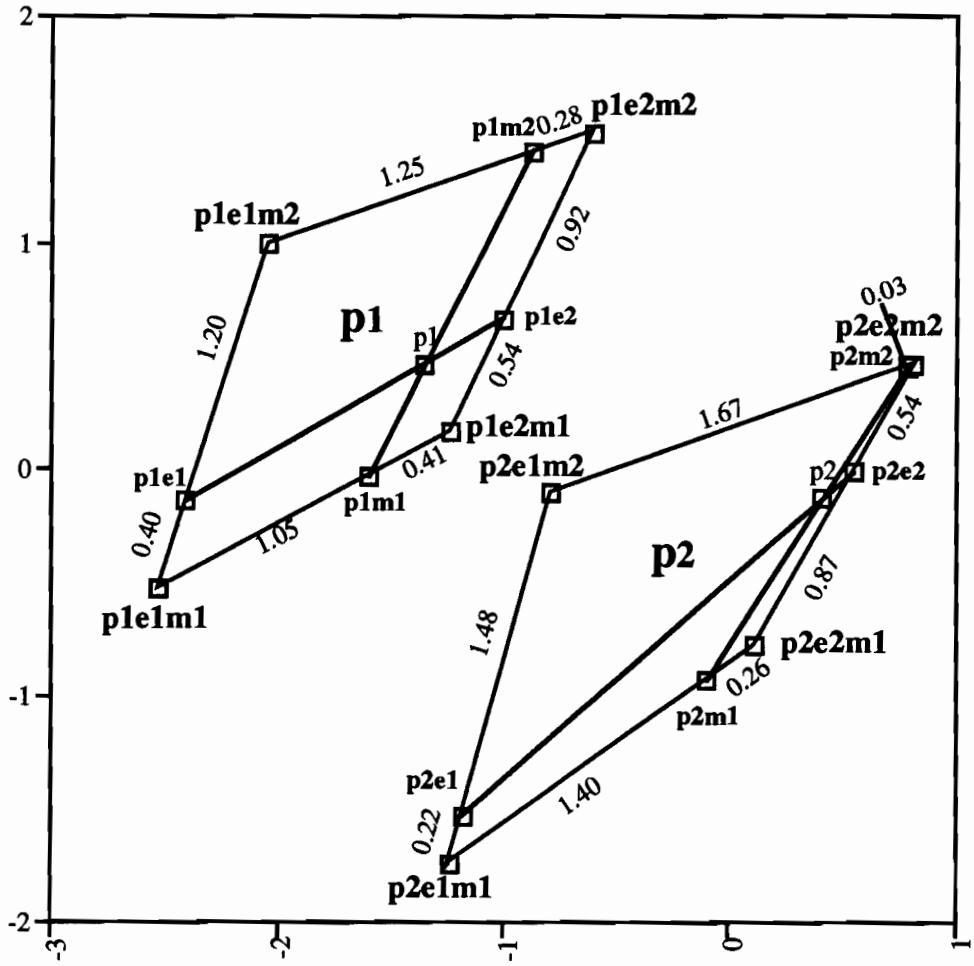

Figure 2. Display of three-way interaction PEM: Category points for Extramarital Sex and Marital Status connected for the two levels of Premarital Sex.

We translate this EM estimated odds ratio for the two levels of P in a spatial model (see Figure 2); in this figure we have used only the category points relevant for this particular interaction. For the two different levels of P, p_1 and p_2, category points for levels of E and M are connected, to form two diamond shapes. Along the edges of each diamond, the distances are given between the four three-way points and their centroids, the two-way points. The closer a two-way point to a three-way point, the more respondents are in that particular three-way point. So we see in Figure 2, that p_1m_2 (premarital sex, married) is closer to $p_1e_2m_2$ (no extramarital sex) than to $p_1e_1m_2$ (extramarital sex); and that p_2e_2 (no premarital sex, no extramarital sex) is closer to $p_2e_2m_2$ (married) than to $p_2e_2m_1$ (divorced). From this we would deduce that extramarital sex is not beneficial to marriage. If we compute the distance ratio for p_1 with respect to the p_1e_1 and p_1e_2 centroids,

$$\frac{d(p_1e_1m_2,\ p_1e_1)\ d(p_1e_2m_1,p_1e_2)}{d(p_1e_1m_1,\ p_1e_1)\ d(p_1e_2m_2,p_1e_2)} = 1.76 \cong \frac{1.20 \times 0.54}{0.40 \times 0.92} \qquad (19)$$

(where $\cong$ denotes equality upto rounding errors) and with respect to centroids p_1m_2 and p_1m_1,

$$\frac{d(p_1e_1m_2,\ p_1m_2)\ d(p_1e_2m_1,\ p_1m_1)}{d(p_1e_2m_2,\ p_1m_2\ d(p_1e_1m_1,\ p_1m_1)} = 1.76 \cong \frac{1.25 \times 0.41}{0.28 \times 1.05} \qquad (20)$$

we note that the equality between the two different ratios is preserved. So, in the sequel we need to look at only one of each pair of ratios. We also remark that the estimated odds ratio of 1.82 reported by Agresti (1990) is indeed close to 1.76. If we now inspect the distance ratio for the p_2 category (those respondents who did not report premarital sex), we obtain

$$\frac{d(p_2e_1m_2,\ p_2m_2)\ d(p_2e_2m_1,\ p_2m_1)}{d(p_2e_2m_2,\ p_2m_2)\ d(p_2e_1m_1,\ p_2m_1)} = 10.62 \cong \frac{1.67 \times 0.26}{0.03 \times\ 1.40} \qquad ,(21)$$

and the estimated odds ratio 10.94 reported by Agresti is again very close to this figure. So the main conclusion of the loglinear analysis that the effect of extramarital sex on divorce is much greater for respondents who did not report premarital sex, is displayed graphically in the homogeneity analysis solution. As usual, there are two companion pairs of odds ratios; we first look at the distance ratios for the two categories of the variable Extramarital Sex. The diamonds for levels e_1 and e_2 are to be found in Figure 3, accompanied with the associated distances.

The distance ratio for those who did report extramarital sex is obtained from

$$\frac{d(p_1e_1m_2,\ e_1m_2)\ d(p_2e_1m_1,\ e_1m_1)}{d(p_2e_1m_2,\ e_1m_2)\ d(p_1e_1m_1,\ e_1m_1)} = 0.45 \cong \frac{0.58 \times 0.81}{1.09 \times\ 0.96}. \qquad (22)$$

Agresti gives the estimated *PM* odds ratio for category e_1 as 0.50, so among those who reported extramarital sex, divorce is about two times more likely for respondents with no premarital sex than for those who had premarital sex. For those who did not report extramarital sex,

$$\frac{d(p_1e_2m_2,\ e_2m_2)\ d(p_2e_2m_1,\ e_2m_1)}{d(p_2e_2m_2,\ e_2m_2)\ d(p_1e_2m_1,\ e_2m_1)} = 2.73 \cong \frac{1.52 \times 0.47}{1.17 \times\ 0.23}. \qquad (23)$$

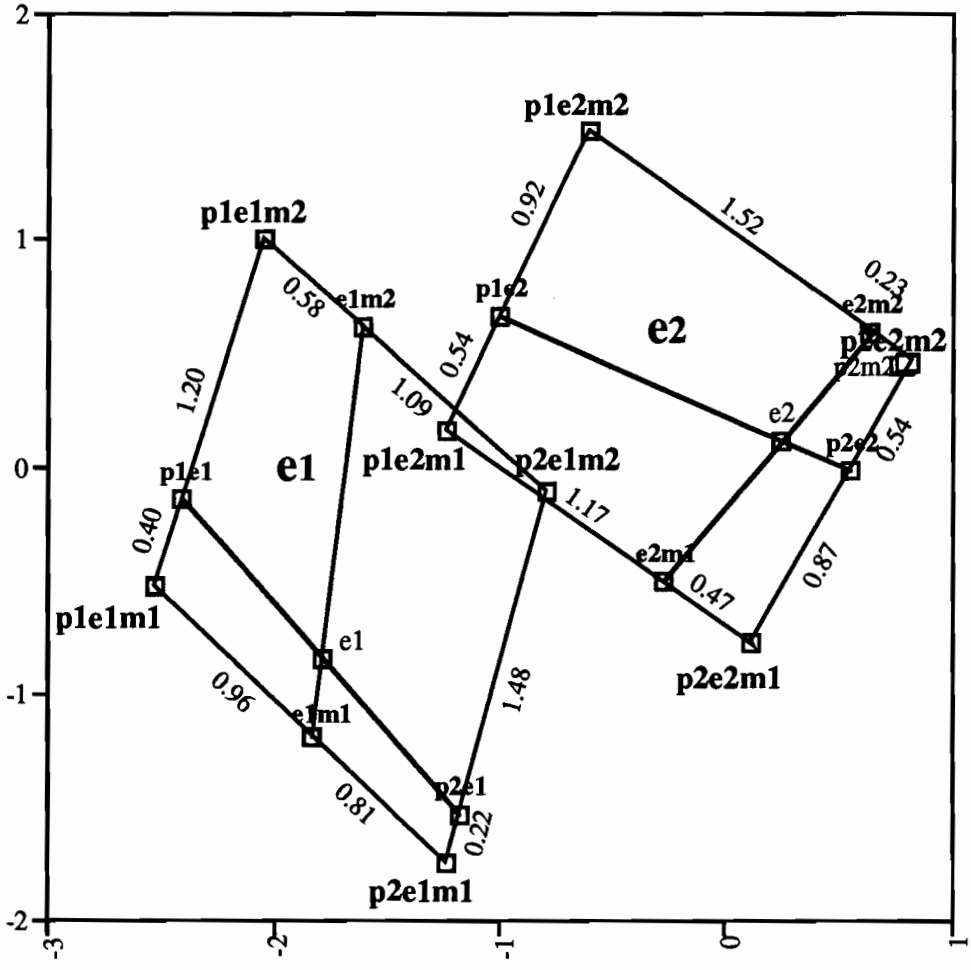

Figure 3. Display of three-way interaction PEM: Category points for Premarital Sex and Marital Status connected for the two levels of Extramarital Sex.

The estimated *PM* odds ratio for category e_2 is reported as 3.00 by Agresti (1990, p. 222): among those who did not report extramarital sex, divorce is much more likely for respondents who had premarital sex than for those who had no premarital sex. Finally, with respect to the levels m_1 and m_2, Agresti reports the *PE* estimates as 1.82 for divorced and 10.95 for married respondents; from the distances, we recover the odd ratios

$$\frac{d(p_2e_1m_1,\, p_2m_1)\; d(p_1e_2m_1,\, p_1m_1)}{d(p_2e_2m_1,\, p_2m_1)\; d(p_1e_1m_1,\, p_1m_1)} = 2.10 \cong \frac{1.40 \times 0.41}{0.26 \times\; 1.05}; \tag{24}$$

$$\frac{d(p_2e_1m_2,\, p_2m_2)\; d(p_1e_2m_2,\, p_1m_2)}{d(p_2e_2m_2,\, p_2m_2\; d(p_1e_1m_2,\, p_1m_2)} = 12.65 \cong \frac{1.67 \times 0.28}{0.03 \times\; 1.25} \tag{25}$$

for m_1 and m_2, respectively. The corresponding diamonds are given in Figure 4.

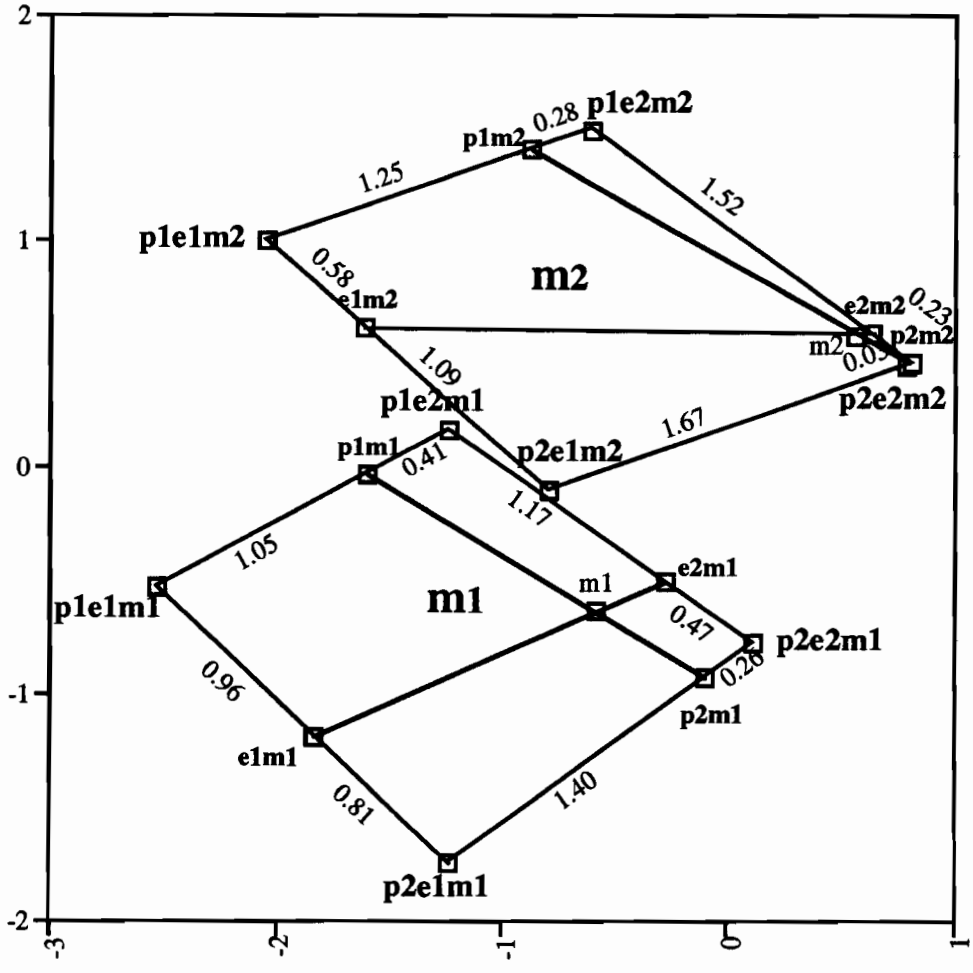

Figure 4. Display of three-way interaction PEM: Category points for Premarital Sex and Extramarital Sex connected for the two levels of Marital Status.

As a final illustration of this very special property of category quantifications in terms of three-way interactions, we inspect the *GPE* interaction as well, which should be identified, following Agresti, as a "no three-way interaction" case. The *GP* distance ratio was obtained as 0.283 for those who reported extra-marital sex and 0.286 for those who did not. So there is only two-way interaction, to the effect that about 3.6 times more men than women had premarital sex. The *GE* distance ratio is 0.695 for those who reported pre-marital sex, and 0.704 for those who did not. Again, there is only two-way interaction: 1.4 times more men than women had extramarital sex.

Finally, the *PE* distance ratio is 3.56 for women and 3.61 for men: those who had premarital sex were 3.6 times more likely to have extramarital sex compared to those who had not, but gender has no effect on the relation between *P* and *E*.

6. Discussion

A major point of this paper is that the use of homogeneity analysis does not need to rely on the assumption that the higher-order interactions among the categorical variables are nonsignificant. We first proposed a procedure that uses homogeneity analysis to display the higher-order interactions in a $2 \times 2 \times 2 \times 2$ contingency table directly. The multiway contingency table was first transformed into a profile frequency table. Then higher-way cross-classification variables were added in a completely balanced way. We demonstrated from the solution of such an extended homogeneity analysis how the higher-way interactions are represented in the visual display.

It was shown that the condition "no two-way interaction" could be expressed exactly in terms of the discrimination measures; if two variables are independent, their discrimination measures add up to the discrimination measure of their cross-classification variable. "No two-way interaction" is also expressed in ratios between distances between particular category points. If there is no two-way interaction, the latter ratio is equal to 1.00. These distance ratios were then shown to be the major diagnostic for identifying three-way interaction. If the second-order interaction is significant, the distance ratio based on each pair of two variables will differ substantially for the different levels (categories) of the third variable.

To simplify the computation while using existing software, we would recommend not to use standard multiple correspondence analysis or homogeneity analysis programs (as in the SPSS HOMALS procedure). This would amount to an analysis of a much larger matrix than is actually required, because the number of profiles is much smaller than the total number of individual objects.

Table 4.
Weighted Indicator Supermatrix for Marital Status Data to be Used as Input for Simple Correspondence Analysis

17	0	17	0	17	0	17	0
4	0	4	0	4	0	0	4
54	0	54	0	0	54	54	0
25	0	25	0	0	25	0	25
36	0	0	36	36	0	36	0
4	0	0	4	4	0	0	4
214	0	0	214	0	214	214	0
322	0	0	322	0	322	0	322
0	28	28	0	28	0	28	0
0	11	11	0	11	0	0	11
0	60	60	0	0	60	60	0
0	42	42	0	0	42	0	42
0	17	0	17	17	0	17	0
0	4	0	4	4	0	0	4
0	68	0	68	0	68	68	0
0	130	0	130	0	130	0	130

Instead, simple correspondence analysis could be applied, for example, the SPSS ANACOR procedure. To apply simple correspondence analysis, we would only have to replace each column in the profile matrix by its indicator matrix, collect the indicator matrices in an indicator supermatrix, and premultiply the latter with a diagonal matrix, containing the corresponding profile frequency on its main diagonal. Such a table for the four original variables in our example is given in Table 4.

In the empirical cases we have analyzed so far, the row scores in an analysis with only the main effect variables were always very similar to those with all cross-classification variables added. This suggests that in practice we would not need to include all the cross-classification variables, but could derive the higher-order category quantifications from a simple analysis, by computing the appropriate centroids of profile points afterwards. The resulting visual displays of interaction through the use of diamonds must be very similar as well. In fact, although the diamonds may be slightly different (due to a somewhat different projection from high-dimensional space), the distance ratios they display are identical.

From the row scores of a simple correspondence analysis of a table as shown in Table 4, the higher-way centroid for the category point for $g_1p_1e_1$, for example, is obtained as 17 times the first row score (for profile 1111) plus 4 times the second row score (for profile 1112) divided by 21;

the category point for $g_1e_1m_1$ is 17 times the first row score (for profile 1111) plus 36 times the fifth row score (for profile 1211) divided by 17+36, and so on. Which combinations should be taken follows from the associated columns in the extended profile frequency matrix in Table 2.

To compare our procedure with already existing ones, the following observations are important. First, our spatial representation is totally different from the usual geometric model used in the theory of loglinear analysis (Fienberg and Gilbert, 1970), which considers the distribution of mass over the cells of the table as one point in a regular polygon. Relationships among the two would be a subject for further study. Second, there are at least two other approaches to the use of cross-classification variables, which are, however, different from ours. In Gifi (1990), it is proposed to replace two of the original (main effect) variables by one cross-classification variable, with the aim of removing "uninteresting" association with respect to the main object of study. Van der Heijden and De Leeuw (1985) use the idea of cross-classification with the aim to study residuals from higher-order independence models. They apply simple correspondence analysis to a matrix of order, say, $n_A \times n_{BC}$; as they remark, it is often not obvious which two out of three variables should be cross-classified. We have shown in this paper that when using our method, no such choice has to be made, while at the same time possible effects can be identified.

Bishop, Fienberg and Holland (1975, p. 24) remark about the possibility of a linear (additive) model in the cell probabilities instead of their logarithms: "We conclude that the difficulty of relating the additive model to the concept of independence makes it less attractive than the loglinear model". The profile scores from homogeneity analysis are additive combinations of category quantifications, and we have seen that they are related to the concept of independence in a rather simple way. The apparent contradiction is resolved once we realize that the scores from homogeneity analysis do not represent the cell probabilities, but the cell itself (the profile). The spatial representation aims at predicting an answer pattern given the profile score, not a probability given the answer pattern.

We have not elaborated on the case where variables contain more than two categories. Some experimentation has shown that the special properties in terms of discrimination measures are preserved for the multi-category case. With respect to the much more complicated distance ratios,

the promising results obtained by applying the general approach proposed in this paper to the multi-category case are currently being scrutinized.

7. References

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.

Bartlett, M.S. (1935). Contingency table interactions. *Journal of the Royal Statistical Society, Supplement 2*, 248-252.

Benzécri, J.-P. (Ed.) (1973). *L'Analyse des Donneés, Tome (Vol.) 2: L'Analyse des Correspondances*. Paris: Dunod.

Birch, M.W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society, Series B, 25*, 220-233.

Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: The MIT Press.

Darroch, J.N. (1962). Interaction in multi-factor contingency tables. *Journal of the Royal Statistical Society, Series B, 24*, 251-263.

Fienberg, S.E. (1980). *The Analysis of Cross-Classified Categorical Data* (2nd ed.). Cambridge, MA: The MIT Press.

Fienberg, S.E., and Gilbert, J.P. (1970). The geometry of a 2×2 contingency table. *Journal of the American Statistical Association, 65*, 694-701.

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. New York: Wiley.

Gilbert, G.N. (1981). *Modelling Society*. London: George Allen and Unwin.

Greenacre, M.J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In: P. Horst (Ed.), *The Prediction of Personal Adjustment*, pp. 319-348. New York: Social Science Research Council.

Nishisato, S. (1994). *Elements of dual scaling: An introduction to practical data analysis*. New Jersey: Lawrence Erlbaum.

Thornes, B., and Collard, J. (1979). Who divorces? London: Routledge & Kegan Paul.

Van de Geer, J.P. (1993). *Analysis of Categorical Data (2 vols.).* Newbury Park, CA: Sage.

Van der Heijden, P.G.M., and de Leeuw, J. (1985). Correspondence analysis used complementary to loglinear analysis. *Psychometrika, 50,* 429-447.

Yule, G.U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society, 75,* 579-642.