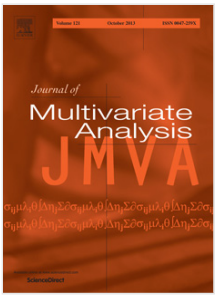

Binary distributions of concentric rings

Nanny Wermuth, Giovanni M. Marchetti, Piotr Zwiernik

PII: S0047-259X(14)00119-5
DOI: http://dx.doi.org/10.1016/j.jmva.2014.05.010
Reference: YJMVA 3756

To appear in: *Journal of Multivariate Analysis*

Received date: 8 December 2013

Please cite this article as: N. Wermuth, G.M. Marchetti, P. Zwiernik, Binary distributions of concentric rings, *Journal of Multivariate Analysis* (2014), http://dx.doi.org/10.1016/j.jmva.2014.05.010

Binary distributions of concentric rings

Nanny Wermuth[1]

Departm. of Math. Sciences, Chalmers University of Technology, Sweden, and Departm. of Medical Psychology and Medical Sociology, Johannes Gutenberg-University, Germany

Giovanni M. Marchetti

Dipartm. di Statist., Informatica, Applicazioni, "G. Parenti", University of Florence, Italy

Piotr Zwiernik

Department of Statistics, University of California, Berkeley, USA

Abstract: *We introduce families of jointly symmetric, binary distributions that are generated over directed star graphs whose nodes represent variables and whose edges indicate positive dependences. The families are parametrized in terms of a single parameter. It is an outstanding feature of these distributions that joint probabilities relate to evenly spaced concentric rings. Kronecker product characterizations make them computationally attractive for a large number of variables. We study the behaviour of different measures of dependence and derive maximum likelihood estimates when all nodes are observed and when the inner node is hidden.*

Key words: Conditional independence; Graphical Markov models; Jointly symmetric distributions; Labeled trees; Latent class models; Phylogenetic trees; Star graphs; Symmetric variables.

1. Introduction

We define and study a family of distribution for $p = 1, 2, \ldots$ binary random variables, denoted by $A_1, \ldots, A_Q, L$. Each variable has equally probable levels, so that the variables are symmetric. There are Q response variables $A_1, \ldots, A_Q$, to a single common explanatory variable L, named the signal and having the levels strong or weak. The possible responses are to succeed or to miss. We use as a convention that success for A_q is coded 1 and that a strong signal of L is also coded 1. For the low level, we use either -1 or 0. Of special interest are situations in which the signal cannot be directly observed, it is instead hidden or latent, but the aim is to understand and estimate the joint structure including L. In that case, we have $t = 1, \ldots, 2^Q$ level combinations.

We let $K_t = a_1 + \cdots + a_Q$ denote the number of ones in any given sequence of response-level combinations, $(a_1, \ldots, a_Q)$, and define a normalizing constant, $c_Q = 2(1+\alpha)^Q$ for $1 \leq \alpha < \infty$, to write with $\{0, 1\}$ coding, also known as baseline coding, for the joint p-dimensional distribution

$$\pi(a_1, \ldots, a_Q, l)\, c_Q \quad = \quad \begin{cases} \alpha^{K_t} & \text{for } l = 1, \\ \alpha^{(Q-K_t)} & \text{for } l = 0\,. \end{cases} \tag{1}$$

[1]Corresponding author: wermuth@chalmers.se

For the $\{-1, 1\}$ coding of the levels, known also as effect coding, the symmetry of each of the binary variables implies zero mean and unit variance. For L, we write

$$pr(L = 1) = pr(L = -1) = \tfrac{1}{2}, \quad E(L) = 0, \quad E(L^2) = 1.$$

For any such binary variable pair (A, L), the correlation coefficient ρ, which is

$$\rho \;=\; \mathrm{cov}(A, L) \;=\; E(A\,L),$$

ranges in $0 \le \rho < 1$ and

$$\alpha = (1 + \rho)/(1 - \rho), \qquad \rho = (\alpha - 1)/(\alpha + 1)\,. \tag{2}$$

The correlation ρ is also the regression coefficient in a projection of A on L. Furthermore, independence of A from L, denoted by $A \perp\!\!\!\perp L$, relates to α and ρ via

$$A \perp\!\!\!\perp L \iff (\alpha = 1) \iff (\rho = 0).$$

This last case would give a degenerate model in equation (1), hence it is excluded for some purposes. The following Table 1 shows how two types of sequences of ratios for ρ generate all possible even and odd positive integers for α and hence proper counts in equation (1).

Table 1: An integer valued α for symmetric binary variables in concentric-ring models

α	1	3	5	7	9	11	13	15	...
ρ	0	1/2	2/3	3/4	4/5	5/6	6/7	7/8	...
α	2	4	6	8	10	12	14	16	...
ρ	1/3	3/5	5/7	7/9	9/11	11/13	13/15	15/17	...

As will be shown, a model with density given by equation (1) has several attractive features that were not previously identified even though it is a special case of a number of models that have been intensively studied. For instance, it is a distribution generated over a labeled tree (Castelo and Siebes, 2003), hence a lattice-conditional-independence model (Perlman and Wu, 1999) and a directed-acyclic-graph model (Wermuth and Lauritzen, 1983; Pearl, 1988) or a Markov field for binary variables (Darroch, Lauritzen and Speed, 1980), an Ising model of ferromagnetism, a binary quadratic exponential distribution (Besag, 1974; Cox and Wermuth, 1994) and a triangular system of symmetric binary variables (Wermuth, Marchetti and Cox, 2009).

With L in equation (1) unobserved, the resulting model may be regarded as a simplest case for constructing phylogentic trees; see Zwiernik and Smith (2011), Allman et al. (2014) and the previous extensive literature in this area. Or, it can be viewed as a special latent-class model (Lazarsfeld, 1950; Linzer and Lewis, 2011), the one with the closest analogy to a Gaussian factor analysis model having a single factor.

A **star graph** is a directed-acyclic graph with one inner node, L, from which Q arrows start and point to the uncoupled, outer nodes, $1, \ldots, Q$. For $p = 6$, the left of

Figure 1 shows such a star graph, having equal regression coefficients ρ when regressing each A_q on L, for $q = 1, \ldots, Q$.

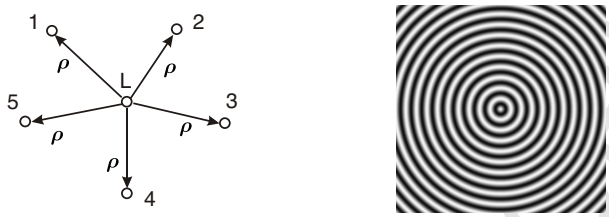

Figure 1: Star graph with equal dependences of five leaves on one common root (left) and a graph of evenly-spaced concentric rings (right).

For Gaussian and for binary distributions generated over star graphs as those in Figure 1, the correlation matrices of the p variables are of identical form; see Wermuth and Marchetti (2014). For $p = 5$, such correlation matrices are in Table 2, with '.' indicating a symmetric entry.

Table 2: Correlation matrix for $p = 5$; left: to equation (1), right: to a binary latent class model

$$\begin{pmatrix} 1 & \rho^2 & \rho^2 & \rho^2 & \rho \\ . & 1 & \rho^2 & \rho^2 & \rho \\ . & . & 1 & \rho^2 & \rho \\ . & . & . & 1 & \rho \\ . & . & . & . & 1 \end{pmatrix} \qquad \begin{pmatrix} 1 & \rho_1\rho_2 & \rho_1\rho_3 & \rho_1\rho_4 & \rho_1 \\ . & 1 & \rho_2\rho_3 & \rho_2\rho_4 & \rho_2 \\ . & . & 1 & \rho_3\rho_4 & \rho_3 \\ . & . & . & 1 & \rho_4 \\ . & . & . & . & 1 \end{pmatrix}.$$

Another feature of the joint probabilities in (1) is that the conditional odds-ratios for each pair A_q, L given the remaining $Q-1$ variables are equal to α^2. When one interprets these as equal distances, concentric rings such as those on the right of Figure 1 may result. The number of rings increases with an increase of Q as illustrated with Table 6 in Section 3. This explains the chosen name of this family of distributions. First, we generate the distributions over star graphs.

2. Generating a concentric-ring model over a star graph

To shorten descriptions and notation, we call both, the outer nodes of the star graph and the corresponding binary variables $A_1, \ldots, A_Q$, **leaves** and identify them sometimes by their indices $1, \ldots, Q$. Similarly we call both the inner node of the star graph, and variable L, the **root**. Often the root is unobserved, that is latent or hidden, and one main task is then to estimate the joint p-dimensional distribution from observations on only the Q leaves.

The first main feature of a joint distribution of concentric rings is mutual conditional independence of the leaves given the root, written as

$$(1 \perp\!\!\!\perp 2 \perp\!\!\!\perp \ \ldots \ \perp\!\!\!\perp Q) \,|L. \tag{3}$$

Any density generated over a star graph, irrespective of the types of variables, is defined by Q conditional densities, $f_{q|L}$, and a marginal density, f_L, of the root. In the condensed node notation, with node set $N = \{1, \ldots, Q, L\}$ of size p, the joint density f_N factorizes as

$$f_N = f_{1|L} \cdots f_{Q|L} f_L. \tag{4}$$

For binary variables, f_N denotes the joint probability distribution, so that equation (4) becomes $\pi(a_1, \ldots, a_Q, l) = \pi(a_1|l) \cdots \pi(a_Q|l)\pi(l)$ where $\pi(a_q|l) = \pi(a_q, l)/\pi(l)$ are obtained from the bivariate probabilities of each leave, A_q, and the root, L. In Table 3, we show probabilities for any binary pair (A, L) and for each variable of the the pair being symmetric.

Table 3: A 2×2 table of a general (A, L) and in the special case of symmetric binary variables

	L		
A	weak	strong	sum
miss	$\pi_{\boldsymbol{mw}}$	$\pi_{\boldsymbol{ms}}$	$\pi_{\boldsymbol{m}}$
succeed	$\pi_{\boldsymbol{sw}}$	$\pi_{\boldsymbol{s}s}$	$1 - \pi_{\boldsymbol{m}}$
sum	$\pi_{\boldsymbol{w}}$	$\pi_{\boldsymbol{s}}$	1

	L		
A	weak	strong	sum
miss	$\frac{1}{4}(1+\rho)$	$\frac{1}{4}(1-\rho)$	$\frac{1}{2}$
succeed	$\frac{1}{4}(1-\rho)$	$\frac{1}{4}(1+\rho)$	$\frac{1}{2}$
sum	$\frac{1}{2}$	$\frac{1}{2}$	1

Several standard measures of dependence, that are in common use, are defined in Table 4 by using Table 3 and equation (2), both for a general binary pair (A, L) and for it being symmetric.

Table 4: Measures in a general 2×2 table and in the special case of symmetric binary variables

definition	interpretation in general and for two symmetric binary variables	
$\pi_{\boldsymbol{ss}}/\pi_{\boldsymbol{ms}}$	odds of succeeding versus missing given a strong signal:	α
$\pi_{\boldsymbol{sw}}/\pi_{\boldsymbol{mw}}$	odds of succeeding versus missing given a weak signal:	$1/\alpha$
$(\pi_{\boldsymbol{ss}}\pi_{\boldsymbol{mw}})/(\pi_{\boldsymbol{ms}}\pi_{\boldsymbol{sw}})$	odds-ratio for success or cross-product ratio:	α^2
$\pi_{\boldsymbol{s}|\boldsymbol{s}} = \pi_{\boldsymbol{ss}}/\pi_{\boldsymbol{s}}$	chance to succeed given a strong signal of L:	$(1+\rho)/2$
$\pi_{\boldsymbol{s}|\boldsymbol{w}} = \pi_{\boldsymbol{sw}}/\pi_{\boldsymbol{w}}$	chance to succeed given a weak signal of L:	$(1-\rho)/2$
$\pi_{\boldsymbol{s}|\boldsymbol{s}} - \pi_{\boldsymbol{s}|\boldsymbol{w}}$	chance difference in succeeding:	ρ
$\pi_{\boldsymbol{s}|\boldsymbol{s}}/\pi_{\boldsymbol{s}|\boldsymbol{w}}$	relative chance for success:	α

For both A, L symmetric, the parameter $\rho > 0$ relates also directly to the probabilities via

$$\rho = (\pi_{\boldsymbol{ss}} + \pi_{\boldsymbol{mw}}) - (\pi_{\boldsymbol{sw}} + \pi_{\boldsymbol{ms}}), \tag{5}$$

and the odds of succeeding versus missing given a strong signal of L coincides with the relative chance for success. Independence of any binary pair (A, L) requires in general, that the odds-ratio equals one, the relative chance equals one and the chance difference equals zero.

For the relation of α to conditional independence given L, we only look at pair (A_1, A_2) at both levels of L in Table 5, since the mutual independence in equation (3) implies independence of each pair of leaves from the remaining $Q-2$ leaves given L, in particular $(1,2) \perp\!\!\!\perp (3,\ldots,Q)|L$.

The conditional independence $1 \perp\!\!\!\perp 2|L$ is directly reflected in the equal-one odds-ratios within the subtables for each level of L. The same holds for the relative chances, while the chance difference and the correlation coefficient in each subtable for (A_1, A_2) are zero for $1 \perp\!\!\!\perp 2|L$. Table 5 shows in addition the **joint symmetry of the distribution** since the probability for any given level combination of the variables remains unchanged after switching all the levels.

Table 5: Probabilities with $1 \perp\!\!\!\perp 2|L$ multiplied by $c_2 = 2(1+\alpha)^2$ for pair (A_1, A_2) given L

	weak L		strong L		
A_1	A_2 miss	A_2 succeed	A_2 miss	A_2 succeed	sum
miss	α^2	α	1	α	$(1+\alpha)^2$
succeed	α	1	α	α^2	$(1+\alpha)^2$
sum	$\alpha(1+\alpha)$	$(1+\alpha)$	$(1+\alpha)$	$\alpha(1+\alpha)$	$2(1+\alpha)^2$
odds-ratio	1		1		

Joint symmetry also holds in general, as can be derived directly from (1). It follows that the marginal distribution of each (A_q, L) is symmetric and does not depend on q. For $\{-1,1\}$ coding, we have then from this symmetry and equation (4), for the generated joint distribution in (1) and with $q = 1,\ldots,Q$

$$\pi(a_1,\ldots,a_Q,l) \;=\; 2^{-p}\textstyle\prod_q (1+\rho\, a_q l). \tag{6}$$

3. Kronecker product representations of joint probabilities

We now introduce for $p \geq 3$ a vector representation. For this, we write for instance $\pi_{111} = pr(A_1 = 1, A_2 = 1, L = 1)$. Then, by using again $N = \{1,\ldots,Q,L\}$ and the $\{0,1\}$ coding and letting the levels of the first variable change fastest, the column vector of probabilities, $\boldsymbol{\pi}_{3,N}$, is in transposed form

$$\begin{aligned} \boldsymbol{\pi}_{3,N}^{\mathrm{T}} &= (\pi_{000},\ \pi_{100},\ \pi_{010},\ \pi_{110},\ \pi_{001},\ \pi_{101},\ \pi_{011},\ \pi_{111}) \\ &= (\alpha^2,\ \alpha,\ \alpha,\ 1,\ 1,\ \alpha,\ \alpha,\ \alpha^2)/c_2, \end{aligned}$$

where $c_2 = 2(1+\alpha)^2$ and we take in this notation always the last variable to coincide with L.

For an integer-valued α, we illustrate next how the concentric rings are generated and increase with the number of variables. One way to generate the probabilities after an increase from p to $p+1$ nodes, is to start with the probabilities at the strong signal of L for the given p, multiplied by $c_Q = 2(1+\alpha)^Q$, to obtain first a vector of powers of α such as in Table 6.

This vector is appended next by the same vector modified just by increasing the power of each α by one. The joint probabilities for a strong signal of L for $p+1$ nodes result after dividing by the new normalizing constant $c_{(Q+1)} = 2(1+\alpha)^Q$ and repeating the probabilities in reverse order for the lower half of the table.

Table 6: Integer parametrization of the upper half of the probability vector for $p = 1$ up to $p = 5$ variables; with the sum of the integers equal to $2(1+\alpha)^Q$, the number of leaves equal to $Q = p-1$

p	moving from p to $p+1$ using powers of α															
1	α^0															
2	α^0	α^1														
3	α^0	α^1	α^1	α^2												
4	α^0	α^1	α^1	α^2	α^1	α^2	α^2	α^3								
5	α^0	α^1	α^1	α^2	α^1	α^2	α^2	α^3	α^1	α^2	α^2	α^3	α^2	α^3	α^3	α^4

For large p, the row vector $\boldsymbol{\pi}_{p,N}^{\mathrm{T}}$ has a computationally attractive representation in terms of Kronecker products. Let $v = (1,\ \alpha)$, $w = (\alpha,\ 1)$ and $c_Q = 2(1+\alpha)^Q$, then $\boldsymbol{\pi}_{p,N}^{\mathrm{T}}$ may be obtained from

$$(\underbrace{w \otimes \cdots \otimes w}_{p-1},\ \underbrace{v \otimes \cdots \otimes v}_{p-1})/c_Q\ . \tag{7}$$

From the given form of the joint distribution, it can be checked directly that for any $p > 2$ and any selected pair (A_q, L), the conditional cross-product ratios equal α^2, the conditional relative chances for success equal α and the conditional chance differences in succeeding equal ρ, that is in all subtables formed by the level combinations of the remaining leaves.

Collapsibility results for the three measures show that these three measures remain unchanged after marginalizing over some or all of the remaining leaves if these are conditionally independent of A_q given L; see Wermuth (1987) and Xie, Ma and Geng (2008). The common strength of dependence of each A_q on L gives an increase of the number of concentric rings as p increases.

To compute moments and other features of the distribution in a fast way, we show in the next section that Kronecker products based on special 2×2 matrices are particularly helpful, since for instance the inverses of such products are the Kronecker products of the inverses.

4. Moments, interactions and sums of level combinations of the leaves

The $\{0, 1\}$ coding of binary variables is well suited to understand the change from raw and from central moments, in general, to those of the concentric-ring model. With

$$\mathcal{B}_p = \underbrace{\mathcal{B} \otimes \cdots \otimes \mathcal{B}}_{p}, \qquad \mathcal{B} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

the column vector of raw moments is, in general binary-star-graph models,

$$\boldsymbol{m}_{p,N} = \mathcal{B}_p\ \boldsymbol{\pi}_{p,N}\ . \tag{8}$$

For the concentric-ring distribution and $p = 3$, the raw moments in $\{0\ 1\}$ coding reduce barely, with e.g. $\pi_{11+} = pr(A_1 = 1, A_2 = 1) = \sum_l \pi_{11l}$ and $\pi_{1++} = \pi_{11+} + \pi_{10+}$, as follows,

$$\begin{aligned} \boldsymbol{m}_{3,N}^{\mathrm{T}} &= (1,\ \pi_{1++},\ \pi_{+1+},\ \pi_{11+},\ \pi_{++1},\ \pi_{1+1},\ \pi_{+11},\ \pi_{111}) \\ &= (1,\ \tfrac{1}{2},\ \tfrac{1}{2},\ \beta,\ \tfrac{1}{2},\ \gamma, \gamma,\ \delta), \end{aligned}$$

where $\beta = (1+\alpha^2)/c_2,\ \ \gamma = \alpha(1+\alpha)/c_2,\ \ \delta = \alpha^2/c_2,\ \ c_2 = 2(1+\alpha)^2$.

Another Kronecker product leads to central moments of $\{0,1\}$-coded binary variables; see Teugels (1990). For instance with $q = 1, \ldots, Q$ and $\boldsymbol{\mathcal{T}}_{p,N} = \boldsymbol{\mathcal{T}}_1 \otimes \cdots \otimes \boldsymbol{\mathcal{T}}_Q \otimes \boldsymbol{\mathcal{T}}_L$, where

$$\boldsymbol{\mathcal{T}}_q = \begin{pmatrix} 1 & 1 \\ -\,pr(A_q = 0) & pr(A_q = 1) \end{pmatrix} \quad \boldsymbol{\mathcal{T}}_L = \begin{pmatrix} 1 & 1 \\ -\,pr(L = 0) & pr(L = 1) \end{pmatrix},$$

the vector of central moments is

$$\boldsymbol{\mu}_{p,N} = \boldsymbol{\mathcal{T}}_{p,N}\ \boldsymbol{\pi}_{p,N}\,. \tag{9}$$

For concentric-ring distribution and $p = 3$, the central moments reduce with $\gamma = \rho/4$ to

$$\begin{aligned} \boldsymbol{\mu}_{3,N}^{\mathrm{T}} &= (1,\ 0,\ 0,\ \mu_{12},\ 0,\ \mu_{13},\ \mu_{23},\ \mu_{123}) \\ &= (1,\ 0,\ 0,\ 4\gamma^2,\ 0,\ \gamma,\ \gamma,\ 0)\,. \end{aligned}$$

By the mixed-product property of Kronecker products, simple relations result, such as for instance

$$\boldsymbol{\mu}_{p,N} = \boldsymbol{\mathcal{T}}_1 \boldsymbol{\mathcal{B}}^{-1} \otimes \cdots \otimes \boldsymbol{\mathcal{T}}_Q \boldsymbol{\mathcal{B}}^{-1} \otimes \boldsymbol{\mathcal{T}}_L \boldsymbol{\mathcal{B}}^{-1}\, \boldsymbol{m}_{p,N}\,.$$

By contrast, the $\{-1, 1\}$ coding of binary variables is well suited to express the change from general log-linear interactions to those that are much simpler in the concentric-ring model. With

$$\boldsymbol{\mathcal{E}}_p = \underbrace{\boldsymbol{\mathcal{E}} \otimes \cdots \otimes \boldsymbol{\mathcal{E}}}_{p}, \qquad \boldsymbol{\mathcal{E}} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$

the vector of log-linear interactions for the probabilities, at the combinations of levels one, are

$$\boldsymbol{\lambda}_{p,N} = \underbrace{\boldsymbol{\mathcal{E}}^{-1} \otimes \cdots \otimes \boldsymbol{\mathcal{E}}^{-1}}_{p} \log(\boldsymbol{\pi}_{p,N})\,. \tag{10}$$

For the concentric-ring distribution and $p = 4$, the log-linear interactions reduce as follows

$$\begin{aligned} \boldsymbol{\lambda}_{4,N}^{\mathrm{T}} &= (\lambda_-,\ \lambda_1,\ \lambda_2,\ \lambda_{12},\ \lambda_3,\ \lambda_{13},\ \lambda_{23},\ \lambda_{123},\ \lambda_4,\ \lambda_{14},\ \lambda_{24},\ \lambda_{124},\ \lambda_{34},\ \lambda_{134},\ \lambda_{234},\ \lambda_{1234}) \\ &= (\beta,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ \gamma,\ \gamma,\ 0, \gamma,\ 0,\ 0,\ 0), \end{aligned}$$

with $\gamma = \frac{1}{2}\log(\alpha),\ \beta = 3\gamma - \log(c_3)$.

In general, only the 2-factor terms that include L and the overall normalizing constant λ_- are nonzero. In the log-linear parametrisation, conditional independence of any pair implies that all higher-order interaction terms involving this pair are vanishing as well;

see e.g. Fienberg (2007). Thus, the independences of equation (3) lead to all other log-linear interaction terms being zero.

For binary variables, the linear interactions may in general be defined with the same Kronecker product matrix as used for the log-linear interactions in equation (10)

$$\boldsymbol{\xi}_{p,N} = \boldsymbol{\mathcal{E}}_p \, \boldsymbol{\pi}_{p,N} \, . \tag{11}$$

These linear interactions reduce for the concentric-ring distribution and $p = 4$ as follows

$$\begin{aligned} \boldsymbol{\xi}_{4,N}^{\mathrm{T}} &= (1,\ \xi_1,\ \xi_2,\ \xi_{12},\ \xi_3,\ \xi_{13},\ \xi_{23},\ \xi_{123},\ \xi_4,\ \xi_{14},\ \xi_{24},\ \xi_{124},\ \xi_{34},\ \xi_{134},\ \xi_{234},\ \xi_{1234}) \\ &= (1,\ 0,\ 0,\ \rho^2,\ 0,\ \rho^2,\ \rho^2,\ 0,\ 0,\ \rho,\ \rho,\ 0,\ \rho,\ 0,\ 0,\ \rho^3) \, . \end{aligned}$$

From equations (9), and (11) and from $\boldsymbol{\mathcal{E}}\boldsymbol{\mathcal{T}}_q^{-1}$ being of diagonal form, the linear-interaction terms in $\boldsymbol{\xi}_{p,N}$ are just rescaled versions of the central moments $\boldsymbol{\mu}_{p,N}$. They are the standardized central moments, that result after transforming the $\{0, 1\}$ coded binary variables A_q, say, with mean $\frac{1}{2}$ and variance $\frac{1}{4}$, into their standardized form with $\{-1, 1\}$-coding, that is into $A_q^* = 2A_q - 1$.

This central moment representation is more complex than the log-linear formulation because of the non-vanishing 4-factor interaction term ρ^3. For $Q > 3$, each odd-order interaction is zero, an even-order k-factor interaction involving the root as the last variable, is ρ^{k-1} and it is ρ^k, otherwise. As we shall see, this advantage of the log-linear interactions disappears in the marginal distribution of the leaves which has no independences.

For later use, we introduce the sum S, and the average $\bar{S}$, of the Q standardized variables A_q^*. Under the concentric ring model, each pair has the same correlation ρ^2, see Table 2, so that

$$\mathrm{var}(S) = Q + 2\binom{Q}{2}\rho^2, \quad Q\,\mathrm{var}(\bar{S}) = 1 + (Q-1)\,\rho^2. \tag{12}$$

Also directly from the right of Table 3, one sees that $E(A_q^*|L = 0) = -\rho$ and $E(A_q^*|L = 1) = \rho$ so that

$$E(\bar{S}|L = 1) - E(\bar{S}|L = 0) = 2\rho. \tag{13}$$

5. Marginal distributions of the leaves

By the joint symmetry, marginalizing over the common root returns a symmetric distribution by construction. This is illustrated in Table 7 for $Q = 3$ leaves.

Table 7: Marginalising over L for $Q = 3$ by adding α's for corresponding level combinations of the leaves; in the table each probability is multiplied by $c_3 = 2(1 + \alpha)^3$

2^3 levels:	000	100	010	110	001	101	011	111
at level $l = 0$:	α^0	α^1	α^1	α^2	α^1	α^2	α^2	α^3
at level $l = 1$:	α^3	α^2	α^2	α^1	α^2	α^1	α^1	α^0
margin over L :	$1 + \alpha^3$	$\alpha + \alpha^2$	$\alpha + \alpha^2$	$\alpha + \alpha^2$	$\alpha + \alpha^2$	$\alpha + \alpha^2$	$\alpha + \alpha^2$	$1 + \alpha^3$

In general, after marginalizing over L, the distribution of the remaining Q leaves is given, with K_t denoting again the number of ones in any sequence of levels, $(a_1, \ldots, a_Q)$, by

$$\pi(a_1, \ldots, a_Q) = \frac{1}{c_Q} \left(\alpha^{K_t} + \alpha^{(Q-K_t)} \right). \tag{14}$$

One also obtains the linear interaction vector for the joint distribution of the leaves, $\boldsymbol{\xi}_{p,N\setminus L}$ as the lower half of $\boldsymbol{\xi}_{p,N}$, where in this notation, we do not distinguish between an element L and the singleton $\{L\}$.

Equivalently, with $\xi_0 = 1$ and $\mathcal{I} \subseteq \{1, \ldots, Q\}$ such that $q \in \mathcal{I}$ if and only if $a_q = 1$ is in $(a_1, \ldots, a_Q)$, as used before in the example to equation (11), the other elements of $\boldsymbol{\xi}_{p,N\setminus L}$ may be written as

$$\xi_{\mathcal{I}} = \begin{cases} \rho^{K_t} & \text{for even } K_t, \\ 0 & \text{otherwise .} \end{cases} \tag{15}$$

Also $\boldsymbol{\lambda}_{p,N\setminus L} = \boldsymbol{\mathcal{E}}_Q^{-1} \log\{(\boldsymbol{\mathcal{E}}_Q)^{-1} \boldsymbol{\xi}_{p,N\setminus L}\}$ has zero values in the same positions as $\boldsymbol{\xi}_{p,N\setminus L}$. Thus, all odd-order log-linear interactions vanish and all odd-order (standardized) central moments vanish. The even-factor terms are functions of ρ^2 which is the induced marginal correlation for any pair of leaves and, at the same time, the induced difference in chances for success.

6. The conditional distribution of the root given the leaves

From equations (1) and (14), the conditional distribution, $\pi(l|a_1, \ldots, a_Q)$, of the root, L, given the leaves, $A_1, \ldots A_Q$, satisfies in terms of α, the number of ones, K_t, in the leaf-level sequence $(a_1, \ldots, a_Q)$

$$c_Q\, \pi(a_1, \ldots, a_Q)\, \pi(l|a_1, \ldots, a_Q) = \begin{cases} \alpha^{K_t} & \text{for } l = 1, \\ \alpha^{(Q-K_t)} & \text{for } l = 0\,. \end{cases} \tag{16}$$

Functions of the odds-ratio are known to be the only measures of dependence in 2×2 tables that are variation independent of the margins; see Edwards (1963). By using the concentric-ring model, we illustrate now how the relative chance and the chance difference may give strongly distorted impressions of equal conditional dependences.

When the roles of explanatory variable L and responses in the given generating process are exchanged, the odds-ratios stay constant, equal chance differences appear to be of sharply reduced strengths and equal relative risks appear to be strongly unequal.

To see this, we compare the dependences of A_2 on L given A_1 using the odds-ratio, $\mathrm{odr}(A_2, L|A_1)$, the chance differences to succeed, $\mathrm{chd}(A_2, L|A_1)$, and the relative chances to succeed, $\mathrm{rch}(A_2, L|A_1)$, with the corresponding dependences of L on A_2 given A_1.After exchanging the ordering to $(A_2,\ L,\ A_1)$ in Table 5 and taking as an example $\alpha = 9$:

$$\mathrm{odr}(A_2, L\,|A_1 = a_1) = \alpha^2 = 81,\ \ \mathrm{chd}(A_2, L|A_1 = a_1) = \rho = 0.80,\ \ \mathrm{rch}(A_2, L|A_1 = a_1) = 9,$$

while from Table 8 with L as the first variable and A_2 as the second, one obtains

$$\mathrm{odr}(L, A_2\,|A_1 = 1) = 81, \qquad \mathrm{chd}(L, A_2\,|A_1 = 1) = 0.49, \qquad \mathrm{rch}(L, A_2\,|A_1 = 1) = 41,$$

$$\mathrm{odr}(L, A_2\,|A_1 = 0) = 81, \qquad \mathrm{chd}(L, A_2\,|A_1 = 0) = 0.49, \qquad \mathrm{rch}(L, A_2\,|A_1 = 0) = 1.98.$$

Table 8: Probabilities multiplied by $2(1+a)^2$ for (A_1, A_2, L) in Table 5 reordered as (L, A_2, A_1)

	A_1 miss		A_1 succ.	
level l of L	A_2 miss	A_2 succeed	A_2 miss	A_2 succ.
0 := weak	α^2	α	α	1
1 := strong	1	α	α	α^2
sum	$(1+\alpha^2)$	2α	2α	$(1+\alpha^2)$
odds-ratio for $l=1$, $a_2=1$; $\mathrm{odr}(L, A_2\|A_1)$	α^2		α^2	
relative chance for $l=1$; $\mathrm{rch}(L, A_2\|A_1)$	$(1+\alpha^2)/2$		$2\alpha^2/(1+\alpha^2)$	
chance difference for $l=1$; $\mathrm{chd}(L, A_2\|A_1)$	$\frac{1}{2} - 1/(1+\alpha^2)$		$\alpha^2/(1+\alpha^2) - \frac{1}{2}$	

We notice next that in a logit regression of L on the leaves, A_q, the regression parameters are functions of the conditional odds-ratios for (L, A_q) since they may be obtained from twice the log-linear parameters $\boldsymbol{\lambda}_{p,N}$ in equation (4) that do not involve L. This follows from the definition of the joint probabilities in (1) and the logit representation

$$\text{logit}\,\{\pi(l|a_1, \ldots a_Q)\} = \log \pi(a_1, \ldots, a_Q, 1) - \log \pi(a_1, \ldots, a_Q, 0).$$

Thus, the odds-ratio and this logistic regression coefficient are unaffected by switching the roles of A_2 and L, while the strength of dependence measured with the chance difference is reduced in the example from 0.80 by almost 40% to 0.49 and the dependences measured with equal relative chances of 9 for A_2 on L , are modified into 41 and about 2, thus clearly into strongly different strengths of dependence at the two levels of A_1. As Q increases, the relative chance for a strong signal, comparing succeeding to missing in A_1, increases even to $(1+\alpha^Q)/2$ at $Q-1$ misses of the remaining variables.

Such changes illustrate potential problems for machine learning and causal conclusions, for interpretations of some case-control studies and for some uses of the propensity score.

7. Maximum-likelihood estimates

One of the most attractive properties of the maximum-likelihood estimate of a set of parameters in a given model is that the maximum-likelihood estimate of any other set of parameters, related to the original ones by a one-to-one (1–1) transformation, is given by the same 1–1 transformation for the estimates; see Fisher (1922). Thus here, given the maximum-likelihood estimate $\hat{\rho}$ of ρ, all other measures of dependence are defined by the relevant 1–1 transformations. Given $\hat{\alpha}$, the maximum-likelihood estimates of the log-linear interactions are also given. Furthermore, other estimated interactions of interest, as well as the joint probabilities, result via the 1–1 transformations of Section 4.

Given the observed frequencies, for a pair (A, L) of symmetric binary variables that sum to n in vector $\boldsymbol{n}_{2,N}^{\mathrm{T}} = (n_{00},\ n_{10},\ n_{01},\ n_{11})$, one obtains with equations (6) and (5)

$$\hat{\rho} = \{(n_{00} + n_{11}) - (n_{01} + n_{10})\}/n := \ \mathrm{csd}_{AL}, \tag{17}$$

where 'csd' abbreviates '**cross-sum difference**', a term introduced by G.M. Marchetti in recent unpublished work. For symmetric variables A_1, A_2, L observed and satisfying $1 \perp\!\!\!\perp 2 \mid L$ and $E(A_1\, L) = E(A_2\, l) = \rho$, given the vector of counts $\boldsymbol{n}_{3,N}^{\mathrm{T}}$, we get the average of the two cross-sum differences as the unique maximum-likelihood estimate

$$\hat{\rho} = \tfrac{1}{2}(\mathrm{csd}_{1L} + \ \mathrm{csd}_{2L})$$

of the common correlation. Similarly, for observations $n(a_1, \ldots, a_Q, l)$ on $A_1, \ldots, A_Q, L$ of a concentric-ring model, the closed-form maximum-likelihood estimate of ρ equals the average of the $q = 1, \ldots, Q$ cross-sum differences in counts for each leaf-root pair (A_q, L):

$$\hat{\rho} = \frac{1}{Q} \sum_q \mathrm{csd}_{qL}\,. \tag{18}$$

When L is hidden, it can be shown for $Q = 2$, that the maximum-likelihood estimate of ρ^2 equals the observed cross-sum difference, and for $Q = 3$, that it equals the average of the three observed cross-sum differences. For $Q > 3$, there is in general no closed-form solution of the likelihood equation to estimate ρ^2, but a method-of-moment estimator $\tilde{\rho}^2$ of ρ^2 is obtained from equation (12) as

$$\tilde{\rho}^2 = (Q\, \hat{v} - 1)/(Q - 1) \tag{19}$$

where $\hat{v}$ is any sample estimate of $\mathrm{var}(\bar{S})$, and v and $\bar{S}$ are as defined for equation (12).

An EM algorithm (Dempster, Laird and Rubin, 1977) for ρ in the concentric ring model can be defined with closed-form solutions both for the E(expectation) and for the M(maximization) steps. In an E-step, the 2^p joint estimated counts $\tilde{n}(a_1, \ldots, a_Q, l)$ are from the observed 2^Q marginal counts of the leaves, $n(a_1, \ldots, a_Q)$, and the conditional distribution of the root given the leaves:

$$\tilde{n}(a_1, \ldots, a_Q, l) = n(a_1, \ldots, a_Q)\tilde{\pi}(l|a_1, \ldots, a_Q)$$

using a current estimate of α and equation (16). In an M-step, the estimated correlation coefficient results with the new 2^p joint counts $\tilde{n}(a_1, \ldots, a_Q, l)$ via equation (13).

Also, the two steps can be combined into a single updating equation for the correlation coefficient. For this, we denote by $\rho(m)$ the value of the correlation coefficient at iteration step m and start with an initial estimate from equation (19), $\rho(0) = (\tilde{\rho}^2)^{1/2}$. Let $n_t = n(a_1, \ldots, a_Q)$ and $s_t = a_1^* + \cdots + a_Q^*$ be the marginal counts and the associated sum in $\{-1, 1\}$ coding, respectively. Then, the updated estimate $\rho(m+1)$ is, with $t = 1, \ldots, 2^Q$, such that

$$\rho(m+1) \ = \frac{1}{nQ} \sum_t T_t(m)\, n_t, \tag{20}$$

where we use the relation between α and ρ in equation (2) to lead to

$$T_t(m) = s_t\{\alpha(m)^{s_t} - 1\}/\{\alpha(m)^{s_t} + 1\}. \tag{21}$$

Notice that, from a table of counts for all p variables and equation (13), an estimate of ρ is

$$\frac{1}{nQ} \sum_t s_t\, \{n(a_1, \ldots, a_Q, 1) - n(a_1, \ldots, a_Q, 0)\}.$$

Then, from equations (14) and (16), at a given iteration of the EM algorithm, we can write

$$\tilde{n}(a_1,\ldots,a_Q,1)-\tilde{n}(a_1,\ldots,a_Q,0)=\frac{\alpha^{K_t}-\alpha^{(Q-K_t)}}{\alpha^{K_t}+\alpha^{(Q-K_t)}}\,n_t=\frac{\alpha^{s_t}-1}{\alpha^{s_t}+1}\,n_t\,,$$

so that equations (20) and (21) follow.

To see that $T_t(m)$ in equation (20) is for $\rho(0)>0$ always nonnegative, note that if $s_t\geq 0$ in equation (21) then also $(\alpha(m)^{s_t}-1)/(\alpha(m)^{s_t}+1)\geq 0$ because $\alpha(m)\geq 1$. Similarly, if $s_t<0$ then $(\alpha(m)^{s_t}-1)/(\alpha(m)^{s_t}+1)\leq 0$.

The algorithm converges to a stationary point of the likelihood and the standard error of the estimate can be found using one of the methods discussed in Tanner (1996, Sect. 4.4, p. 74). In extensive simulations under the model with $Q=4$, the number of iterations required for convergence, for ρ in the range of most interest, in $0.5<\rho<0.8$, was with a tolerance of $\epsilon=10^{-4}$ at most 4 and with a tolerance of $\epsilon=10^{-7}$ at most 20. The absolute difference between $\hat{\rho}$ and ρ was less than 0.1 and less than 0.05, in samples of size 300 and 1000 respectively.

8. Discussion

A family of jointly symmetric distributions in equally probable binary variables has been defined, where for each given number of variables, a distribution is characterized by a single parameter. The family is shown to have several attractive features that were not previously identified even though it is a special case of a number of models that have been intensively studied, such as Ising models of ferromagnetism, latent class structures and models for constructing phylogenetic trees.

In particular, such a distribution is a graphical Markov model, generated over a star graph with $p-1$ leaves and one common root. A positive dependence of each leaf on the root equals a positive Pearson's correlation coefficient, ρ. When p increases with ρ kept fixed, the model leads to an increasing number of evenly-spaced concentric rings.

An integer parametrization shows which sample size is needed so that the smallest count is expected to equal one. This information helps to plan for observed positive distributions, that is for a sufficient condition that the intersection property (see e.g. Pearl, 1988) holds for a given set of observations on symmetric binary variables.

A closed-form maximum-likelihood estimate $\hat{\rho}$ of ρ is obtained when the root is observed in addition to the leaves. Otherwise, a closed form method-of-moment estimate $\tilde{\rho}$ of ρ is derived. This estimate is a good starting value for the EM algorithm which reduces to a single updating equation to obtain $\hat{\rho}$. Simulations suggest that $\tilde{\rho}$ and $\hat{\rho}$ agree often up to the second decimal place, that the likelihood function for ρ has a unique maximum and that it is quite flat only for $\rho\leq 1/3$ that is for the rather small dependences among each leave pair of only $\rho^2\leq 1/9$. With $\hat{\rho}$ estimated just from observations on the leaves, the joint probabilities or interactions including the root are available in terms of Kronecker products of small matrices even for many variables.

The models are also used to illustrate how conditional relative chances and chance differences can change strongly, when the roles of a regressor variable and the response are exchanged, while odds-ratios and logit regression coefficients capture the unchanged

equal dependences given the remaining leaves. This problem occurs more generally but is convincingly demonstrated using this special binary family of distributions.

For two binary variables, in general, Pearson's correlation coefficient, ρ, is a multiple of the cross-product difference of the probabilities; see for instance equation (10) in Wermuth and Marchetti (2014). Only for symmetric binary variables, $\rho > 0$ reduces to the cross-sum difference in equation (5) and becomes a 1-1 function of the odds-ratio. The cross-sum difference of counts in equation (17), arises also as the nonparametric measure of dependence, studied by Blomqvist (1950) for continuous random variables: in the special case of symmetry in the observed 2×2 table that may result after median-dichotomizing the bivariate observations. Extensions of this measure and relations to copulas have been investigated by Schmid and Schmidt (2007) and Genest, Carabarin-Aguirre and Harvey (2013).

The one-parameter model considered here may be generalized in several ways. One possibility is to abandon the assumption of symmetry. For binary variables, this leads to the model studied for example in Allman et al. (2014). However even for this minimally extended model, it is much more complex to provide detailed insight into maximum-likelihood inference. In future work, we intend to study symmetric variables with more than two levels, concentric rings of binary variables with unequal spacings and maximization of the empirical likelihood functions.

References

Allman, E.S., Rhodes, J.A., Sturmfels, B. and Zwiernik, P. (2014). Tensors of nonnegative rank two. ArXiv:1305.0539 and *Linear Algebra and Applic.: Special Issue on Statistics.* To appear.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. B* **36**, 192–236.

Blomqvist, N. (1950). On a measure of dependence between two random variables. *Ann. Math. Stat.* **21**, 593–600.

Castelo, R. and Siebes. A. (2003). A characterization of moral transitive acyclic directed graph Markov models as labeled trees. *J. Stat. Plan. Inf.* **115**, 235–259.

Cox, D.R. and Wermuth, N. (1994). A note on the quadratic exponential binary distribution. *Biometrika* **81**, 403–406.

Darroch, J.N., Lauritzen, S.L. and Speed, T.P. (1980) Markov fields and log-linear models for contingency tables. *Ann. Statist.* **8**, 522–539.

Dempster, A.P. , Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B* **39**, 1–38.

Edwards, A. W. F. (1963). The measure of association in a 2 × 2 table. *J. Roy. Statist. Soc. A* **126**, 109–114.

Edwards, D. (1996). *Introduction to Graphical Modelling.* Springer, New York.

Fienberg, S.E. (2007). *The Analysis of Cross-classified Categorical Data*, 2nd ed. Springer, New York.

Fisher, R.A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philos. Trans. Roy. Soc. London Ser. A* **222**, 309–368.

Genest, C., Carabarín-Aguirre, A. and Harvey, F. (2013). Copula parameter estimation using Blomqvist's beta. *J. Soc. Franç. Statist.* **154**, 5–24.

Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis. *Measurement Prediction* **4**, 362–412.

Linzer, D.A. and Lewis J.B. (2011). poLCA: An R package for polytomous variable latent class analysis *J. Statist. Softw.* **42**.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo.

Perlman, M. and Wu, L (1999). Lattice conditional independence models for contingency tables with non-monotone missing data patterns. *J. Statist. Plan. Inference* **79**, 259–287.

Schmid, F. and Schmidt, R. (2007). Nonparametric inference on multivariate versions of Blomqvist's beta and related measures of tail dependence. *Metrika* **66**, 323–354.

Tanner, M.A. (1996). *Tools for Statistical Inference*, 3rd ed. Springer, New York.

Teugels, J.L. (1990). Some representations of the multivariate Bernoulli and binomial distributions. *J. Multiv. Analysis* **32**, 256-268.

Wermuth, N. (1987). Parametric collapsibility and the lack of moderating effects in contingency tables with a dichotomous response variable. *J. Roy. Statist. Soc. B*, **49**, 353–364.

Wermuth, N. and Lauritzen, S.L. (1983). Graphical and recursive models for contingency tables. *Biometrika*, **70**, 537–552.

Wermuth, N. and Marchetti, G.M. (2014). Star graphs induce tetrad correlations: for Gaussian as well as for binary variables. *Electr. J. Statist.* **8**, 253–273.

Wermuth, N., Marchetti, G.M. and Cox, D.R. (2009). Triangular systems for symmetric binary variables. *Electr. J. Statist.* **3**, 932–955. %vspace-3.3mm

Xie, X.C., Ma, Z.M. and Geng, Z. (2008). Some association measures and their collapsibility. *Statist. Sinica*. **18**, 1165–1183.

Zwiernik, P. and Smith, J.Q. (2011). Implicit inequality constraints in a binary tree model. *Electr. J. Statist.* **5**, 1276–1312.