Palindromic Bernoulli distributions

Article *in* Electronic Journal of Statistics · October 2015

DOI: 10.1214/16-EJS1175 · Source: arXiv

2 authors:

Giovanni Maria Marchetti
University of Florence
77 PUBLICATIONS 709 CITATIONS

Nanny Wermuth
Chalmers University of Technology
123 PUBLICATIONS 4,662 CITATIONS

Palindromic Bernoulli distributions

Giovanni Marchetti

Dipartimento di Statistica, Informatica, Applicazioni "G. Parenti", Florence,

and **Nanny Wermuth**

Mathematical Statistics, Chalmers University of Technology, Gothenburg, Sweden and

Medical Psychology and Medical Sociology, Gutenberg–University, Mainz

Abstract

We introduce and study a subclass of joint Bernoulli distributions which has the palindromic property. For such distributions the vector of joint probabilities is unchanged when the order of the elements is reversed. We prove for binary variables that the palindromic property is equivalent to zero constraints on all odd-order interaction parameters, be it in parameterizations which are log-linear, linear or multivariate logistic. In particular, we derive the one-to-one parametric transformations for these three types of model specifications and give simple closed forms of maximum likelihood estimates. Some special cases and a case study are described.

Keywords: Central symmetry; Linear-in-probability models; Log-linear models, Multivariate logistic models; Median-dichotomization; Orthant probabilities; Odd-order interactions

1 Introduction

A sequence of characters, such as `QR*-TS`, becomes a palindromic sequence when the order of the characters is reversed and appended, here to give `QR*-TSST-*RQ`. The notion is used in somewhat modified forms, among others, in musicology, biology and linguistics. An example of a palindromic sentence which respects the spacings between words is 'step on no pets'.

Here, we adapt the term to Bernoulli distributions. For a single binary variable, the distribution is palindromic if it is uniform, that is if both levels occur with probability 1/2. For a Bernoulli distribution of d binary variables $A_1, \ldots, A_d$, having a probability mass function $p(a)$ with a in the set of all binary d-vectors, the distribution is palindromic if $p(a) = p(\sim a)$ for all a, where $\sim a$ is the complement of a; for instance, $\sim a = (0, 1, 0)$ for $a = (1, 0, 1)$.

With $\alpha, \beta, \gamma, \delta$ denoting probabilities, bivariate and trivariate palindromic Bernoulli distributions can be written, as in the following tables:

A_1 A_2:	0	1	sum
0	α	β	1/2
1	β	α	1/2
sum	1/2	1/2	1

A_3:	0	0	1	1	
A_1 A_2:	0	1	0	1	sum
0	α	γ	δ	β	1/2
1	β	δ	γ	α	1/2
sum	$\alpha+\beta$	$\gamma+\delta$	$\gamma+\delta$	$\alpha+\beta$	1

Continuous distributions may also be palindromic. For instance, a bivariate Gaussian distribution centred at zero is palindromic because its probabilities coincide in the first, positive quadrant and in the fourth, negative quadrant. Similarly, for the second and the third quadrant, these probabilities agree. For more than two variables, quadrants are replaced by orthants and a joint distribution is said to palindromic if their orthant probabilities repeat in reverse order. Examples are mean-centred Gaussian distributions, spherical distributions and, more generally, distributions with central symmetry; see Serfling (2006).

Edwards (2000, App. C) studied the properties of the dichotomized normal distribution and introduced the term multivariate symmetry. We change the terminology to stress the special structure of these distributions and to distinguish them from the class of *completely symmetric* tables, defined in the literature on contingency tables by Bhapkar & Darroch (1990). These satisfy the weaker condition $p(a) = p(\sigma(a))$, for any a and for any permutation σ of the indices.

The orthant probabilities of a joint Gaussian distribution give a nonhierarchical log-linear model in which all interactions involving an odd number of factors vanish, that is they are zero. In this paper, we study properties of palindromic Bernoulli distributions. In particular, we prove that the vanishing of all odd-order log-linear interactions is not only a necessary but also a sufficient condition. We show the same characterization for models linear-in-probabilities, Streitberg (1990), and for the multivariate logistic parametrization, Glonek & McCullagh (1995), and explain why palindromic Bernoulli distributions with Markov structure are in the regular exponential family.

2 Characterization in terms of interaction parameters

2.1 Notation

Let $A = (A_1, \dots, A_d)$ be a random vector with a *multivariate Bernoulli* distribution. Thus, A takes values $a = (a_1, \dots, a_d)$ in the set $\mathcal{I} = \{0, 1\}^d$ with probabilities

$$p(a) = \Pr(A_1 = a_1, \dots, A_p = a_p), \quad \textstyle\sum_{a \in \mathcal{I}} p(a) = 1.$$

For simplicity, we assume $p(a) > 0$ for all a. The probability distribution of A is determined by the $2^d \times 1$ vector π containing all the probabilities $p(a)$ and belonging to the $(2^d - 1)$-dimensional simplex. We list vectors a in a lexicographic order such that the first index in a runs fastest, then the second changes and the last index runs slowest. Cells of a corresponding contingency table are in vector $b \in \mathcal{I}$.

Given a subset $M \subseteq V$ of the variables, the marginal distribution of the variables A_v, for $v \in M$ has itself a joint Bernoulli distribution, in the same lexicographic order:

$$p_M(a_M) = \Pr(A_v = a_v, \text{ for all } v \in M).$$

We use three well-studied parameterizations for joint Bernoulli distributions, that is the log-linear, the linear and the multivariate logistic parameterizations and show how and why they differ even for palindromic Bernoulli distributions.

In general, a parameterization of A is a smooth one-to-one transformation, mapping π into a $2^p \times 1$ vector $\theta = G(\pi)$, say, whose entries θ_b, are called *interaction parameters*. To index interactions, it is useful to have a one-to-one mapping between the cells in b and subsets of $V = \{1, \dots, p\}$. For $p = 3$:

	Lexicographic order							
cells in b:	000	100	010	110	001	101	011	111
subset of V:	$\emptyset$	1	2	12	3	13	23	123
θ_b :	$\theta_\emptyset$	θ_1	θ_2	θ_{12}	θ_3	θ_{13}	θ_{23}	θ_{123}

The cardinality of the set b, denoted by $|b| = \sum_v b_v$, gives the number of ones in vector b. Depending on $|b|$ being odd or even, an interaction parameter θ_b is said to be of odd or even order. For instance, the even-order θ_{13} is a two-factor interaction of A_1 and A_3.

2.2 Log-linear parameters

Log-linear parameters are contrasts of log probabilities, that is linear combinations of $\log p(a)$, with weights adding to zero. The vector of the log-linear parameters is

$$\lambda = \mathcal{H}_d^{-1} \log \pi \tag{2.1}$$

where

$$\mathcal{H}_d = \underbrace{\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}}_{d}.$$

is a $2^d \times 2^d$ symmetric design matrix whose generic entry is $h_{ab} = (-1)^{a \cdot b}$ for $(a, b) \in \mathcal{I} \times \mathcal{I}$. Its inverse, the contrast matrix, is $\mathcal{H}_d^{-1} = 2^{-d}\mathcal{H}_d$. The special form of $\mathcal{H}_d$, chosen here, uses so-called *effect coding*; see Wermuth & Cox (1992). The individual interactions can be written as

$$\lambda_b = 2^{-d} \textstyle\sum_{a \in \mathcal{I}_V} (-1)^{a \cdot b} \log p(a), \tag{2.2}$$

where $a \cdot b = a_1 b_1 + \cdots + a_p b_p$ is the inner product of the two binary vectors a and b; see Haberman (1973, p. 619). In equation (2.2), the symbol b is interpreted for λ_b as a subset of V and in the expression $(-1)^{a \cdot b}$ as binary vector.

The inverse mappings from λ to π may be explicitly computed as

$$\pi = \exp(\mathcal{H}_d \lambda), \qquad p(a) = \exp\left[\textstyle\sum_{b \in \mathcal{I}} (-1)^{a \cdot b} \lambda_b\right]. \tag{2.3}$$

Bernoulli distributions with positive cell probabilities belong to the so-called regular exponential family with the vector λ containing the canonical parameters.

2.3 Linear interactions or moment parameters

In contrast to log-linear models, the linear-in-probability models, discussed for instance by Cox & Wermuth (1992, Appendix 2), and their interactions are based on moments. The vector $\xi = \mathcal{H}_p \pi$ is a *moment parameter vector* and the mapping between ξ and λ is one-to-one and differentiable; see Barndorff-Nielsen (1978, p. 121).

The moment vector ξ is proportional to the expected value of the sufficient statistics for λ. With y denoting the vector of the frequencies and by the symmetry of $\mathcal{H}_d$ of equation (2.1), this vector of sufficient statistics is $\mathcal{H}_d y$. The elements of ξ, called also *linear interactions*, are

$$\xi_b = \textstyle\sum_{a \in \mathcal{I}} (-1)^{a \cdot b}\, p(a) \tag{2.4}$$

and gives as inverse transformations

$$\pi = 2^{-d}\, \mathcal{H}_d\, \xi, \qquad p(a) = 2^{-d} \textstyle\sum_{b \in \mathcal{I}} (-1)^{a \cdot b} \xi_b. \tag{2.5}$$

For the transformed random variables $D_v = (-1)^{A_v}$, which take value 1 if $A_v = 0$ and -1 if $A_v = 1$, the individual interactions are

$$\xi_b = E\Big(\textstyle\prod_{v \in V} D_v^{b_v}\Big). \tag{2.6}$$

Because the element $(-1)^{a \cdot b}$ in equation (2.4) may be written with $d_v = (-1)^{a_v}$ as

$$(-1)^{a \cdot b} = (-1)^{a_1 b_1} \times \cdots \times (-1)^{a_p b_p} = \textstyle\prod_{v \in V} (d_v)^{b_v},$$

equation (2.4) gives the expected value of this product with respect to $p(a)$.

Equation (2.6) implies that each moment parameter, ξ_b, is a *marginal parameter*, defined in the marginal distribution of the random vector $(A_v)_{v \in b}$, while the log-linear parameter λ_b is defined in the joint distribution. Therefore there is, in general, no simple relation between the log-linear parameter λ_b^M, say, in the marginal distribution $p_M(a_M)$ and λ_b, the log-linear parameter in the joint distribution, but there are exceptions for instance as in Example 3.2. By contrast, the moment vector ξ_b^M defined in $p_M(a_M)$ coincides with ξ_b.

There is the following important result for regular exponential families by Barndorff-Nielsen (1978, pp. 121–122). For an arbitrary partition of the parameter vectors λ and ξ in two sub-vectors such that $\lambda = (\lambda_{\mathcal{A}}, \lambda_{\mathcal{B}})$ and $\xi = (\xi_{\mathcal{A}}, \xi_{\mathcal{B}})$, the distribution π is uniquely parameterized by the mixed vector $(\lambda_{\mathcal{A}}, \xi_{\mathcal{B}})$ or by $(\xi_{\mathcal{A}}, \lambda_{\mathcal{B}})$ and there is a diffeomorphism between this *mixed parameterization* and the log-linear parameter λ or the moment parameter ξ.

2.4 Multivariate logistic parameters

The *multivariate logistic parametrization*, introduced by Glonek & McCullagh (1995), is defined by the highest order log-linear parameters, considered here under effect coding, in each possible marginal distribution of A. The parameters are given by the vector $\eta = (\eta_b)_{b \subseteq V}$ where

$$\eta_b = \lambda_b^b, \qquad b \subseteq V. \tag{2.7}$$

Kauermann (1997) showed that the mapping $T : \lambda \mapsto \eta$ from the log-linear to the multivariate logistic parameters is a diffeomorphism by proving that T is a composition of smooth transformations between the canonical, the moment and the mixed parameters. More details are given below in subsection 2.5.

Let $\Lambda = \mathbb{R}^{2^p-1}$ be the parameter space for the log-linear parameters λ. Then the parameter space $E = T(\Lambda)$ for η is the image of λ under the transformation T. Explicit forms for the inverse function $T^{-1} : E \to \Lambda$ are known for $p = 1$ or $p = 2$ and in special cases, such as in Example 3.2. An algorithm provided by Qaqish & Ivanova (2006) detects simultaneously whether the vector η is compatible with a proper probability vector π.

2.5 Properties of the palindromic Bernoulli distributions

Palindromic Bernoulli distributions are closed under marginalization.

Proposition 2.1. *If $p(a)$ is a palindromic Bernoulli distribution then, for any subset M of the variables, the marginal distribution $p_M(a_M)$ is a palindromic.*

Proof. Define the partition $a = (a_N, a_M)$ and let $p_M(a_M) = \sum_{a_N \in \{0,1\}^{|N|}} p(a_N, a_M)$. Then if the distribution is palindromic, $p(a_N, a_M) = p(\sim a_N, \sim a_M)$ and

$$p_M(a_M) = \textstyle\sum_{\sim a_N \in \{0,1\}^{|N|}} p(\sim a_N, \sim a_M) = p_M(\sim a_M).$$

□

Next, we characterise the distribution by zero constraints on interactions.

Proposition 2.2. *A Bernoulli distribution is palindromic if and only if, with $\theta_b = \xi_b$ or $\theta_b = \lambda_b$, all odd-order linear or log-linear interactions vanish, that is if and only if*

$$\theta_b = 0, \qquad \text{for all } b \subseteq V \text{ with } |b| \text{ odd}.$$

Proof. 1) (If A is palindromic then all odd-order $\xi_b = 0$.) Any linear interaction can be written as

$$\xi_b = \textstyle\sum_{a \in \mathcal{I}_1} (-1)^{a \cdot b} p(a) + \sum_{a \in \mathcal{I}_1} (-1)^{(\sim a) \cdot b} p(\sim a), \tag{2.8}$$

where $\mathcal{I}_1$ denotes the subset of cells having a one as first element. Thus $\mathcal{I}_1$ contains half of the cells. If the distribution is palindromic, $p(\sim a) = p(a)$ and $(-1)^{(\sim a) \cdot b)} = (-1)^{|b|}(-1)^{a \cdot b}$. Thus,

$$\xi_b = \textstyle\sum_{a \in \mathcal{I}_1} (-1)^{a \cdot b} p(a) + (-1)^{|b|} \sum_{a \in \mathcal{I}_1} (-1)^{a \cdot b} p(a). \tag{2.9}$$

When $|b|$ is odd then $(-1)^{|b|} = -1$ and $\xi_b = 0$; see also Edwards (2000, App. C).

2) (If all odd-order $\xi_b = 0$, then $p(\sim a) = p(a)$.) If all odd-order interactions vanish, then

$$p(a) = \frac{1}{2^d} \textstyle\sum_{b \in \mathcal{I}_{\text{even}}} (-1)^{a \cdot b} \xi_b, \tag{2.10}$$

where $\mathcal{I}_{\text{even}}$ is the subset of the cells b such that $|b|$ is even. Thus,

$$p(\sim a) = \frac{1}{2^p} \textstyle\sum_{b \in \mathcal{I}_{\text{even}}} (-1)^{(\sim a) \cdot b} \xi_b = \frac{1}{2^p} \sum_{b \in \mathcal{I}_{\text{even}}} (-1)^{|b|} (-1)^{a \cdot b} \xi_b = p(a), \tag{2.11}$$

because $|b|$ is even. So the distribution is palindromic.

3) The same arguments apply for the log-linear parameterization. The distribution is palindromic if and only if $\log p(a) = \log p(\sim a)$ for all a. Therefore, using equation (2.2), and the previous lines of reasoning, $\lambda_b = 0$ whenever $|b|$ is odd. Conversely, if all odd-order log-linear parameters λ_b vanish, then from

$$\log p(a) = \textstyle\sum_{b \in \mathcal{I}} (-1)^{a \cdot b} \lambda_b$$

we get $\log p(a) = \log p(\sim a)$ and the distribution is palindromic. □

By equation (2.6) and Proposition 2.2, the joint distribution of A is palindromic if and only if all the odd-order moments of $D = (-1)^A$ are zero. Also as the palindromic property is characterized by linear constraints on the canonical pararameters λ we have the following result.

Corollary 2.3. *Palindromic Bernoulli distributions are a regular exponential family.*

We show next a similar characterization for the multivariate logistic parametrization.

Proposition 2.4. *A Bernoulli distribution is palindromic if and only if all odd-order multivariate logistic parameters vanish, that is if and only if*

$$\eta_b = 0, \qquad \textit{for all } b \subseteq V \textit{ with } |b| \textit{ odd.}$$

The following proof uses a transformation $T \colon \lambda \mapsto \eta$ introduced by Kauermann (1997, p. 265). The composition of smooth one-to-one transformations T_M gives T, for each nonempty subset $M \subseteq V$. The functions T_M operate on parameter transformations between the canonical and the moment parametrizations, as follows. If $M = V$,

$$T_M(\lambda_{\mathcal{P}(V)\setminus V}, \lambda_V) = (\xi_{\mathcal{P}(V)\setminus V}, \lambda_V).$$

If $M \subset V, |M| \neq 1$:

$$T_M(\dots, \xi_{\mathcal{P}(V)\setminus V}, \xi_M, \dots) = (\dots, \xi_{\mathcal{P}(M)\setminus M}, \eta_M, \dots)$$

and the remaining parameters, which are not listed, are left unchanged. Finally, if $|M| = 1$,

$$T_M(\xi_M, \dots) = (\eta_M, \dots).$$

For instance, to clarify, to get $\eta = T(\lambda)$ for three variables we define

$$T(\lambda) = T_1 \circ T_2 \circ T_3 \circ T_{12} \circ T_{13} \circ T_{23} \circ T_{123}(\eta)$$

and Table 1 gives the details of the required transformations T_M.

Proof. Let $\text{odd} = \{b \in \{0,1\}^{|V|} : |b| \text{ odd}\}$ be the subset of all odd-order interactions. Then, below we show that $\eta_{\text{odd}} = 0$ if and only if $\boldsymbol{\lambda}_{\text{odd}} = 0$.

From Proposition 2.2 we know that a binary distribution is palindromic if and only if $\lambda_{\text{odd}} = 0$. If π is palindromic then all the marginal distributions $p_b(a_b), b \subseteq V$ are palindromic and thus in each of them, such that $|b|$ is odd, $\lambda_b^b = \eta_b = 0$. Thus, $\eta_{\text{odd}} = 0$. Let $\text{even} = \mathcal{P}(V) \setminus \text{odd}$. Then

$$T(\lambda_{\text{even}}, \lambda_{\text{odd}} = 0) = (\eta_{\text{even}}, \eta_{\text{odd}} = 0). \tag{2.12}$$

Table 1: *The sequence of transformations required to obtain the multivariate logistic parameter η from the log-linear parameter λ.*

Transformation	Parameters							Intermediate result
$T_{123}(\lambda)$	ξ_1	ξ_2	ξ_{12}	ξ_3	ξ_{13}	ξ_{23}	η_{123}	$\theta^{(1)}$
$T_{23}(\theta^{(1)})$	ξ_1	ξ_2	ξ_{12}	ξ_3	ξ_{13}	η_{23}	η_{123}	$\theta^{(2)}$
$T_{13}(\theta^{(2)})$	ξ_1	ξ_2	ξ_{12}	ξ_3	η_{13}	η_{23}	η_{123}	$\theta^{(3)}$
$T_{12}(\theta^{(3)})$	ξ_1	ξ_2	η_{12}	ξ_3	η_{13}	η_{23}	η_{123}	$\theta^{(4)}$
$T_3(\theta^{(4)})$	ξ_1	ξ_2	η_{12}	η_3	η_{13}	η_{23}	η_{123}	$\theta^{(5)}$
$T_2(\theta^{(5)})$	ξ_1	η_2	η_{12}	η_3	η_{13}	η_{23}	η_{123}	$\theta^{(6)}$
$T_1(\theta^{(6)})$	η_1	η_2	η_{12}	η_3	η_{13}	η_{23}	η_{123}	η

Conversely, if $\eta_{\text{odd}} = 0$, let η_{even} be arbitrarily chosen such as $(\eta_{\text{even}}, \eta_{\text{odd}}) = 0) \in E_0 \subset E$ (the parameter space of the ηs). As E_0 is connected we can directly use equation (2.12) and the smoothness of the inverse transformation T^{-1} to get

$$T^{-1}(\eta_{\text{even}}, \eta_{\text{odd}} = 0) = (\lambda_{\text{even}}, \lambda_{\text{odd}} = 0),$$

and thus the distribution π is palindromic. □

Table 2 illustrates the different parameters with a 2^3 table.

Table 2: *Illustration of the the different parameters with a 2^3 table; constant terms omitted.*

	Lexicographic order							
cells b :	000	100	010	110	001	101	011	111
$80\,\pi$:	15	9	1	15	15	1	9	15
subsets of V :	$\emptyset$	1	2	12	3	13	23	123
ξ :	–	0	0	1/2	0	−1/5	1/5	0
λ :	–	0	0	log(5)/2	0	−1/5	1/5	0
η :	–	0	0	log(3)/2	0	− log(3)/2	log(3)/2	0

Next, we state a result connected with binary probability distributions *generated by a linear triangular system*, as studied in Wermuth, Marchetti & Cox (2009). Their joint probabilities may be defined by the recursive factorization

$$\Pr(A_1 = a_1, \ldots, A_d = a_d) = \Pr(A_1 = a_1)\prod_{s=2}^{d}\Pr(A_s = a_s \mid A_1 = a_1, \ldots, A_{s-1} = a_{s-1})$$

with uniform margins. With $\beta_{s,j}$ denoting linear regression coefficients,

$$\Pr(A_s = a_s \mid A_1 = a_1, \ldots, A_{s-1} = a_{s-1}) = \tfrac{1}{2}(1 + \textstyle\sum_{1=s-1}^{d} \beta_{s,j}(-1)^{a_s + a_j}). \quad (2.13)$$

The conditional expected values of A_s given variables $A_{[s-1]} = (A_1, \ldots, A_{s-1})$ are linear regressions with only main effects and no constant term. For these distributions, all the even-order linear interactions are known functions of the marginal correlations. Here, we prove in Appendix 1 the following result and get back to such systems later.

Proposition 2.5. *If a binary probability distribution is generated by a linear triangular system, then it is palindromic.*

3 Some special cases

3.1 Median dichotomization

Let (X_1, X_2) have a joint distribution function F_{12} with marginal distributions functions F_1 and F_2. Let further $U_1 = F_1(X_1)$ and $U_2 = F_2(X_2)$ be the probability integral transforms of X_1 and X_2, so that U_1 and U_2 are uniform. Also let $\widetilde{X}_j$ be the medians of X_j, $j = 1, 2$. Consider now the median dichotomized variables,

$$A_1 = \mathbb{I}[U_1 > \tfrac{1}{2}], \quad A_2 = \mathbb{I}[U_2 > \tfrac{1}{2}] \tag{3.1}$$

where $\mathbb{I}[\cdot]$ is the indicator function. Then, the joint distribution of A_1 and A_2 is a bivariate palindromic Bernoulli distribution, as given in Section 1, with $\alpha = P(U_1 > \frac{1}{2}, U_2 > \frac{1}{2})$.

The variables $D_1 = (-1)^{A_1}$ and $D_2 = (-1)^{A_2}$, taking values $1, -1$, have mean zero and unit variance, so that $\xi_{12} = E(D_1 D_2)$, the correlation coefficient between D_1 and D_2, becomes the *cross-sum difference* of the joint probabilities

$$\xi_{12} = 2\alpha - 2\beta = 4\alpha - 1. \tag{3.2}$$

Thus the correlation between two binary variables, which is a multiple of the cross-product difference, coincides in a bivariate palindromic Bernoulli distribution with the cross-sum difference. This was not noted, when ξ_{12} was proposed as a measure of dependence between any two random variables X_1 and X_2 by Blomqvist (1950):

$$\begin{aligned} 4\alpha - 1 &= \Pr\{(X_1 - \widetilde{X}_1)(X_2 - \widetilde{X}_2) > 0\} - \Pr\{(X_1 - \widetilde{X}_1)(X_2 - \widetilde{X}_2) < 0\} \\ &= 2\Pr(U_1 \le \tfrac{1}{2}, U_2 \le \tfrac{1}{2}) + 2\Pr(U_1 > \tfrac{1}{2}, U_2 > \tfrac{1}{2}) - 1. \end{aligned}$$

Remark 3.1. The probability α may be interpreted as the copula $C(\frac{1}{2}, \frac{1}{2})$ of the random vector (X_1, X_2), where the function $C(u, v) = \Pr(U_1 \le u, U_2 \le v))$, $0 \le u \le 1$, $0 \le v \le 1$.

By using the linear interaction expansion of equation (2.5), the distribution of D_1, D_2 is

$$P(D_1 = i, D_2 = j) = \tfrac{1}{4}(1 + \xi_{12} ij), \qquad i, j = 1, -1. \tag{3.3}$$

After median-dichotomizing $d > 2$ continuous variables, the resulting binary variables $A_v, v = 1, \ldots, d$ are still marginally uniform, but their joint distribution is palindromic only for centrally symmetric variables, that is when $X_v - \widetilde{X}_v$ has the same distribution as $-(X_v - \widetilde{X}_v)$ for each $X_v, v = 1, \dots d$.

With $d = 3$, the joint distribution of the median-dichotomized variables is palindromic with parameters α, β, γ and δ, as given in section 1. Their marginal correlations are

$$\xi_{12} = 4\alpha + 4\delta - 1, \quad \xi_{13} = 4\alpha + 4\gamma - 1, \quad \xi_{23} = 4\alpha + 4\beta - 1$$

and the joint probability distribution is, with $i, j, k = 1, -1$.

$$P(D_1 = i, D_2 = j, D_3 = k) = \tfrac{1}{8}(1 + \xi_{12}ij + \xi_{13}ik + \xi_{23}jk).$$

Example 3.2. The following example gives the orthant probabilities of a trivariate, mean-centred Gaussian distribution having equal correlations: $-1/2 < \rho < 1$. The joint probability vector of the median-dichotomized variables is:

$$8\pi = (1 + 3\xi, \quad 1 - \xi, \quad 1 - \xi, \quad 1 - \xi, \quad 1 - \xi, \quad 1 - \xi, \quad 1 - \xi, \quad 1 + 3\xi)$$

and the explicit transformations between the three types of parameters result with

$$\xi = \frac{2}{\pi} \arcsin \rho, \quad \lambda = \frac{1}{4} \log \frac{1 + 3\xi}{1 - \xi}, \quad \eta = \operatorname{atanh} \xi. \tag{3.4}$$

The arcsin transformation is due to Sheppard (1898) and the obtained distribution is a *concentric ring model*; see Wermuth, Marchetti & Zwiernik (2014).

Proposition 3.3. *If X has a d-variate Gaussian distribution with mean zero and correlation matrix $R = [\rho_{st}]$ for $s, t = 1, \dots, d$ and A is the binary random vector obtained by median dichotomizing X, with linear interaction parameters ξ_b, then R can be reconstructed from the correlation matrix $R_A = [\xi_{st}]$ between the binary variables A by*

$$\rho_{st} = \sin\{(\pi/2)\xi_{st}\}, \quad s, t = 1, \dots, d.$$

The proof results by inverting the arcsin transformation of the quadrant probability

$$\xi_{st} = 4 \Pr(X_s \le 0, X_t \le 0) - 1 = 2\pi^{-1} \arcsin \rho_{st}.$$

As a consequence one may to reconstruct the original correlations from the palindromic Bernoulli distribution derived via the orthant probabilities.

3.2 Maximum likelihood estimation

For a palindromic Bernoulli distribution, given a random sample of size n, one has as counts, that is as observed cell frequencies: $n(a)$, $a \in \mathcal{I}$. The likelihood is

$$\textstyle\prod_{a \in \mathcal{I}} p(a)^{n(a)} = \prod_{a \in \mathcal{I}_0} p(a)^{n(a)} p(\sim a)^{n(\sim a)} = \prod_{a \in \mathcal{I}_0} p(a)^{n(a) + n(\sim a)} \tag{3.5}$$

where $\mathcal{I}_0$ is the set of half of the cells a such that $a_1 = 0$. The sufficient statistics are thus the set of the 2^{d-1} frequencies $n(a) + n(\sim a)$, obtained by summing each cell and its complement image. The maximum likelihood estimate of a cell probability (or of a cell count) is the average of the two proportions (or of counts):

$$\widehat{p}(a) = \{n(a) + n(\sim a)\}/(2n), \quad \widehat{n}(a) = \{n(a) + n(\sim a)\}/2 \tag{3.6}$$

For palindromic Bernoulli distributions, Wilks' likelihood ratio test statistic is

$$w = 2 \textstyle\sum_{a \in \mathcal{I}} n(a) \log \left(\dfrac{2\, n(a)}{n(a) + n(\sim a)} \right). \tag{3.7}$$

It has an asymptotic χ^2 distribution with 2^{d-1} degrees of freedom; see Edwards (2000, App. C). The maximum likelihood estimates of the linear interaction parameters are

$$\widehat{\xi}_b = \begin{cases} 0 & \text{if } |b| \text{ odd,} \\ \sum_{a \in \mathcal{I}} (-1)^{a \cdot b}\, n(a)/n. & \text{if } |b| \text{ even.} \end{cases}$$

Thus, the estimated distribution matches the observed to the fitted moments and the estimator $\widehat{\xi}_b$ for b of even cardinality contains just marginal correlations, computed from the fitted table. For $|b| = 2$, the estimated marginal correlation, $\widehat{\xi}_{12}$ coincides with the correlation coefficient in the fitted table $\widehat{p}(a)$, hence results as a cross-sum difference of the counts

$$\widehat{\xi}_{12} = (n_{00} + n_{11}) - (n_{01} + n_{10}). \tag{3.8}$$

Since the log-linear and the multivariate logistic parameters are in a one-to-one relation to the linear interactions, the maximum likelihood estimates of their parameter vectors, result by the same transformations that hold for the parameters; see Fisher (1922). For the special transformations that apply here, see equations (2.3), (2.5).

In the following, we speak of maximum likelihood estimates simply as 'estimates'. Estimates may simplify further, when the distribution satisfies independence constraints in such a way that they lead to a graphical Markov model; see, for an overview of these models, Wermuth (2015).

Example 3.4 (A Markov chain)**.** Let A_1, A_2 and A_3 be three binary random variables where A_1 and A_3 are conditionally independent given A_2, so that the probabilities satisfy

$$p_{ijk} = p_{+j+}^{-1}\, p_{ij+}\, p_{+jk} \text{ for } i, j, k = \pm 1\,.$$

Its undirected graph, called a concentration graph, 1 —— 2 —— 3, has a missing edge for nodes 1 and 3, representing A_1 and A_2, and it is a simplest type of a graphical Markov model, a Markov chain in 3 variables.

The log-linear parameters are constrained by $\lambda_{13} = \lambda_{123} = 0$ for the conditional independence of pair $(1, 3)$. If, in addition, the distribution is palindromic, the odd-order parameters are zero so that also $\lambda_1 = \lambda_2 = \lambda_3 = 0$. In general Bernoulli distributions, the minimal sufficient statistics are the observed counts corresponding to the cliques of the graph, here these are just the node pairs (1,2) and (2,3). However for a palindromic Bernoulli distribution, the minimal sufficient statistics are the estimated counts $\widehat{n}_{ij+}$ and $\widehat{n}_{+jk}$ for margins $(1,2)$ and $(2,3)$, defined as in equation (3.6) so that

$$\widehat{p}_{ijk} = n^{-2}\,\widehat{n}_{ij+}\widehat{n}_{+jk} \text{ for } i,j,k = \pm 1.$$

This example illustrates how independence constraints, conditionally given all remaining variables, simply add to the linear constraints on canonical parameters of a palindromic Bernoulli distribution. For each involved variable pair, the two-factor and all its higher-order log-linear interactions have to vanish.

When a joint Bernoulli distribution has no higher than two-factor interactions, it is called in the literature either an Ising model or a binary quadratic-exponential distribution. A palindromic Ising distribution is therefore defined by having exclusively two-factor log-linear interactions, in addition to a constant term. When the model is decomposable, since its concentration graph is chordal, see e.g. Wermuth (2015), it can be generated by a linear triangular system; see end of Section 2. We get a further result after noting that the maximal complete subsets of a concentration graph are its cliques and that a concentration graph can be constructed by starting from a complete undirected graph, in which each pair is connected by a line, and then deleting the edge for each pair that is to be conditionally independent given all remaining variables.

Proposition 3.5. *If the concentration graph of a palindromic Bernoulli distribution has maximal clique size three, then it is an Ising model.*

4 A case study

For a sample of grades obtained at the University of Florence, we aim at predicting grades in Physics in terms of given grades in Algebra, Analysis and Geometry. The passing grades range in each subject from 18 to 31. We use sums of grades over exams in three successive years and have data for $n = 78$ students who reached in each of the subjects a sum of at least 60 points. Instructors expect positive correlations for each pair of these grades and no sign reversal for the correlations at fixed level combinations of the other variables. The data are in Appendix 2.

The four summed grades are closely bell-shaped, each of their scatter plots shows a nearly elliptic form as well as the plots of residual pairs obtained after linear least squares regression of each grade on the other three. There is also no evidence for nonlinear relations in the probability plots of Cox & Wermuth (1994). Thus, there is substantive and empirical support for assuming a joint Gaussian distribution.

After replacing for pairs (1,4) and (2,4) the observed correlations by $\widehat{r}_{14} = r_{13}r_{34}$, $\widehat{r}_{24} = r_{23}r_{34}$, in Table 3 we have the estimate of the correlation matrix, which has zeros for pairs (1,4) and (2,4) in its inverse, in its concentration matrix; see e.g. Wermuth, Marchetti & Cox (2009), equation (2.8).

Wilks' likelihood-ratio test statistic on 2 degrees of freedom shows with a value of $w = 2.8$ a good fit to the model with generating sets $\{\{1,2,3\},\{3,4\}\}$. This implies conditional independence of

Table 3: *For four fields and 78 students, observed marginal correlations, r_{ij} (below the diagonal), concentrations on the diagonal and partial correlations, $r_{ij.kl}$ (above the diagonal).*

	Analysis	Algebra	Geometry	Physics
1:=Analysis	2.64	0.27	0.34	0.17
2:=Algebra	0.72	3.03	0.51	0.04
3:=Geometry	0.76	0.80	4.07	0.38
4:=Physics	0.62	0.60	0.71	2.09

the grade in Physics from those in Analysis and Algebra given the grade in Geometry. This follows directly, for instance, with the corresponding concentration graph, on the left of Fig. 4. It has the cliques $\{1,2,3\}$ and $\{3,4\}$ for which node 3 separates node 4 from nodes 1,2 since to reach nodes 1,2 from node 4, one has to pass via node 3.

Similarly, after replacing the marginal correlations for pairs $(1,2)$, $(1,3)$ and $(2,3)$ by their average $\widehat{r} = 0.76$, we have for the submatrix of (1,2,3) the conventional estimate of an equicorrelation matrix; see Olkin and Pratt (1958, Section 3). This is here well-fitting since $w = 3.4$ on 2 degrees of freedom. The grade in Physics, correlates with this sum score as 0.706, even slightly less than with the grade in Geometry alone, where $r_{34} = 0.709$. This is plausible in view of the well-fitting Markov structure.

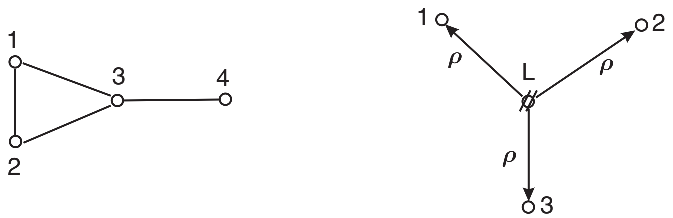

Figure 1: Left: the well-fitting concentration graph for the Florence grades; right: a possible generating graph for grades $1, 2, 3$.

A possible generating graph for the Gaussian equicorrelation matrix is the star graph displayed on the right of Fig. 4. In it mathematical ability is represented by the unobserved inner node, L, and the three grades are the outer nodes of the graph, shown as responses to L by arrows starting at L and pointing to the uncoupled nodes 1,2,3; each arrow has assigned to it the same positive correlation ρ. After marginalizing over L, each outer pair is correlated like ρ^2. We shall see next how well these results are reflected in the dichotomized data.

After median-dichotomizing the grades with jittering, we generate precisely uniform binary variables, the marginal distributions of which differ only little from those obtained by simple median-dichotomizing. One obtains the estimate of a palindromic contingency table in closed form using equation (3.6) and as we shall see, the same well-fitting concentration graph as on the left of Fig. 4.

The observed contingency table is given next, together with, in the additional rows and in the following order, the estimates of palindromic counts, and of the counts after imposing, in addition, conditional independences for the model of Fig. 4 on the left, and the estimates of the corresponding log-linear interactions.

Table 4: *Cells $ijkl$, levels of interactions, counts n_{ijkl}, MLEs of palindromic counts, of the palindromic concentration graph counts and of the log-linear interactions under the latter model*

0000	1000	0100	1100	0010	1010	0110	1110	0001	1001	0101	1101	0011	1011	0110	1111
∅	1	2	12	3	13	23	123	4	14	24	124	34	134	234	1234
22	3	3	0	1	0	1	9	6	2	2	1	3	2	1	22
22	2	2.5	1.5	1	1	1.5	7.5	7.5	1.5	1.0	1.0	1.5	2.5	2.0	22.0
21.2	2.5	2.5	1.8	0.7	1.0	1.0	8.3	8.3	1.0	1.0	0.7	1.8	2.5	2.5	21.2
0.90	0	0	0.45	0	0.62	0.62	0	0	0	0	0	0.47	0	0	0

The palindromic concentration graph model fits well, with $w = 10.3$ on 11 degrees of freedom. This decomposes into $w = 9.1$ on 9 degrees of freedom for the saturated palindromic model and $w = 1.2$ on 2 degrees of freedom for the additional independence constraints. Values of the studentized log-linear interactions are 2.5, 3.5, 3.5 and 3.7 for λ_{12}, λ_{13}, λ_{23} and λ_{34}, respectively. Thus, the same independences as for the underlying joint Gaussian distribution fit also the median-dichotomized data and further simplifications are not compatible given the sizes of the remaining studentized interactions. The partial correlations implied by the well-fitting palindromic Markov structure has also zeros for pairs $(1, 4)$. This illustrates a result by Loh & Wainwright (2013): if the largest separating set of a concentration graph contains a single node, then the conditional independences show also as zero partial correlations for the independent pairs.

The sum score of the median-dichotomized grades 1,2,3 leads as in the underlying Gaussian distribution not to an improved prediction of grades in Physics. To our starting question, we get two summarizing answers. Given a grade below the median in Geometry, one predicts that 72% of these students will have a grade below the median in Physics and, similarly, given a grade above the median in Geometry, one predicts that 72% will have a grade above the median in Physics.

5 Discussion

We say that centrally-symmetric Bernoulli distributions are palindromic since their probabilities, at the fixed level of one of the variables repeat in reverse order for the second level of this variable and thereby mimic palindromic sequences of characters as introduced in linguistics.

A palindromic Bernoulli distribution is characterized by the vanishing of all odd-order log-linear interactions. Hence, such zero constraints lead to a non-hierarchical, log-linear model which give centrally symmetric probabilities. Until now, it was only known that in centrally-symmetric Bernoulli distributions, all odd-order log-linear parameters vanish; see Edwards (2000, App. C). With these linear constraints, distributions result which are in the regular exponential family.

Palindromic Bernoulli distributions may also be parameterized with all odd-order interactions vanishing in a linear-in-probability model and in a multivariate-logistic model. The parameters in the three types of model are in one-to-one relations; see Section 2. These relations are now available in closed form for the linear and the log-linear formulations, while in general, iterative procedures are needed when the multivariate logistic formulation is involved. In any case, equivalent parameterizations assure that the maximum-likelihood estimates of the parameters are in the same one-to-one relation; see Fisher (1922).

It is remarkable that a palindromic Bernoulli distribution can be expressed precisely as a log-linear and as a linear model, since log-linear parameters use the notion of multiplicative interactions and the linear-in-probability models are based instead on the notion of additive interactions as discussed, for instance by Darroch & Speed (1983).

The log-linear parameterization shows that positive palindromic Bernoulli distributions are in the regular exponential family with and without additional independence constraints in its concentration graph. A palindromic Ising model may have only log-linear two-factor interactions as non-vanishing canonical parameters, while in their linear-model formulations higher-order interactions may be present. The palindromic property is preserved under marginalizing over any subset of the variables; see Proposition 2.1, even though one may no longer have an Ising model after marginalizing over some of the variables.

Another property is important for applications. In palindromic Bernoulli distributions, many other measures of dependence of a variable pair are one-to-one functions of the odds-ratio; in particular the relative risk, used mainly in epidemiology, and the risk difference, employed almost exclusively in the literature on causal modelling. Only if a measure of dependence is a function of the odds-ratio, it varies independently of its margins; see Edwards (1963) and only then, measures of bivariate dependence become directly comparable under different sampling schemes, for instance when the overall count is fixed as in a cross-sectional study or one of the margins is fixed as in a prospective study or the other margin is fixed as in a retrospective study.

In the future, it is desirable to investigate in more detail the relations between the palindromic and the totally positive distributions with additional conditional independence constraints. We expect also, that with a direct extension of the palindromic property to discrete variables of more levels, similar attractive properties can be obtained as for the palindromic Bernoulli distribution.

Appendix 1. Proof of Proposition 2.5

Proof. Our proof of Proposition 2.5 is by induction. We know that A_1 has a palindromic distribution. For $s = 2, \ldots, d$ we assume that the random vector $A_{[s-1]} = (A_1, \ldots, A_{s-1})$ has a palindromic distribution, and then we show that the distribution of $A_{[s]} = (A_1, \ldots, A_s)$ is palindromic. Let $\mathcal{I}_{\text{even}}$ denote the subset of $\{0, 1\}^s$ with even order and split it in two parts

$$\mathcal{I}_0 = \{a \in \mathcal{I}_{\text{even}} : a_s = 0\},\ \mathcal{I}_1 = \{a \in \mathcal{I}_{\text{even}} : a_s = 1\}.$$

We then start from the identity

$$\Pr(A_{[s]} = a_{[s]}) = \Pr(A_{[s-1]} = a_{[s-1]})\Pr(A_s = a_s \mid A_{[s-1]} = a_{[s-1]})$$

and after substituting equations (2.5) and (2.13) and taking into account that by assumption $A_{[s-1]}$ has a palindromic distribution and thus $\xi_b = 0$ for all $b \in \mathcal{I}_1$, we have

$$\Pr(A_{[s]} = a) = 2^{-s}{\textstyle\sum_{b\in\mathcal{I}_0}}\xi_b(-1)^{a\cdot b} \cdot \{1 + {\textstyle\sum_{j=1}^{s-1}}\beta_{s,j}(-1)^{a\cdot e_{s,j}}\}$$

where $e_{s,j}$ is a binary vector of dimension s with ones exactly in positions s and j. After multiplying and collecting terms we get with

$$\xi_b = {\textstyle\sum_{v\in I_0\colon v\triangle\{s,j\}=b}}\,\xi_v\beta_{s,j}, \qquad \text{for } b \in \mathcal{I}_1, \tag{5.1}$$

$$\Pr(A_{[s]} = a) = 2^{-s}\Big({\textstyle\sum_{b\in\mathcal{I}_0}}\xi_b(-1)^{a\cdot b} + {\textstyle\sum_{b\in\mathcal{I}_1}}\xi_b(-1)^{a\cdot b}\Big),$$

where $\triangle$ denotes the symmetric difference of sets. Therefore $A_{[s]}$ has a linear parameterization with exclusively even order interactions and hence is palindromic. Thus, by induction, the distribution of $A_{[d]} = A$ is palindromic. From the recursive equation (5.1), each linear interaction is a linear function of the regression parameters $\beta_{s,j}$. □

Appendix 2. The data for the case study

The columns of the following table contain sums of grades of three exams in four subjects for $n = 78$ mathematics students at the University of Florence.

Table 5: Summed grades over 3 exams in the order: Analysis, Algebra, Geometry and Physics.

78	78	74	80	88	77	79	85	82	82	74	89	85	77	93	85	79	85	74	69	78	88	67	92	85	69
76	75	71	77	81	79	77	90	79	72	62	90	75	83	92	88	80	88	68	80	75	88	70	89	88	75
82	81	74	71	85	74	81	83	73	73	71	86	84	84	93	82	78	90	70	79	71	89	68	91	91	62
85	77	80	80	79	80	75	82	71	71	72	87	82	69	90	75	75	82	70	78	72	77	69	93	87	68
79	92	76	88	73	91	76	71	65	74	80	71	78	77	70	83	89	72	82	77	91	92	75	90	90	93
78	92	78	87	68	85	78	79	68	76	89	74	81	74	68	89	81	76	81	74	92	92	69	79	82	93
71	92	84	88	64	83	82	69	71	75	80	71	85	69	67	88	83	75	83	82	93	92	72	90	89	93
79	90	86	78	69	75	82	71	63	72	78	74	81	67	66	72	82	75	79	76	92	87	75	79	78	89
92	87	81	82	76	86	92	87	79	91	88	90	90	92	89	83	77	69	89	92	86	76	68	79	76	88
93	83	69	70	75	71	80	70	70	77	88	92	85	92	84	83	82	74	83	92	74	71	62	68	66	89
93	87	74	67	80	69	87	77	69	92	83	91	82	91	86	83	80	83	83	90	78	71	65	74	83	91
89	77	81	79	84	72	80	81	70	79	77	72	88	81	86	81	78	76	77	79	73	69	69	72	80	85

Acknowledgement

We thank G. Bulgarelli at the University of Florence for the data and G. Ottaviani for helpful discussions. For computations we used Matlab®, 2015, for exact median-dichotomizing the Matlab function 'jitter' is by R. Cotton who adapted the R version by W. Stahel and M. Maechler, ETH Zurich.

References

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families in Statistical Theory.* New York: Wiley and Sons.

Bhapkar, V. P. & Darroch, J. N. (1990). Marginal symmetry and quasi symmetry of general order. *J. Multiv. Analysis* **34**, 173–184.

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

Cox, D. R. & Wermuth, N. (1992). Response models for mixed binary and quantitative variables. *Biometrika* **79**, 441–461.

Cox, D. R. & Wermuth, N. (1994). Tests of linearity, multivariate normality and adequacy of linear scores. *Applied Statistics, J. Roy. Statist. Soc. C*, **43**, 347–355.

Darroch, J. N., & Speed, T. P (1983). Additive and multiplicative models and interactions. *Ann. Statist.* **(**11), 724–738.

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

Edwards, D. (2000). *Introduction to graphical modelling.* New York: Springer.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A* **222**, 309–368.

Glonek, G. F. V. & McCullagh, P. (1995). Multivariate Logistic Models. *J. Roy. Statist. Soc. Ser. B* **57**, 533–546.

Haberman, S. J. (1973). Log-linear models for frequency data: Sufficient statistics and likelihood equations. *Ann. Statist.* **1**, 617–632.

Loh, P. & Wainwright, M. J. (1973). Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *Ann. Statist.* **41**, 3022–3049.

Kauermann, G. (1997). A note on multivariate logistic models for contingency tables. *Aust. J. Stat.* **39**, 261–276.

Olkin, I. and Pratt, J.W. (1958). Unbiased estimation of certain correlation coefficients. *Ann. Math. Statist.* **29**, 201–211.

Qaqish, B. F. & Ivanova, A. (2006). Multivariate logistic models. *Biometrika* **93**, 1011–1017.

Serfling, R. J. (2006). Multivariate symmetry and asymmetry. In: Kotz, S., Balakrishnan, N, Read,C. B. & Vidakovic, B.(eds.) *Encyc. Statist. Sciences*, 2nd ed. **8**, 5538–5345.

Sheppard, W.F. (1898). On the geometrical treatment of the 'normal curve' of statistics, with special reference to correlation and to the theory of error. *Proc. Roy. Soc. London*, **62**, 171–173.

Streitberg, B. (1990). Lancaster interactions revisited. *Ann. Statist.* **18**, 1878–1885.

Wermuth, N. (2015). Graphical Markov models, unifying results and their interpretation. In: (Balakrishnan, N. et al., eds.) *Wiley Statsref: Statistics Reference Online*; also on arXiv: 1505.02456.

Wermuth, N. & Cox, D. R. (1992). On the relation between interactions obtained with alternative codings of discrete variables. *Methodika* **VI**, 76–85.

WERMUTH, N. & MARCHETTI, G. M. (2014). Star graphs induce tetrad correlation for Gaussian as well as for binary distributions. *Electron. J. Statist.* **8**, 253–273; also on arXiv: 1307.5396.

WERMUTH, N., MARCHETTI, G. M. & COX, D. R. (2009). Triangular systems for symmetric binary variables. *Electron. J. Statist.* **3**, 932–955.

WERMUTH, N., MARCHETTI, G. M. & ZWIERNIK, P. (2014). Binary distributions of concentric rings. *J. Multiv. Analysis* **130**, 252–260; also on arXiv: 1311.5655.