Journal of Multivariate Analysis 130 (2014) 252–260

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Binary distributions of concentric rings

Nanny Wermuth [a,b,*], Giovanni M. Marchetti [c], Piotr Zwiernik [d]

[a] Department of Mathematical Sciences, Chalmers University of Technology, Sweden

[b] Department of Medical Psychology and Medical Sociology, Johannes Gutenberg-University, Germany

[c] Dipartimento di Statistica, Informatica, Applicazioni, "G. Parenti", University of Florence, Italy

[d] Department of Statistics, University of California, Berkeley, USA

ARTICLE INFO

Article history:
Received 8 December 2013
Available online 29 May 2014

AMS subject classifications:
62E10
62H17
62H20

Keywords:
Conditional independence
Graphical Markov models
Jointly symmetric distributions
Labeled trees
Latent class models
Phylogenetic trees
Star graphs
Symmetric variables

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 behavior of different measures of dependence and derive maximum likelihood estimates when all nodes are observed and when the inner node is hidden.

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 $1 \le \alpha < \infty$, to write with $\{0, 1\}$ coding, also known as baseline coding, for the joint p-dimensional distribution

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

* Corresponding author at: Department of Medical Psychology and Medical Sociology, Johannes Gutenberg-University, Germany.
E-mail addresses: wermuth@chalmers.se, nanny.wermuth@uni-mainz.de (N. Wermuth).

http://dx.doi.org/10.1016/j.jmva.2014.05.010

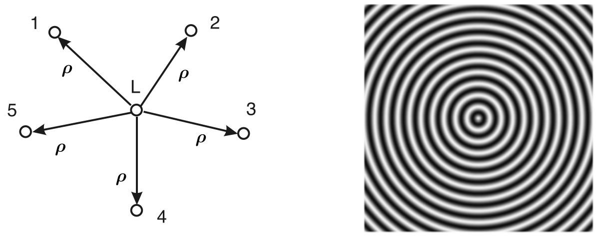

Fig. 1. Star graph with equal dependences of five leaves on one common root (left) and a graph of evenly-spaced concentric rings (right).

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

$$\mathrm{pr}(L = 1) = \mathrm{pr}(L = -1) = \frac{1}{2}, \qquad E(L) = 0, \qquad E(L^2) = 1.$$

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

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

ranges in $0 \leq \rho < 1$ 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 Eq. (1), hence it is excluded for some purposes. 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 Eq. (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 Eq. (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 [4], hence a lattice-conditional-independence model [15] and a directed-acyclic-graph model [20,14] or a Markov field for binary variables [6], an Ising model of ferromagnetism, a binary quadratic exponential distribution [2,5] and a triangular system of symmetric binary variables [22].

With L in Eq. (1) unobserved, the resulting model may be regarded as a simplest case for constructing phylogenetic trees; see [24,1] and the previous extensive literature in this area. Or, it can be viewed as a special latent-class model [12,13], 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 Fig. 1 shows such a star graph, having equal regression coefficients ρ when regressing each A_q on L, for $q = 1, \ldots, Q$.

For Gaussian and for binary distributions generated over star graphs as those in Fig. 1, the correlation matrices of the p variables are of identical form; see [21]. For $p = 5$, such correlation matrices are in Table 2, with '·' indicating a symmetric entry.

Table 2
Correlation matrix for $p = 5$; left: to Eq. (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 Fig. 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.