Algebraic Descriptions of Nominal Multivariate Discrete Data

J.L. Teugels[1] & J. Van Horebeek[1,2]

[1] Department of Mathematics, Katholieke Universiteit Leuven, Heverlee, Belgium

[2] CIMAT, Guanajuato, Mexico

Abstract

Traditionally, multivariate discrete data are analyzed by means of log-linear models. In this paper we show how an algebraic approach leads naturally to alternative models, parametrized in terms of the moments of the distribution. Moreover we derive a complete characterization of all meaningful transformations of the components and show how transformations affect the moments of a distribution. It turns out that our models provide the necessary formal description of longitudinal data; moreover in the classical case, they can be considered as an analysis tool, complementary to log-linear models.

1 Introduction

We start with a given multivariate discrete nominal variable $\mathbf{X}$. Questions of interest about $\mathbf{X}$ can be roughly divided into two groups. One group is related to conditional characteristics such as conditional independencies or questions concerning the sign and/or magnitude of *log-odds* ratios. The other group focuses on marginal characteristics such as marginal independencies or multivariate moments like covariances.

As indicated by Goodman [5], measuring interactions between variables in terms of *log-odds* ratios should be considered complementary to those in terms of covariance/correlation. In practice one often resorts to a *log-linear* model because (*i*) it is very suitable in the detection of conditional characteristics, (*ii*) it has very attractive properties and (*iii*) it allows several modifications to incorporate, up to a certain level, characteristics of the marginal distribution (e.g. [14], [15] and [20]).

Nevertheless, in some situations one requires an exhaustive model in terms of the marginal characteristics. This may be caused by the design of the experiment where, for example, subsampling was used keeping some marginals fixed at given values. Or the investigator of some categorical longitudinal data is interested in testing hypotheses such as marginal homogeneity [9], in pairwise independence [7], symmetry, etc.

In the first part of the paper, we will define algebraic operators that lead to a parametrization in terms of the moments. We show how the operators transform the cell probabilities into new parameters that can be easily characterized. We review the basic ideas as formulated in [16], [18], [19] and independently in [4], and formulate a unifying framework. The underlying motivation is to develop a conceptually rich and general model, instead of focusing on the numerical conditions of how to adapt a *log-linear* model to test the above mentioned hypotheses (see [1] for a recent overview).

In the second part we similarly develop a complete characterization of all meaningful transformations of nominal data and show its impact on the parametrization by means of moments. Consequently, we will obtain a generalization of the results derived by [3]. Both parts will be illustrated with concrete data.

In the sequel we use the following notations for a given multivariate discrete variable: $\mathbf{X} = (X_1, \cdots, X_n)$, $\mathbf{X} \in MD(r_1, \cdots, r_n) \Leftrightarrow X_i \in \{0, \cdots, r_i - 1\}, 1 \leq i \leq n$; further $p_{i_1,\cdots,i_n} := P(X_1 = i_1, \cdots, X_n = i_n)$ for the joint distribution and $X_i \perp X_j | X_k$ iff X_i and X_j are conditionally independent, given X_k.

2 Block models

Suppose that a discrete multivariate distribution is given by its cell probabilities $\{p_{i_1,\ldots,i_n}\}$. We need to express the distribution into other, more interpretative quantities that shed some light on the interactions between the marginals. We can order the cell probabilities in a huge vector and - assuming that interesting transformations are linear- look for an underlying transformation matrix $\mathbb{A}$ as is shown in the next figure.

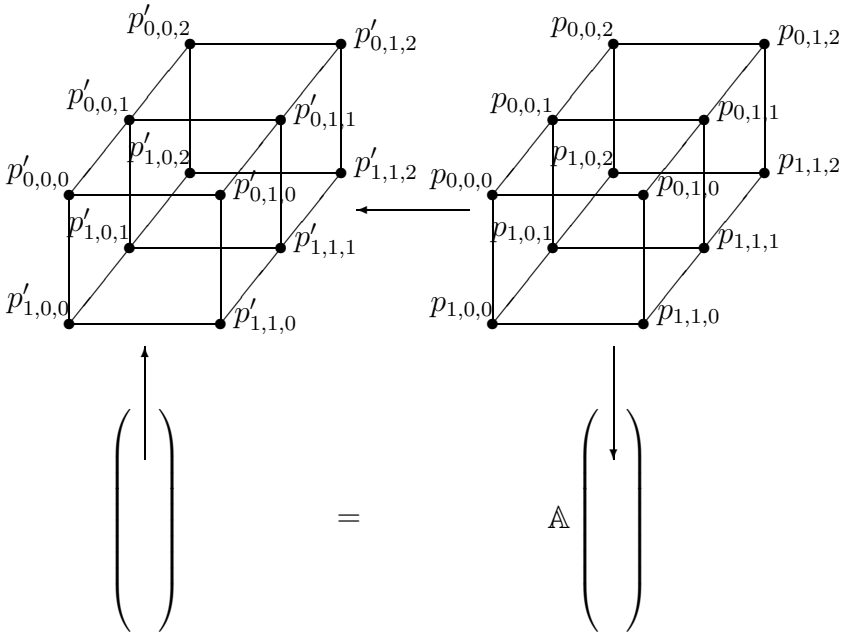


Figure 1

Due to the vectorization, the spatial structure of the variable is completely lost. In this section we introduce *blocks* to solve this problem. A block can be considered as a matrix-type structure with potentially more than two indices. Because of the similarity with matrices, many properties of and operators on matrices can be generalized to blocks.

2.1 Blocks and flats

Definition 2.1 *Define* $M(r_1, \cdots, r_n) = \mathbb{R}^{\{0,\cdots,r_1-1\}\times\cdots\times\{0,\cdots,r_n-1\}}$ *where* $\mathbb{R}$ *denotes the set of real numbers; any member of* $M(r_1, \cdots, r_n)$ *is called a block.*

The case $n = 2$ corresponds to matrices. The definition of equality of blocks is obvious. We accept the convention that indices of blocks always start at 0. Blocks can be represented graphically in different forms. In what follows, we will make use of hypercubes as shown in Figure 1 and Figure 2 below. An alternative can be found in [8].

We define two auxiliary concepts involving matrices.

Definition 2.2 *A flat* $(\mathbb{A}^1|\cdots|\mathbb{A}^n)$ *is an ordered sequence of matrices where* $\mathbb{A}^i \in M(r_i, s_i)$, $1 \leq i \leq n$. *We call a matrix a scrambler if its elements belong to* $\{0, 1\}$ *and if in every column there is exactly one* 1. *A flat is a scrambler-flat if it is built up with scramblers.*

It will be useful to define a straightforward addition and multiplication between flats.

1. *addition of flats*

   if $A = (\mathbb{A}^1|\cdots|\mathbb{A}^n)$ and $B = (\mathbb{B}^1|\cdots|\mathbb{B}^n)$ with $\mathbb{A}^i, \mathbb{B}^i \in M(r_i, s_i)$, define:

$$A + B = (\mathbb{A}^1 + \mathbb{B}^1|\cdots|\mathbb{A}^n + \mathbb{B}^n);$$

2. *multiplication of flats*

   if $A = (\mathbb{A}^1|\cdots|\mathbb{A}^n)$ and $B = (\mathbb{B}^1|\cdots|\mathbb{B}^n)$ with $\mathbb{A}^i \in M(r_i, s_i)$, $\mathbb{B}^i \in M(s_i, t_i)$, define:

$$A.B = (\mathbb{A}^1\mathbb{B}^1|\cdots|\mathbb{A}^n\mathbb{B}^n).$$

The first two of the following concepts are familiar in matrix calculus and allow an easy extension to blocks; the third seems to be new.

Definition 2.3 *If $\mathcal{B} \in M(r_1, \cdots, r_n)$, define the vectorization operator $vec(\mathcal{B})$ as*

$$vec(\mathcal{B})_k = \mathcal{B}_{k_1,\ldots,k_n}$$

with $k = k_1 + k_2 r_1 + k_3 r_1 r_2 + \cdots + k_n \prod_{i=1}^{n-1} r_i$ and $0 \leq k_i < r_i$.

Definition 2.4 *For any two blocks $\mathcal{B}^1 \in M(r_1, \cdots, r_n)$ and $\mathcal{B}^2 \in M(s_1, \cdots, s_n)$ we define the Kronecker product $\mathcal{B}^1 \otimes \mathcal{B}^2 \in M(r_1 s_1, \cdots, r_n s_n)$ as*

$$(\mathcal{B}^1 \otimes \mathcal{B}^2)_{i_1,\ldots,i_n} = \mathcal{B}^1_{j_1,\ldots,j_n}\mathcal{B}^2_{k_1,\ldots,k_n}$$

with $i_l = j_l s_l + k_l$, $1 \leq l \leq n$ and $0 \leq k_l < s_l$.

Definition 2.5 *If A is a flat $(\mathbb{A}^1|\cdots|\mathbb{A}^n)$ with $\mathbb{A}^i \in M(r_i, s_i)$ and $\mathcal{B} \in M(s_1, \cdots, s_n)$, define the flat-product $A \triangleright \mathcal{B} \in M(r_1, \cdots, r_n)$ as*

$$(A \triangleright \mathcal{B})_{i_1,\ldots,i_n} = \sum_{k_1=0}^{s_1-1} \mathbb{A}^1_{i_1,k_1} \sum_{k_2=0}^{s_2-1} \mathbb{A}^2_{i_2,k_2} \cdots \sum_{k_n=0}^{s_n-1} \mathbb{A}^n_{i_n,k_n} \mathcal{B}_{k_1,\ldots,k_n}.$$

Many pleasant properties for the above concepts can now be derived; most of them illustrate how blocks are generalizations of matrices.

Property 2.1

1. *If $A = (\mathbb{A}), \mathbb{A} \in M(r, s), \mathcal{C} \in M(s)$:*

$$A \triangleright \mathcal{C} = \mathbb{A}\mathcal{C}; \tag{1}$$

2. *If $\mathbb{A} \in M(r_1, s_1), \mathbb{B} \in M(r_2, s_2)$ and $\mathcal{C} \in M(s_1, s_2)$:*

$$(\mathbb{A}|\mathbb{B}) \triangleright \mathcal{C} = \mathbb{A}\mathcal{C}\mathbb{B}^t; \tag{2}$$

3. *In general:*

$$A \triangleright (\alpha\mathcal{C} + \beta\mathcal{D}) = \alpha A \triangleright \mathcal{C} + \beta A \triangleright \mathcal{D}; \tag{3}$$

$$vec(A \triangleright \mathcal{B}) = (\mathbb{A}^n \otimes \cdots \otimes \mathbb{A}^1) vec(\mathcal{B}); \tag{4}$$

$$A \triangleright (B \triangleright \mathcal{C}) = (A.B) \triangleright \mathcal{C}. \tag{5}$$

Proof:

Relations (1), (2), (3) and (5) are derived by applying the definition.

Relation (4) is obtained as follows:

$$vec(A \triangleright \mathcal{B})_k = (A \triangleright \mathcal{B})_{k_1,\ldots,k_n}$$

with $k = k_1 + k_2 r_1 + k_3 r_1 r_2 + \cdots + k_n \prod_{i=1}^{n-1} r_i$, $0 \leq k_i < r_i$ and

$$(A \triangleright \mathcal{B})_{k_1,\ldots,k_n} = \sum_{l_1=0}^{s_1-1} \mathbb{A}^1_{k_1,l_1} \sum_{l_2=0}^{s_2-1} \mathbb{A}^2_{k_2,l_2} \cdots \sum_{l_n=0}^{s_n-1} \mathbb{A}^n_{k_n,l_n} \mathcal{B}_{l_1,\ldots,l_n}. \tag{6}$$

We also know that:

$$((\mathbb{A}^n \otimes \cdots \otimes \mathbb{A}^1) vec(\mathcal{B}))_k = \sum_{i=0}^{s_1 s_2 \cdots s_n - 1} (\mathbb{A}^n \otimes \cdots \otimes \mathbb{A}^1)_{k,i} vec(\mathcal{B})_i \tag{7}$$

with $vec(\mathcal{B})_i = \mathcal{B}_{i_1,\ldots,i_n}$ and $i = i_1 + i_2 s_1 + i_3 s_1 s_2 + \cdots + i_n \prod_{i=1}^{n-1} s_i$.

Note that

$$(\mathbb{A}^n \otimes \cdots \otimes \mathbb{A}^1)_{k,i} = \prod_{u=1}^{n} \mathbb{A}^u_{k_u,i_u} \tag{8}$$

with $k = k_1 + k_2 r_1 + k_3 r_1 r_2 + \cdots + k_n \prod_{i=1}^{n-1} r_i$ and $i = i_1 + i_2 s_1 + i_3 s_1 s_2 + \cdots + i_n \prod_{i=1}^{n-1} s_i$.

Substitute (8) into (7) and split the summation to get

$$((\mathbb{A}^n \otimes \cdots \otimes \mathbb{A}^1) vec(\mathcal{B}))_k = \sum_{i_1=0}^{s_1-1} \cdots \sum_{i_n=0}^{s_n-1} \prod_{u=1}^{n} \mathbb{A}^u_{k_u,i_u} \mathcal{B}_{i_1,\ldots,i_n}. \tag{9}$$

We conclude that (6) equals (9).

□

Let us note in particular that a combination of (4) and (5) but applied to matrices, leads to the famous *mixed product rule* for Kronecker products.

2.1.1 Blocks built up with moments and central moments

As mentioned in the introduction, our next step is to define blocks using ingredients from the marginal characteristics; in this fashion hypotheses of interest can be formulated immediately in terms of the elements of such a block. The association of blocks to a sequence of stochastic variables is rather general.

Definition 2.6 *With a sequence of random variables* $Z_{i,k}$, $1 \leq k \leq n$ *and* $0 \leq i \leq r_k - 1$, *we associate a block as follows:*

$$\mathcal{B}_{i_1,\cdots,i_n}(\mathbf{Z}) = E(Z_{i_1,1} \cdots Z_{i_n,n}). \tag{10}$$

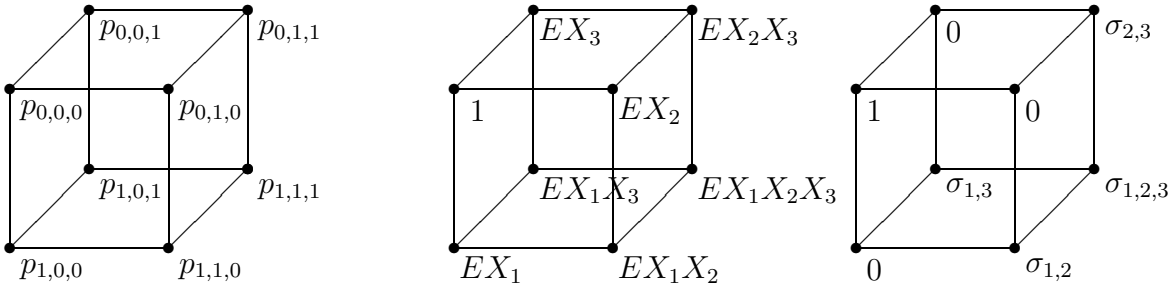
To be more specific, suppose now that we have a multivariate discrete variable $\mathbf{X} \in MD(r_1, \cdots, r_n)$. As shown in the next example, we can construct from $\mathbf{X}$ sequences $Z_{i,k}$ and subsequent blocks in a variety of ways.

Example

- Choose in (10), $Z_{i,k} = I(X_k = i)$, we obtain a block built up with the cell probabilities since $E(I(X_1 = i_1) \cdots I(X_n = i_n)) = p_{i_1,\cdots,i_n}$. We denote this block by $\mathcal{B}^p(\mathbf{X})$;

- Choose in (10), $Z_{i,k} = X_k^i$, we get the block built up with a given set of moments. This block is denoted by $\mathcal{B}^\mu(\mathbf{X})$;

- Choose in (10), $Z_{0,k} = 1$ and $Z_{i,k} = X_k^i - EX_k^i$, we get a block built up with a given set of central moments. We denote this block by $\mathcal{B}^\sigma(\mathbf{X})$.



(a):$\mathcal{B}^p(\mathbf{X})$ *(b)*: $\mathcal{B}^\mu(\mathbf{X})$ *(c)*:$\mathcal{B}^\sigma(\mathbf{X})$

with $\sigma_{i,j} = Cov(X_i, X_j)$ *and* $\sigma_{1,2,3} = E(X_1 - EX_1)(X_2 - EX_2)(X_3 - EX_3)$.

Figure 2

It is important to emphasize that this list is in no way exhaustive. Other choices using for example factorial moments are possible (see [16]).

2.1.2 Transformation formulas

We now apply the operators of the previous section to obtain formulas that express the original cell probabilities in terms of the new representation and vice versa. In other words we reparametrize the cell probabilities.

Property 2.2 *Suppose that* $\mathbf{X} \in MD(r_1, \cdots, r_n)$. *The operator to transform* $\mathcal{B}^p(\mathbf{X})$ *into* $\mathcal{B}^\mu(\mathbf{X})$ *or* $\mathcal{B}^\sigma(\mathbf{X})$ *and vice versa is the flat-product where the flats are defined as in the following scheme:*

transformation	*flat*
(a) $\mathcal{B}^p(\mathbf{X}) \rightarrow \mathcal{B}^\mu(\mathbf{X})$	$(\mathbb{A}^1\vert\cdots\vert\mathbb{A}^n)$
	with $\mathbb{A}^k \in M(r_k, r_k) : \mathbb{A}^k_{i,j} = j^i$
(b) $\mathcal{B}^p(\mathbf{X}) \rightarrow \mathcal{B}^\sigma(\mathbf{X})$	$(\mathbb{B}^1\vert\cdots\vert\mathbb{B}^n)$
	with $\mathbb{B}^k \in M(r_k, r_k) : \mathbb{B}^k_{i,j} = \begin{cases} 1 & i = 0 \\ j^i - EX^i_k & i \neq 0 \end{cases}$
(c) $\mathcal{B}^\mu(\mathbf{X}) \rightarrow \mathcal{B}^p(\mathbf{X})$	$(\mathbb{C}^1\vert\cdots\vert\mathbb{C}^n)$
	with $\mathbb{C}^k \in M(r_k, r_k) : \mathbb{C}^k = \begin{bmatrix} 1 & -e^T\mathbb{Z}^{r_k-1} \\ 0 & \mathbb{Z}^{r_k-1} \end{bmatrix}$
	and $e \in M(r_k - 1, 1) : e = [1, \cdots, 1]^T$,
	$\mathbb{Z}^t \in M(t,t) : \mathbb{Z}^t_{i,j} = \frac{(-1)^{i+t}}{(i+1)!(t-i-1)!} \sum_{k=1}^{j+1} (i+1)^{k-j-2} \begin{bmatrix} t+1 \\ k \end{bmatrix}$
(d) $\mathcal{B}^\sigma(\mathbf{X}) \rightarrow \mathcal{B}^p(\mathbf{X})$	$(\mathbb{D}^1\vert\cdots\vert\mathbb{D}^n)$
	with $\mathbb{D}^k \in M(r_k, r_k) : \mathbb{D}^k_{i,j} = \begin{cases} \sum_{s=0}^{r_k-1} EX^s_k \mathbb{C}^k_{i,s} & j = 0 \\ \mathbb{C}^k_{i,j} & j \neq 0 \end{cases}$

where $\begin{bmatrix} t+1 \\ k \end{bmatrix}$ *represents a Stirling number of the first kind defined by the relation:*

$$x(x-1)\cdots(x-t) = \sum_{k=1}^{t+1} \begin{bmatrix} t+1 \\ k \end{bmatrix} x^k. \tag{11}$$

Proof:

(a): Apply the definition of a flat-product and $EX_1^{i_1} \cdots X_n^{i_n}$.

(b): This is similar to the above case.

(c): Because of (5), it is sufficient to take $\mathbb{C}^k$ as the inverse matrix of $\mathbb{A}^k$. Partition $\mathbb{A}^k$ as follows:

$$\begin{bmatrix} 1 & e^T \\ 0 & \mathbb{V}^{r_k-1} \end{bmatrix}$$

with $\mathbb{V}^t \in M(t,t)$ the *Vandermonde Matrix*: $\mathbb{V}^t_{i,j} = (j+1)^{i+1}$.

Consequently, it suffices to show that $\mathbb{Z}^t$ is the inverse of $\mathbb{V}^t$:

$$\forall i,j : \sum_{k=0}^{t-1} \mathbb{Z}^t_{i,k} \mathbb{V}^t_{k,j} = \delta_{i,j}.$$

To do that, define the functions $f_i, 0 \leq i \leq t-1$:

$$f_i(x) = \frac{(-1)^{i+1+t}}{(i+1)!(t-i-1)!} \prod_{0 \leq m \neq i+1 \leq t} (x-m). \tag{12}$$

Since $f_i(j+1) = \delta_{i,j}$, the elements of $\mathbb{Z}^t$ satisfy the following relation:

$$\sum_{k=0}^{t-1} \mathbb{Z}^t_{i,k} x^{k+1} = \frac{(-1)^{i+1+t}}{(i+1)!(t-i-1)!} \prod_{0 \leq m \neq i+1 \leq t} (x-m).$$

Multiplying both sides by $(x-(i+1))$, using the expansion (11) for the right hand side and equating the coefficients of equal powers of x on the left and right hand side, we get the following recursion:

$$(i+1)\mathbb{Z}^t_{i,m} - \mathbb{Z}^t_{i,m-1} = \frac{(-1)^{i+t}}{(i+1)!(t-i-1)!} \begin{bmatrix} t+1 \\ m+1 \end{bmatrix},$$

with $\mathbb{Z}^t_{i,-1} = 0$.

Multiplying both sides by $(i+1)^{m-1}$ and defining $w_m := (i+1)^m \mathbb{Z}^t_{i,m}$, we get:

$$w_m - w_{m-1} = \frac{(-1)^{i+t}(i+1)^{m-1}}{(i+1)!(t-i-1)!} \left[\begin{array}{c} t+1 \\ \\ m+1 \end{array} \right].$$

This can be solved by adding the terms in the right hand side.

(d): From the previous part we know:

$$\begin{pmatrix} I(X_k = 0) \\ I(X_k = 1) \\ \vdots \\ I(X_k = r_k - 1) \end{pmatrix} = \mathbb{C}_k \begin{pmatrix} 1 \\ X_k \\ X_k^2 \\ \vdots \\ X_k^{r_k - 1} \end{pmatrix} =$$

$$(\mathbb{C}_k \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ EX_k & 1 & 0 & \cdots & 0 \\ EX_k^2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ EX_k^{r_k-1} & 0 & 0 & \cdots & 1 \end{pmatrix}) \begin{pmatrix} 1 \\ X_k - EX_k \\ X_k^2 - EX_k^2 \\ \vdots \\ X_k^{r_k-1} - EX_k^{r_k-1} \end{pmatrix}.$$

Now apply the definition of a flat-product.

□

For the case of binary variables, the above relations simplify as follows:

flat for transf. *from* $\downarrow$ *to* $\rightarrow$	$\mathcal{B}^p(\mathbf{X})$
$\mathcal{B}^p(\mathbf{X})$	$(\mathbb{I}\vert\cdots\vert\mathbb{I})$
$\mathcal{B}^\mu(\mathbf{X})$	$(\begin{pmatrix} 1 & -1 \\ & \\ 0 & 1 \end{pmatrix}\vert\cdots\vert\begin{pmatrix} 1 & -1 \\ & \\ 0 & 1 \end{pmatrix})$
$\mathcal{B}^\sigma(\mathbf{X})$	$(\begin{pmatrix} 1-EX_1 & -1 \\ & \\ EX_1 & 1 \end{pmatrix}\vert\cdots\vert\begin{pmatrix} 1-EX_n & -1 \\ & \\ EX_n & 1 \end{pmatrix})$

flat for transf. *from* $\downarrow$ *to* $\rightarrow$	$\mathcal{B}^\mu(\mathbf{X})$
$\mathcal{B}^p(\mathbf{X})$	$(\begin{pmatrix} 1 & 1 \\ & \\ 0 & 1 \end{pmatrix}\vert\cdots\vert\begin{pmatrix} 1 & 1 \\ & \\ 0 & 1 \end{pmatrix})$
$\mathcal{B}^\mu(\mathbf{X})$	$(\mathbb{I}\vert\cdots\vert\mathbb{I})$
$\mathcal{B}^\sigma(\mathbf{X})$	$(\begin{pmatrix} 1 & 0 \\ & \\ EX_1 & 1 \end{pmatrix}\vert\cdots\vert\begin{pmatrix} 1 & 0 \\ & \\ EX_n & 1 \end{pmatrix})$

flat for transf. *from* $\downarrow$ *to* $\rightarrow$	$\mathcal{B}^\sigma(\mathbf{X})$
$\mathcal{B}^p(\mathbf{X})$	$(\begin{pmatrix} 1 & 1 \\ & \\ -EX_1 & 1-EX_1 \end{pmatrix}\vert\cdots\vert\begin{pmatrix} 1 & 1 \\ & \\ -EX_n & 1-EX_n \end{pmatrix})$
$\mathcal{B}^\mu(\mathbf{X})$	$(\begin{pmatrix} 1 & 0 \\ & \\ -EX_1 & 1 \end{pmatrix}\vert\cdots\vert\begin{pmatrix} 1 & 0 \\ & \\ -EX_n & 1 \end{pmatrix})$
$\mathcal{B}^\sigma(\mathbf{X})$	$(\mathbb{I}\vert\cdots\vert\mathbb{I})$

Table I

2.2 Block models in practice

2.2.1 Example

Consider the following data taken from Grizzle (1969). Each subject is classified according to its reaction (favorable "0" or not favorable "1") after treatment by three kinds of drugs resp. X_1, X_2, X_3.

	$X_3 = 0$		$X_3 = 1$	
	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	*6*	*2*	*2*	*6*
$X_1 = 1$	*16*	*4*	*4*	*6*

Table II

Among the points of interest we mention differences (or similarities) in efficiency between the treatments (marginal homogeneity) and their interaction. In Table *III* we have collected a summary of some relevant groups of hypotheses together with one possible explicitation of each hypothesis and its formulation in terms of the moment parametrization (i.e. in terms of $EX_1, EX_2, EX_3, \sigma_{1,2}, \sigma_{1,3}, \sigma_{2,3}, \sigma_{1,2,3}$).

type of hypothesis	*example*	*formulation with moments*
marg. homog.	$X_1 \stackrel{\mathcal{D}}{=} X_2 \stackrel{\mathcal{D}}{=} X_3$	$EX_1 = EX_2 = EX_3$
independency	$X_1 \perp X_2 \perp X_3$	$\sigma_{1,2} = \sigma_{1,3} = \sigma_{2,3} = \sigma_{1,2,3} = 0$
symmetry	$p_{i_1,i_2,i_3} = p_{\Pi(i_1,i_2,i_3)}$ *for all perm.* Π	$EX_1 = EX_2 = EX_3$, $EX_1X_2 = EX_1X_3 = EX_2X_3$
2-nd order marg. homog.	$(X_1, X_2) \stackrel{\mathcal{D}}{=} (X_1, X_3)$	$EX_2 = EX_3$, $\sigma_{1,2} = \sigma_{1,3}$
2-nd order symmetry	$(X_1, X_2) \stackrel{\mathcal{D}}{=} (X_2, X_1)$	$EX_1 = EX_2$
no pairwise interaction	$X_1 \perp X_2$, $X_1 \perp X_3$, $X_2 \perp X_3$	$\sigma_{1,2} = \sigma_{1,3} = \sigma_{2,3} = 0$

Table III

The transformation formulas of Section 2.1.2 allow us to specify the cell probabilities in terms of moment parameters; next, we maximize $\sum n_{i_1,i_2,i_3} \log[p_{i_1,i_2,i_3}(EX_1, EX_2, \cdots)]$ over the free parameters of $(EX_1, EX_2, \cdots)$ under H_0 and a classical *goodness-of-fit* statistic can be used. The existence and uniqueness of the *maximum likelihood estimators* is studied in [1].

The estimated moment parameters (together with 95% confidence intervals obtained by applying the traditional $\delta-$method) are: $EX_1 = 0.652(\pm 0.13); EX_2 = 0.391(\pm 0.14); EX_3 = 0.391(\pm 0.14); \sigma_{1,2} = -0.037(\pm 0.06); \sigma_{1,3} = -0.037(\pm 0.06); \sigma_{2,3} = 0.107(\pm 0.06); \sigma_{1,2,3} = -0.0101(\pm 0.04)$.

There is evidence that the efficiency (i.e. the mean) of X_1 is different from that of X_2 and X_3 (the hypothesis of equality has a p-value of 0.04) . Also the interaction between (X_2, X_3) seems to be different from that between (X_1, X_3) and (X_1, X_2) (p-value of 0.01).

In Table *IV* the results of the most important acceptable hypotheses are summarized.

hypothesis	*p-value*
$EX_2 = EX_3$	*1.0*
$\sigma_{1,2} = \sigma_{1,3} = \sigma_{1,2,3} = 0$	*0.62*
$EX_2 = EX_3, \sigma_{1,2} = \sigma_{1,3} = \sigma_{1,2,3} = 0$	*0.77*

Table IV

Finally note that, opposite to a *log-linear* model, the hypothesis of quasi-symmetry can not be tested directly. One has to resort to the decomposition [2]:

$$\text{quasi-symmetry} \cap \text{marginal homogeneity} \leftrightarrow \text{symmetry}. \tag{13}$$

For the above dataset the hypothesis of marginal homogeneity has a p-value of 0.04 such that the use of (13) for the hypothesis of quasi-symmetry is justified [2] (but the hypothesis itself will be rejected with a p-value less then 0.01).

This is to be compared with a classical *log-linear* model where a hypothesis of marginal homogeneity can not be tested directly but the one of quasi-symmetry can. Once more, (13) can be used but now under the restriction that quasi-symmetry is not too implausible. Unfortunately this is hardly acceptable and a direct test, as available with a block model, seems preferable as has already been mentioned in [2]. This observation illustrates the complementarity of block and *log-linear* models.

2.2.2 Practical issues

Scale of measurement

In the case of non-binary variables, the chosen scale of measurement will have an influence on the parameters. Nevertheless, for an important class of hypotheses this does not matter as is shown in the next example taken from Hagenaars (1990). It concerns a study about changes in political preferences during the post-election period February 1977 and March 1977 in the Netherlands. People were asked which party (X_1) and which prime-minister (X_3) they preferred in February and for which party (X_2) and prime- minister (X_4) they would vote if one organizes new elections at that moment (March 1977). The data are given in Table *V*.

Questions of interest are for example: " Did the party or prime-minister preference change between February and March?" or "Is the preference for a prime-minister different from the difference for a party?". Such questions are naturally interpreted in terms of equalities of the underlying random variables such as $X_1 \stackrel{\mathcal{D}}{=} X_2$ or equivalently $EX_1 = EX_2$ & $EX_1^2 = EX_1^2$. As no absolute values are involved, the scale of measurement does not matter (supposing that we use the same scale for X_1 and X_2).
Similarly, we can formulate conditions for symmetry such as $(X_1, X_2) \stackrel{\mathcal{D}}{=} (X_2, X_1)$ which is equivalent to $EX_1 = EX_2$ & $EX_1^2 = EX_2^2$ & $EX_1X_2^2 = EX_1^2X_2$. Finally we can test for constraints of the type $\sum_i a_i P(X_i = k) = c_k$. E.g. "Is there a net change between the turnover in party preference and prime minister preference?": $\forall k : P(X_1 = k) - P(X_2 = k) = P(X_3 = k) - P(X_4 = k)$, which is equivalent to $EX_1 - EX_2 = EX_3 - EX_4$ and $EX_1^2 - EX_2^2 = EX_3^2 - EX_4^2$.

		$X_3=0$			$X_3=1$			$X_3=2$		
	$X_4=$	0	1	2	0	1	2	0	1	2
	$X_2=0$	*84*	*9*	*23*	*6*	*13*	*7*	*24*	*8*	*68*
$X_1=0$	$X_2=1$	*0*	*1*	*0*	*0*	*8*	*1*	*2*	*2*	*3*
	$X_2=2$	*3*	*1*	*2*	*0*	*2*	*3*	*2*	*3*	*9*
	$X_2=0$	*1*	*1*	*0*	*1*	*2*	*2*	*1*	*0*	*1*
$X_1=1$	$X_2=1$	*2*	*4*	*0*	*1*	*293*	*6*	*1*	*22*	*21*
	$X_2=2$	*1*	*0*	*0*	*1*	*8*	*7*	*0*	*0*	*9*
	$X_2=0$	*6*	*1*	*1*	*4*	*5*	*0*	*9*	*1*	*16*
$X_1=2$	$X_2=1$	*0*	*1*	*1*	*0*	*31*	*0*	*2*	*9*	*7*
	$X_2=2$	*14*	*1*	*15*	*3*	*48*	*23*	*12*	*21*	*200*

The coding used for X_1 *and* X_2 *is:* 0 *represents "Christian Democratic";* 1 *represents "Left Wing" and* 2 *represents "Other" . The coding for* X_3 *and* X_4 *is:* 0 *represents "Van Agt" (Christian Democrate),* 1 *represents "Den Uyl" (Left Wing) and* 2 *represents "Other" .*

Table V

By way of illustration, Table *VI* collects a few results on a set of such hypotheses. Mainly because of a significant difference between X_1 and X_2, only the second and last hypothesis is acceptable (with a p-value of resp. 0.14 and 0.60).

hypothesis	*formulation*
Has the party preference changed?	$X_1 \stackrel{\mathcal{D}}{=} X_2$
Has the prime-minister preference changed?	$X_3 \stackrel{\mathcal{D}}{=} X_4$
Has the preference changed in time?	$(X_1, X_3) \stackrel{\mathcal{D}}{=} (X_2, X_4)$
Is there symmetry in party and prime-minister preference at each moment?	$(X_1, X_2) \stackrel{\mathcal{D}}{=} (X_2, X_1)$ & $(X_3, X_4) \stackrel{\mathcal{D}}{=} (X_4, X_3)$
Is the prime-minister preference equal to the party preference?	$(X_1, X_2) \stackrel{\mathcal{D}}{=} (X_3, X_4)$
Is there a net change between the turnover in party preference and prime-minister preference?	$EX_1 - EX_2 = EX_3 - EX_4$ & $EX_1^2 - EX_2^2 = EX_3^2 - EX_4^2$.

Table VI

Log-linear versus block models

Finally we sketch a number of technical differences between classical *log-linear* models and block models.

1. Contrary to the situation with *log-linear* models, moment parameters in block models are also defined in case of structural zeros. As shown in Property 2.2, this difference is caused by the fact that we do not calculate ratios of probabilities but only linear combinations of them.

2. The calculation of *maximum likelihood estimators* is much harder with moment parameters in a block model than in a *log-linear* model. In the latter, one parame-

ter is a normalization constant but there are no further restrictions on the domain of the remaining parameters. This is not the case with moment parametrization where for each parameter the domain is determined by a set of inequality constraints. For example, in the binary case, one always has $EX_1X_2 \leq EX_1$. However, using gradient search it is not difficult to include those restrictions on the domain of the parameter space.

Hence, the only remaining problem (and as it turned out, only relevant for very large datasets) is finding acceptable starting values. We solved this problem by first expanding the likelihood including the constraints by means of the *Lagrange*-method. Before returning to the original likelihood, we applied a gradient search until an acceptable solution was found.

3. The parameters in block models are sums of cell frequencies. As noted in [13], sparse tables will often lead to relatively large values for such parameters in comparison with the observed cell frequencies. Therefore, it might be of interest to carry out a direct estimation method in terms of such parameters rather than to rely on the classical maximum likelihood estimate of the cell probabilities.

3 Transformations of discrete variables

In the second part of the paper we return to the algebraic framework. We will show how it implies a complete characterization of all meaningful transformations on discrete variables. Before giving our main result, we first formulate the problem in a more general context and introduce some additional operators on blocks.

3.1 Formulation of the problem

Consider the following example from Bloomfield [3]. Suppose couples are asked about their favorite party. The data can be described by means of the variables (X_1, X_2) where X_1 (X_2) denotes the party the man (woman) would vote on. In case of binary variables, it is possible that an easier interaction structure is obtained if we look at the pair of variables (Y_1, Y_2), where Y_1 denotes whether they vote on the same party or not, and Y_2 denotes the man's preference. In [3], only linear transformations (modulo 2) on (X_1, X_2) have been considered as they can be easily formulated in terms of the parameters of a *log-linear* model.

We intend to derive all *meaningful transformations* of $\mathbf{X}$ and show how the cell probabilities are related to the blocks $\mathcal{B}^p(\mathbf{Y}), \mathcal{B}^\mu(\mathbf{Y})$ and $\mathcal{B}^\sigma(\mathbf{Y})$. Another motivation to look for such transformations is that they lead to a statistically and mathematically correct *dimension reduction technique* that takes into account the nominal nature of the data. The resulting procedure is in contrast to the classical approach where the categorical variables are treated as metric quantities.

3.2 Rao product

In 1968, Khatri and Rao defined the following operator:

Definition 3.1 *[12] If* $\mathbb{A} \in M(r_1, s)$ *and* $\mathbb{B} \in M(r_2, s)$*, define* $\mathbb{A} \odot \mathbb{B} \in M(r_1 r_2, s)$ *by*

$$\forall i : (\mathbb{A} \odot \mathbb{B})_{|i} = \mathbb{A}_{|i} \otimes \mathbb{B}_{|i}$$

where $'|i'$ *denotes the* i*-th column of a matrix.*

Example

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 1 \end{pmatrix} \odot \begin{pmatrix} 1 & 0 & 2 \\ 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \otimes \begin{pmatrix} 1 \\ 1 \end{pmatrix} & \begin{pmatrix} 2 \\ 1 \end{pmatrix} \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix} & \begin{pmatrix} 3 \\ 1 \end{pmatrix} \otimes \begin{pmatrix} 2 \\ 1 \end{pmatrix} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 6 \\ 1 & 2 & 3 \\ 0 & 0 & 2 \\ 0 & 1 & 1 \end{pmatrix}.$$

Property 3.1 *[12] If* $\mathbb{T}_1 \in M(p, q), \mathbb{T}_2 \in M(n, m), \mathbb{A} \in M(q, s)$ *and* $\mathbb{B} \in M(m, s)$*, then one has the following mixed product rule:*

$$(\mathbb{T}_1 \otimes \mathbb{T}_2)(\mathbb{A} \odot \mathbb{B}) = (\mathbb{T}_1 \mathbb{A}) \odot (\mathbb{T}_2 \mathbb{B}).$$

We extend the above concept to blocks and derive its mixed product rule.

Definition 3.2 *Suppose* $\mathcal{B} \in M(s_1, \cdots, s_n)$, $s = \prod_i s_i$ *and* $A = (\mathbb{A}^1 | \cdots | \mathbb{A}^n)$ *with* $\mathbb{A}^i \in M(r_i, s)$. *Define the Rao product* $A \triangle \mathcal{B} \in M(r_1, \cdots, r_m)$ *as:*

$$A \triangle \mathcal{B} = \mathcal{C} \Leftrightarrow vec(\mathcal{C}) = (\mathbb{A}^n \odot \cdots \odot \mathbb{A}^1) vec(\mathcal{B}).$$

Property 3.2 *If* $\mathcal{C} \in M(t_1, \cdots, t_n)$, $t = \prod_i t_i$, $A = (\mathbb{A}^1 | \cdots | \mathbb{A}^n)$ *and* $B = (\mathbb{B}^1 | \cdots | \mathbb{B}^n)$ *with* $\mathbb{A}^i \in M(r_i, s_i)$ *and* $\mathbb{B}^i \in M(s_i, t)$ *then*

$$A \triangleright (B \triangle \mathcal{C}) = (A.B) \triangle \mathcal{C}. \tag{14}$$

Proof:

Vectorization of the left hand side of (14), gives:

$$((\mathbb{A}^n \otimes \cdots \otimes \mathbb{A}^1)(\mathbb{B}^n \odot \cdots \odot \mathbb{B}^1)) vec(\mathcal{C}).$$

Because of Property 3.1, one obtains

$$(\mathbb{A}^n \mathbb{B}^n \odot \cdots \odot \mathbb{A}^1 \mathbb{B}^1) vec(\mathcal{C})$$

which is exactly the right hand side of (14) after vectorization.

□

3.3 Representation theorem

For the sake of simplicity we restrict ourselves to the multivariate Bernoulli case where all $r_i = 2$ for all i. The more elaborate general case is considered in [19]. In the sequel we also assume that $\mathbf{X}$ has n components.

Definition 3.3 *If* $\mathbf{X}$ *is a multivariate Bernoulli variable, define the Kronecker vector* $K(\mathbf{X})$ *as:*

$$K(\mathbf{X}) = \begin{pmatrix} \bar{X}_n \\ X_n \end{pmatrix} \otimes \begin{pmatrix} \bar{X}_{n-1} \\ X_{n-1} \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} \bar{X}_1 \\ X_1 \end{pmatrix},$$

with $\bar{X}_i = 1 - X_i$.

Using the definition of the Kronecker product and the vectorization operator *vec*, one easily shows that

$$vec(\mathcal{B}^p(\mathbf{X})) = EK(\mathbf{X}).$$

To simplify the formulation of the next theorem we use the following abbreviations

$$\mathbf{X}^* = \begin{pmatrix} 1 & X_1 & X_2 & \cdots & X_n \end{pmatrix}^T$$

and $\mathbb{H}_m^i \in M(2, m+1)$ defined by

$$\mathbb{H}_m^i = \begin{pmatrix} 1 & 0 & \cdots & 0 & -1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}$$

where -1 is found in the $i+1$-th column.

We now formulate our representation theorem of transformations $f()$ of a multivariate Bernoulli variable, $\mathbf{X}$, into a space of multivariate Bernoulli variables where $f()$ is defined as a (measurable) function that maps $(x_1, \cdots, x_n)$ into $\{0,1\}^m$. The proof is deferred to the last section of the paper.

Theorem 3.1 *Suppose* $\mathbf{X} = (X_1 \cdots, X_n)$ *has a multivariate Bernoulli distribution as defined by* $\mathcal{B}^p(\mathbf{X})$, $\mathbf{Y} = (Y_1, \cdots, Y_m)$ *is a transformation of* $\mathbf{X}$,

iff

there exists a $(m+1) \times 2^n$ *matrix* $\mathbb{T}$*:* $\forall j : t_{0,j} = 1$ *and* $\forall i \neq 0, \forall j : t_{i,j} \in \{0, 1\}$ *such that*

$$\mathbf{Y}^* = \mathbb{T}\, K(\mathbf{X}) \tag{15}$$

iff

there exists a scrambler-flat $T = (\mathbb{T}^1 | \cdots | \mathbb{T}^m)$ *with* $\mathbb{T}^i \in M(2, 2^n)$ *such that*

$$\mathcal{B}^p(\mathbf{Y}) = T \triangle \mathcal{B}^p(\mathbf{X}) \tag{16}$$

iff

there exists a $2^m \times 2^n$ *scrambler* $\mathbb{A}$ *so that*

$$EK(\mathbf{Y}) = \mathbb{A} EK(\mathbf{X}). \tag{17}$$

Moreover we have the following relationship between $\mathbb{T}$, T *and* $\mathbb{A}$ *:*

$$\mathbb{T}^i = \mathbb{H}_m^i \mathbb{T} \tag{18}$$

and

$$\mathbb{A} = (\mathbb{H}_m^m \mathbb{T}) \circledcirc \cdots \circledcirc (\mathbb{H}_m^1 \mathbb{T}), \tag{19}$$

It is natural to call $\mathbb{T}$ the *transformation matrix.* Note that Eq. (15) and (16) specify how a transformation can be formulated in terms of the parameters of the block model and vice versa.

3.3.1 Applications of the representation theorem

In this subsection we give a few applications of the above representation theorem.

3.3.1.1 The dimension of a multivariate discrete distribution

Theorem 3.1 allows us to introduce an equivalence relationship on multivariate Bernoulli distributions. This will then naturally lead to a concept of *dimension*, somewhat akin to that of the multivariate normal distribution.

Definition 3.4 *Suppose* $\mathbf{X}$ *and* $\mathbf{Y}$ *are multivariate Bernoulli variables , we call* $\mathbf{X}$ *and* $\mathbf{Y}$ *Bernoulli equivalent, and write* $\mathbf{X} \Leftrightarrow \mathbf{Y}$*, iff there exist transformations* $f(), g()$ *such that* $f(\mathbf{X}) = \mathbf{Y}$*,* $g(\mathbf{Y}) = \mathbf{X}$ *with probability one.*

Definition 3.5 *If* $\mathbf{X}$ *is a multivariate Bernoulli variable then it has dimension* k *iff* k *is the minimal number of components necessary to construct a multivariate Bernoulli variable* $\mathbf{Y}$ *such that* $\mathbf{X} \Leftrightarrow \mathbf{Y}$*.*

Property 3.3 *Suppose* $\mathbf{X}$ *is a multivariate Bernoulli variable with* n *components.* $\mathbf{X}$ *has at most dimension* $n-1$ *iff there are at least* 2^{n-1} *zeros in the block* $\mathcal{B}^p(\mathbf{X})$*.*

Proof:

$\Downarrow$ We show how to construct a transformation of $\mathbf{X}$ to a multivariate Bernoulli variable $\mathbf{Y}$ with $n-1$ components. Because of Theorem 3.1, it suffices to construct a scrambler $\mathbb{A} \in M(2^{n-1}, 2^n)$ for which $EK(\mathbf{Y}) = \mathbb{A}\, EK(\mathbf{X})$ and a scrambler $\mathbb{B} \in M(2^n, 2^{n-1})$ for which $EK(\mathbf{X}) = \mathbb{B}\, EK(\mathbf{Y})$.

Take the identity matrix of dimension $2^n \times 2^n$ as a starting point and remove the 2^{n-1} rows i for which $EK(\mathbf{X})_i = 0$. Replace in some of the remaining rows a 0 by a 1 so that in every column there is exactly one 1. This new matrix is a scrambler with 2^{n-1} rows and 2^n columns.

In order to construct $\mathbb{B}$ we consider the identity matrix of dimension $2^{n-1} \times 2^{n-1}$ and we insert rows with zeros at those places i where $EK(\mathbf{X})_i = 0$.

$\Uparrow$ We know there exists a matrix $\mathbb{A}$ of dimension $2^n \times 2^{n-1}$ for which $EK(\mathbf{X}) = \mathbb{A}\, EK(\mathbf{Y})$. In each column of $\mathbb{A}$ we find exactly one 1. Hence there are 2^{n-1} rows with only zeros in $\mathbb{A}$, consequently $EK(\mathbf{X})$ has at least 2^{n-1} zeros.

□

Example

Let us start from $\mathbf{X}$ with a distribution determined by the cell probabilities $\mathcal{B}^p(\mathbf{X})$:

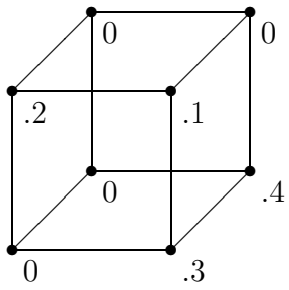


Figure 3

One has $EK(\mathbf{X}) = \begin{pmatrix} .2 & 0 & .1 & .3 & 0 & 0 & 0 & .4 \end{pmatrix}^T$.

The above property guarantees the existence of $\mathbf{Y} = (Y_1, Y_2)$ such that $\mathbf{X} \Leftrightarrow \mathbf{Y}$. If we

use the construction as described in Property 3.3, we obtain for example

$$\mathbb{A} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

Given $\mathbb{A}$, the matrices $\mathbb{T}^i = \mathbb{H}_8^i\mathbb{T}$ can be calculated by means of (19). We find:

$$\mathbb{T}^1 = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

$$\mathbb{T}^2 = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

The flat $T = (\mathbb{T}^1|\mathbb{T}^2)$ satisfies $\mathcal{B}^p(\mathbf{Y}) = T\triangle\mathcal{B}^p(\mathbf{X})$. Since $\mathbb{T}^i = \mathbb{H}_8^i\mathbb{T}$, we have the explicit expression for the transformation matrix

$$\mathbb{T} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

We can finally use $\mathbb{T}$ to write down explicitly a link between $\mathbf{Y}$ and $\mathbf{X}$. By means of (15) we find

$$\begin{aligned} Y_1 &= \bar{X}_1X_2\bar{X}_3 + X_1\bar{X}_2X_3 + \bar{X}_1X_2X_3 + X_1X_2X_3 \qquad (20) \\ &= \bar{X}_1X_2 + X_1X_3 \\ Y_2 &= X_1X_2\bar{X}_3 + \bar{X}_1X_2X_3 + X_1X_2X_3 \\ &= X_1X_2 + \bar{X}_1X_2X_3. \end{aligned}$$

Of course, this choice of $\mathbb{A}$ is not the only possibility. In [19] we developed a technique, based on Karnaugh-Veitch diagrams (see e.g. [10]) by which we can derive all possible such transformations.

3.3.1.2 A dependency measure

As another application of the above results, we construct a new *association measure* for categorical variables.

Definition 3.6 *Suppose* $\mathbf{X}$ *is a multivariate Bernoulli variable with* n *components. Define*

$$S = \sum_{i=1}^{2^{n-1}} p_{(i)}$$

where $p_{(i)}$ *denotes the* i*-th smallest probability among the cell probabilities.*

If S turns out to be zero, it means that $\mathbf{X}$ can be transformed to a lower dimension. This in particular implies that there is a very strong relationship between the components of $\mathbf{X}$. The case where S reaches its maximal value 0.5, corresponds to the situation where all cell probabilities are equal to each other. This means i.a. that the entropy is maximal, implying in turn a very weak relationship between the components;knowledge of one component of $\mathbf{X}$ does not tell us anything about the other components.

Let us compare S with some of the traditional association measures. The quantity S expresses the dependency while classical measures such as *Goodman-Kruskal*'s λ or *Pearson*'s ϕ^2 [17] express the association strength with respect to the case of independent variables. Indeed, the minimal value of the latter is obtained when the variables

are independent. Moreover λ and ϕ^2 measure the association between two specific components while S is invariant under invertible transformations of the components (with $m = n$) and consequently, it primarily measures the dependency in the data.

Finally remark that S plays the same role in categorical data analysis as the variance in ordinary *Principal Component Analysis* in that it expresses the information loss caused by a reduction of the dimensionality of the data.

Example

Consider the following data from [11]. The data refer to 94 graves of an old Indian cemetery. The variables X_1, X_2 and X_3 indicate the absence or presence of Red Ochre, Pottery and Hoe near a grave.

				X_3 :	
				Hoe absent	*Hoe present*
			Pottery absent	*33*	*7*
	Ochre absent	X_2 :			
			Pottery present	*28*	*10*
X_1 :					
			Pottery absent	*1*	*3*
	Ochre present	X_2 :			
			Pottery present	*3*	*9*

Table VII

The hypothesis $X_1 \perp X_2 \perp X_3$ is rejected at any level (p-value < 0.001). Nevertheless one finds many transformations of the data into variables for which the independence assumption is easily accepted. Some of them are listed in Table *VIII*.

transformation	*p-value*
$Y_1 = X_1\bar{X}_3 + X_2\bar{X}_3 + X_1X_2$	
$Y_2 = X_1\bar{X}_3 + \bar{X}_2X_3$	*0.780*
$Y_3 = X_1\bar{X}_2 + X_2X_3$	
$Y_1 = X_1\bar{X}_3 + \bar{X}_1X_2$	
$Y_2 = X_1\bar{X}_3 + \bar{X}_2X_3$	*0.696*
$Y_3 = X_1\bar{X}_2 + X_2X_3$	
$Y_1 = X_1$	
$Y_2 = \bar{X}_1\bar{X}_2X_3 + X_1\bar{X}_3 + X_2\bar{X}_3$	*0.668*
$Y_3 = X_1\bar{X}_2 + \bar{X}_1X_3$	
$Y_1 = X_2\bar{X}_3 + X_1X_3$	
$Y_2 = \bar{X}_2X_3 + X_1\bar{X}_3$	*0.605*
$Y_3 = X_2X_3 + X_1\bar{X}_2$	

Table VIII

The interpretation of the first transformation is shown in Figure 4. It shows $\mathcal{B}^p(\mathbf{Y})$ in terms of the cell probabilities of $\mathbf{X}$, $p_{i,j,k}(\mathbf{X})$ (cf. Figure 2 (a)). As one can see, the transformation defines a reordering of the cell probabilities such that the hypothesis of

independent components of $\mathbf{Y}$ is acceptable.

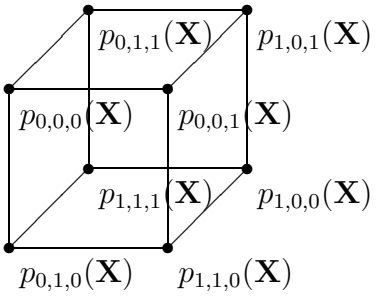


Figure 4

3.3.2 Proof of the representation theorem

We first proof two lemmas.

Lemma 3.1 *Assume that* $\mathbf{X}$ *is a multivariate Bernoulli variable. All possible transformations of* $\mathbf{X}$ *into a multivariate Bernoulli variable* $\mathbf{Y} = (Y_1, \cdots, Y_m)$*, are characterized by the* $(m+1) \times 2^n$ *matrix* $\mathbb{T} : \forall j : t_{0,j} = 1$ *and* $\forall i \neq 0, \forall j : t_{i,j} \in \{0, 1\}$ *such that:*

$$\mathbf{Y}^* = \mathbb{T}\, K(\mathbf{X}). \tag{21}$$

Proof: If f is a function from $\{0,1\}^n \to \{0,1\}$, there exist constants $a_t \in \{0,1\}$:

$$f(x_1, \cdots, x_n) = \sum_{t=0}^{2^n-1} a_t \prod_{i=1}^{n} (1 - x_i)^{1-t_i} x_i^{t_i} \tag{22}$$

with $t_i \in \{0,1\}$ defined by $t = \sum_1^n t_i 2^{i-1}$ and $a_t = f(t_1, \ldots, t_n)$. Apply (22) on each component of $\mathbf{X}$ and make use of the equality $K(\mathbf{X})_t = \prod_i (1 - X_i)^{1-t_i} X_i^{t_i}$ with $t = \sum_i t_i 2^{i-1}$. Noting that $(\mathbb{T}K(\mathbf{X}))_0 = 1$ as $\sum_t K(\mathbf{X})_t = 1$, we get (21).

□

Lemma 3.2 *If* $\mathbf{X}$ *has a multivariate Bernoulli distribution,* $\mathbf{Y} = (Y_1, \cdots, Y_m)$ *is a transformation of* $\mathbf{X}$*, again with a multivariate Bernoulli distribution*

iff

there exists a $2^m \times 2^n$ *scrambler* $\mathbb{A}$ *so that*

$$K(\mathbf{Y}) = \mathbb{A}K(\mathbf{X}) \tag{23}$$

or equivalently

$$EK(\mathbf{Y}) = \mathbb{A}EK(\mathbf{X}) \tag{24}$$

with

$$\mathbb{A} = (\mathbb{H}_m^m \mathbb{T}) \circledast \cdots \circledast (\mathbb{H}_m^1 \mathbb{T}), \tag{25}$$

where $\mathbb{T}$ *is the corresponding transformation matrix.*

Proof:

$\Downarrow$ We determine the scrambler $\mathbb{A}$ for a given transformation as follows.

Define

$$\mathbb{J} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$$

and

$$\mathbb{J}_m^i = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}$$

with $\mathbb{J}_m^i$ a $2 \times (m+1)$ matrix with a 1 in the first column, first row and in the $i+1$-th column, second row.

Since $\mathbb{J}_m^i \mathbf{Y}^* = \begin{pmatrix} 1 \\ \\ Y_i \end{pmatrix}$ and $\mathbb{J} \begin{pmatrix} 1 \\ \\ Y_i \end{pmatrix} = \begin{pmatrix} \bar{Y}_i \\ \\ Y_i \end{pmatrix}$, we get

$$K(\mathbf{Y}) = (\mathbb{J}\mathbb{J}_m^m \mathbf{Y}^*) \otimes (\mathbb{J}\mathbb{J}_m^{m-1} \mathbf{Y}^*) \otimes \cdots \otimes (\mathbb{J}\mathbb{J}_m^1 \mathbf{Y}^*) := \vec{\otimes}_{i=m}^1 (\mathbb{J}\mathbb{J}_m^i \mathbf{Y}^*). \tag{26}$$

Since $\mathbb{H}_m^i = \mathbb{J}\mathbb{J}_m^i$, Eq. (26) can be rewritten by means of Lemma 3.1, as:

$$K(\mathbf{Y}) = \vec{\otimes}_{i=m}^1 (\mathbb{H}_m^i \mathbb{T} K(\mathbf{X})) = \vec{\otimes}_{i=m}^1 (\mathbb{T}^i K(\mathbf{X})) \tag{27}$$

with $\mathbb{T}^i = \mathbb{H}_m^i \mathbb{T}$ a 2×2^n scrambler:

$$t_{0,l}^i = 1 - t_{i,l} \text{ and } t_{1,l}^i = t_{i,l} = 1 - t_{0,l}^i \tag{28}$$

and

$$K(\mathbf{Y})_j = \sum_{l_m=0}^{2^n-1} \sum_{l_{m-1}=0}^{2^n-1} \cdots \sum_{l_1=0}^{2^n-1} t_{j_m,l_m}^m \cdots t_{j_1,l_1}^1 K(\mathbf{X})_{l_1} K(\mathbf{X})_{l_2} \cdots K(\mathbf{X})_{l_m}$$

with $j = \sum_{i=1}^m j_i 2^{i-1}$.

Since $K(\mathbf{X})_{l_1} K(\mathbf{X})_{l_2} \cdots K(\mathbf{X})_{l_m} = K(\mathbf{X})_{l_1}$ if $l_1 = l_2 = ... = l_m$ and 0 otherwise, we obtain

$$K(\mathbf{Y})_j = \sum_{l=0}^{2^n-1} (\prod_{i=1}^m t_{j_i,l}^i) K(\mathbf{X})_l.$$

If we now define

$$\prod_{i=1}^m t_{j_i,l}^i = a_{j,l} \tag{29}$$

we obtain (23) because it is easy to show that $\mathbb{A}$ is a scrambler .

Finally, we prove (25). Eq. (29) implies that

$$\begin{pmatrix} a_{0,l} \\ \vdots \\ a_{2^m-1,l} \end{pmatrix} = \begin{pmatrix} t_{0,l}^m \\ \\ t_{1,l}^m \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} t_{0,l}^1 \\ \\ t_{1,l}^1 \end{pmatrix}. \tag{30}$$

Because of the definition of the $\odot$ operator and $\mathbb{T}^i$, (30) is equivalent to

$$\mathbb{A} = \mathbb{T}^m \odot \cdots \odot \mathbb{T}^1.$$

$\Uparrow$ Given $\mathbb{A}$, we show how to construct the matrix $\mathbb{T}$.

We can go backwards through the proof of $\Downarrow$ if we show that to every scrambler $\mathbb{A}$ there correspond scramblers $\mathbb{T}^i$ so that (30) holds. These $\mathbb{T}^i$ will uniquely determine $\mathbb{T}$ because of the definition of $\mathbb{T}^i$.

For a given l, call l' the index for which $a_{l',l} = 1$. Since $\mathbb{A}$ is a scrambler, l' is determined uniquely by l. Define $\{l_i\}_{i=1}^m$ with $l_i \in \{0, 1\}$ by:

$$l' = \sum_{i=1}^{m} l_i 2^{i-1}.$$

Call $t^i_{0,l} = 1 - l_i$ and $t^i_{1,l} = l_i$. Applying the definition of the Kronecker product, we obtain (30).

□

Proof of Theorem 3.1:

We know that $\mathcal{B}^p(\mathbf{Y}) = T \triangle \mathcal{B}^p(\mathbf{X})$ iff $vec(\mathcal{B}^p(\mathbf{Y})) = (\mathbb{T}^m \odot \cdots \odot \mathbb{T}^1) vec(\mathcal{B}^p(\mathbf{X}))$. Since $vec(\mathcal{B}^p(\mathbf{Y})) = EK(\mathbf{Y})$, we obtain (16) and (18) by applying (24) and (25) from Lemma 3.2. Eq. (15) follows then from Lemma 3.1 and, (17) and (19) from Lemma 3.2.

□

Acknowledgement

The authors like to thank an associate editor for a variety of suggestions that helped in the revision of the paper. Also the comments of the referee have been incorporated in the revision.

References

[1] Bergsma, W. (1997). *Marginal Models for Categorical Data*. Tilburg University Press.

[2] Bishop, Y., Fienberg, S. and Holland, P. (1975). *Discrete Multivariate Analysis*. The MIT Press.

[3] Bloomfield, P. (1974). Linear transformations for multivariate binary data. *Biometrics* **30** 609-617.

[4] Ekholm, A., Smith, P., and MCdonald, J. (1994). *Marginal Regression Analysis of a Multivariate Binary Response*. Technical Report 94-7, University of Southampton.

[5] Goodman, L. (1991). Measures, models and graphical displays in the analysis of cross-classified data. *J. Amer. Statist. Assoc.* **86** 1085-1110.

[6] Grizzle, J., Starmer C. and Koch, G. (1969). Analysis of categorical data by linear models. *Biometrics* **25** 489-504.

[7] Haber, M. (1986). Testing for pairwise independence. *Biometrics* **42** 429-435.

[8] Haberman, S. (1974). *The Analysis of Frequency Data*. The University of Chicago Press.

[9] Hagenaars, J. (1990). *Categorical Longitudinal Data*. Sage Publications.

[10] Hohn, F.E. (1966). *Applied Boolean Algebra: an Elementary Introduction*. The Macmillan Company.

[11] Hojsgaard, S., and Skjoth, F. (1991). *Split Models: an Extension of Graphical Association Models*. Institute for Electronic Systems, University of Aalborg.

[12] Khatri, C.G., and Rao, C.R. (1968). Solutions to some functional equations and their applications to characterizations of probability distributions. *Sankhya* **30** 167-180.

[13] Koch, G., Landis, J., Freeman, J., Freeman, D., and Lehnen, R. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics* **33** 133-158.

[14] Lang, J. (1994). On likelihood methods for generalized loglinear models. In *Proceedings 9th International Workshop on Statistical Modelling, Exeter (U.K.)*. University of Exeter.

[15] Lang, J. and Agresti, A. (1994). Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *J. Amer. Statist. Assoc.* **89** 625-632.

[16] Sadaghiani, M.K. (1991). *Models for Discrete Multivariate Data Analysis*. Doctoral Dissertation, K.U.Leuven.

[17] Reynolds, J. (1976). *Analysis of Nominal Data*. Sage University Paper series on Quantitative Application in the Social Sciences, Sage University.

[18] Teugels, J.L. (1990) . Some representations of the multivariate Bernoulli and binomial distributions. *J. Multivariate Anal.* **32** 256-268.

[19] Van Horebeek, J. (1994). *Het Modelleren van Nominale Categorische Data*. Doctoral Dissertation, K.U.Leuven.

[20] Zhao, L.P., and Prentice, R.L. (1990). Correlated binary regresion using a quadratic exponential model. *Biometrika* **77** 642-648.