Stat

The ISI's Journal for the Rapid Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.155

Explicit, identical maximum likelihood estimates for some cyclic Gaussian and cyclic Ising models

Giovanni M. Marchetti[a] **and Nanny Wermuth**[b,c]*

Received 6 June 2017; Accepted 30 June 2017

Cyclic models are a subclass of graphical Markov models with simple, undirected probability graphs that are chordless cycles. In general, all currently known distributions require iterative procedures to obtain maximum likelihood estimates in such cyclic models. For exponential families, the relevant conditional independence constraint for a variable pair is given all remaining variables, and it is captured by vanishing canonical parameters involving this pair. For Gaussian models, the canonical parameter is a concentration, that is, an off-diagonal element in the inverse covariance matrix, while for Ising models, it is a conditional log-linear, two-factor interaction. We give conditions under which the two different likelihood functions, that is, one for continuous and one for binary variables, permit nevertheless explicit maximum likelihood estimates, and we show that their estimated correlation matrices are identical, provided the relevant starting correlation matrices coincide.

Keywords: canonical parameter; chain graph; graphical Markov model; palindromic Ising model; quadratic exponential distribution

1 Introduction

Graphical Markov models are a large class of multivariate models that permit to model undirected and directed dependences of different types (Studený, 2005; Scutari & Strimmer, 2011; Sadeghi & Lauritzen, 2014; Wermuth, 2015) and have applications in many fields in the natural, social and medical sciences. These models started to be developed after the analogies between undirected association models for continuous variables with joint Gaussian distributions and for discrete distributions had been recognized (Wermuth, 1976; Darroch et al., 1980), based on models that had been studied without graphs by Dempster (1972) and by Bishop et al. (1975), respectively.

In probabilistic graphs, the nodes represent random variables, edges present permit conditional dependences and missing edges capture conditional independence constraints in many different types of graphs; see, for instance, Wermuth & Sadeghi (2012). The conditioning sets depend on the types of graph and edges present; for variable pairs in undirected graphs drawn with full-line edges as in Figure 1, these are all remaining variables.

Distributions that satisfy all independences defining a given graph are said to be generated over this graph. These may result in quite different models when the involved variables are of continuous type as for Gaussian distributions

[a]**Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, viale Morgagni 59, 50134 Florence, Italy**

[b]**Mathematical Sciences, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden**

[c]**Department of Medical Psychology and Medical Sociology, Johannes Gutenberg University of Mainz, Saarstrasse 21, 55099 Mainz, Germany**

*__Email: wermuth@chalmers.se__

(wileyonlinelibrary.com) DOI: 10.1002/sta4.155

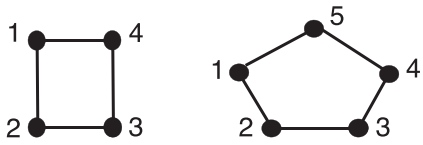

Figure 1. Left, a labelled 4-cycle; right a labelled 5-cycle.

or of binary type as for Ising models. Both of these belong however to the quadratic exponential family because their canonical parameters involve just two variables. For a Gaussian model, the canonical parameters are concentrations, the elements of the inverse covariance matrix, while for an Ising model, these are two-factor, log-linear interactions.

Many missing edges typically simplify graphs and models and their interpretation, but for chordless cycles, this happens only if the missing edges concern the same node, as described next with an example in Figure 1.

Chordless cycles are finite, simple undirected graphs in four or more nodes, which have as many nodes as edges and by starting at any node of a d-cycle, and walking along its d edges, one returns to the starting node. In the 5-cycle of Figure 1, the factorization of any joint density generated over the graph, the missing edges for instance for pairs $(1, 3)$ and $(2, 4)$ do not lead to a simpler factorization of this density. This makes cyclic models more complex than models without chordless cycles even though each node is touched by just two edges, that is, has only two neighbours.

By marginalizing over any node in a chordless d-cycle, one removes this node and replaces the two edges touching this node by a new edge that couples the previously uncoupled neighbours (Wermuth, 2011; Sadeghi, 2016). In this way, a chordless $(d-1)$-cycle results for $d > 4$. In the 5-cycle of Figure 1, the independence of pairs $(1, 3)$ $(1, 4)$ given three remaining variables simplifies to 1 independent of 3 and 4 given $2, 5$ and hence to a missing edge for $(1, 3)$ also in the two marginal 4-cycles of nodes $1, 2, 3, 4$ or $1, 2, 3, 5$. The same holds in more complex graphical Markov models which satisfy the so-called intersection property; see, for example, Sadeghi & Wermuth (2016). This happens, in particular, whenever the distribution is strictly positive, while necessary and sufficient conditions are currently known only for joint distributions of either Gaussian or discrete variables (San Martín et al., 2005).

Chordal graphs have no chordless cycle as a subgraph, and each single-node elimination scheme provides a full ordering of all nodes (Tarjan & Yannakakis, 1984). Therefore, joint distributions generated over chordal graphs can be factorized into recursive sequences of single-response, conditional distributions, also called regressions, and the models are said to be decomposable. Until today, decomposable models are used also to study the more complex models generated over non-chordal graphs; see, for instance, Dahl et al. (2008), Thomas & Green (2009), and Studený & Cussens (2016).

Even for joint Gaussian distributions, maximum likelihood estimation for models generated over non-chordal graphs require in general iterative procedures (Speed & Kiiveri, 1986; Sadeghi & Marchetti, 2012; Lauritzen et al., 2017). Only under constraints that are additional to independences, such as equal edge strength, do explicit estimates become available for cyclic Gaussian models (Højsgaard & Lauritzen, 2007).

To our knowledge, no similar results had been obtained so far for Ising models even though there has been recent intensive work on Ising models in statistics (Foygel Barber & Drton, 2015; Martín del Campo et al., 2017; Bhattacharya & Mukherjee, 2018) and in machine learning (Murphy, 2012; Bresler, 2015; Johnson et al., 2016).

On the history of Ising models in physics, a sequence of three papers has been published by the same author: the second is titled "History of the Lenz-Ising model 1950–1965: from irrelevance to relevance" (Niss, 2009). Something similar may possibly now be stated regarding its relevance in statistics and machine learning given the earlier quoted recent work and more forthcoming insights.

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.155

The ISI's Journal for the Rapid Dissemination of Statistics Research

Of special interest are two versions of the Ising model, which have recently been studied as the general and palindromic Ising models, both with or without conditional independence constraints; see Marchetti & Wermuth (2016). The term palindromic means in linguistics that a sequence of characters or words may be read forward and backward to give the same meaning, as, for instance, in the sentence 'step on no pets'. A joint binary distribution is palindromic if its list of probabilities given in any lexicographic order remains unchanged after fully reversing its listed elements. For Ising models, this type of central symmetry results, surprisingly, when the levels of each variable occur with probability $1/2$. With effect coding, that is with all levels coded as ± 1, this leads to a vector variable of mean zero and a covariance matrix, which coincides with the correlation matrix.

We introduce in Section 2 some more definitions and concepts, derive the two different likelihood functions in general and their solutions for 4-cycles in Section 3, give some extensions in Section 4 and end with a short discussion.

2 Some more definitions and concepts

The arguably best studied class of graphical Markov models, also known as Markov random fields, is specified by an undirected graph $G = (V, E)$, where $V = \{1, 2, \ldots, d\}$, is the finite vertex or node set and $E \subset \{\{s, t\} : s \neq t \in V\}$ denotes the edges present in the graph. The node set indexes the components of a random vector variable $X = (X_s), s \in V$, and a pair $(s, t) \in E$ permits the dependence of X_s and X_t given $X \setminus \{X_s, X_t\}$. When an edge $\{s, t\}$ is missing in G, then the variables X_s and X_t are independent given the remaining variables. The graph G may be represented by a symmetric binary matrix, say A, called its adjacency matrix, which has elements one for edges present, $\{s, t\} \in E$, and zero otherwise.

Two seemingly quite different models are generated over such an undirected graph G. For a continuous random vector X of dimension d, with realization x, the density function $f(x)$ for a mean-centred joint Gaussian distribution is

$$f(x) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp(-\tfrac{1}{2} x^T \Sigma^{-1} x)\,, \tag{1}$$

with Σ the covariance matrix of elements σ_{st}, the inverse of a matrix M denoted by M^{-1} and its determinant by $|M|$.

In this paper, we take the variables to have also unit variances so that Σ coincides with the correlation matrix. In both situations, it holds for the elements of Σ^{-1}, denoted by κ_{st} and called concentrations, that $\{s, t\} \notin E \iff \kappa_{st} = 0$. Together with $\sigma_{ss} > 0$ and $\sigma_{st} \neq 0$ for $\{s, t\} \in E$ of the covariance matrix, these constraints define a unique matrix (Dempster, 1972; Lauritzen et al., 2017). As a consequence for any Gaussian Markov random field, not only the covariances, corresponding to edges present in the graph, but also the variances are matched in a maximum likelihood estimate (MLE) to the observed ones. This implies that working with a correlation matrix instead of a covariance matrix leaves the estimated conditional independences and the estimated correlation matrix unchanged.

For palindromic Ising models, X taking values $x = (i_1, \ldots, i_d)$ with $i_s \in \{-1, 1\}$, the joint probability function $p(x)$ is

$$p(x) = Z(\lambda)^{-1} \exp\Big(\textstyle\sum_{s<t} \lambda_{st} i_s i_t\Big)\,, \tag{2}$$

where λ_{st} are the two-factor, log-linear interactions for which $\{s, t\} \notin E \iff \lambda_{st} = 0$ and λ denotes the vector of λ_{st}. The normalizing constant $Z(\lambda)$ assures that the probabilities add to one. General Ising models may have additional

(wileyonlinelibrary.com) DOI: 10.1002/sta4.155

one-factor, log-linear λ_s terms in the joint probabilities and hence non-zero means. For palindromic Ising models, X has mean zero, unit variance and

$$p(i_1, \ldots, i_d) = p(-i_1, \ldots, -i_d), \qquad \text{for all } (i_1, \ldots, i_d) \in \{-1, +1\}^d \tag{3}$$

so that the list of the joint probabilities in any lexicographical order stays unchanged when it is reversed.

An important characterizing feature of a palindromic binary distributions is that all odd-order log-linear – as well as linear interactions – vanish, that is, they are constrained to be zero (Marchetti & Wermuth, 2016, proposition 2.2), while in palindromic Ising models, all even-order, higher than two-factor, log-linear interactions vanish as well. Arguably, models of equal edge strength, define one of the simplest, non-trivial subclasses.

Palindromic Ising models have been used extensively in statistical physics, where they are known as models of ferromagnetism without external magnetic field, also called zero-mean Ising models; see, for instance, Globerson & Jaakkola (2007) and Johnson et al. (2016). The equal edge-strength models are also known as Curie-Weiss models. Cyclic models of this type are discussed here in Section 3.2. Zero-mean Ising models are also useful to summarize well-fitting Gaussian Markov random fields in terms of median-dichotomized variables, such as for the averaged grades reported in Marchetti & Wermuth (2016). These data happen to display total positivity of order two, and for such data, no effect-reversal can ever arise after median dichotomizing; see Fallat et al. (2017), proposition 3.2(iii).

Because in cycles – triangles or chordless cycles – every node has precisely two neighbours, not only the conditional odds ratios are constant at all levels of the remaining variables in the palindromic Ising model, but also the conditional two-factor expectations are linear and constant at all levels of the remaining variables. This assures that these conditional correlations coincide with a single partial correlation coefficient; see Baba et al. (2004), theorem 1.

3 The two likelihood functions in cyclic models

3.1 The Gaussian cyclic model of equal edge strength

For the Gaussian density of equation (1), where we assumed that X has components with mean zero and unit variance, the covariance matrix, Σ, coincides with the correlation matrix. For a random sample $x_1, \ldots, x_N$ of size N and no independence constraints, the log-likelihood function becomes, disregarding the common factor N,

$$\ell(\Sigma) = \tfrac{1}{2} \log |\Sigma^{-1}| - \tfrac{1}{2} \mathrm{tr}(\Sigma^{-1} R)\,, \tag{4}$$

where $R = \sum_x xx^T/N$ and $\mathrm{tr}(M)$ denotes the trace of a symmetric matrix M. The elements r_{st} of R are simple observed correlation coefficients if the sample values are also standardized to have mean zero and unit variance.

For a chordless cycle, maximization is for symmetric positive definite matrices Σ such that the inverse $K = \Sigma^{-1}$ has elements $\kappa_{ss} > 0$ and $\kappa_{st} = 0$ whenever edge (s, t) is missing in the graph. Partial correlations given all remaining variables are then zero as well because they can be computed from the elements of K as

$$\theta_{st} = -\kappa_{st} (\kappa_{ss} \kappa_{tt})^{-\frac{1}{2}}\,. \tag{5}$$

For a proof of equation (5), see, for instance, Wermuth et al. (2006), section 2.3.

For all Gaussian cyclic models in $d > 3$ variables having equal edge-strength dependences, the partial correlation matrix Θ has equal, non-zero elements, θ, so that the second term in the likelihood equation (4) reduces to

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.155

The ISI's Journal for the Rapid Dissemination of Statistics Research

$$\mathrm{tr}(\Sigma^{-1}R) = d\,(1-2\theta\,\bar{r})\,\kappa, \tag{6}$$

where $\bar{r}$, the sufficient statistic, is the average of the observed correlations corresponding to the edges present in the graph G and κ is one of the identical elements along the diagonal of the concentration matrix Σ^{-1}.

Specifically, for the chordless 4-cycle in Figure 1 with equal edge strength, the model implies

$$\Theta = \begin{pmatrix} 1 & \theta & 0 & \theta \\ . & 1 & \theta & 0 \\ . & . & 1 & \theta \\ . & . & . & 1 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & \rho & \rho^* & \rho \\ . & 1 & \rho & \rho^* \\ . & . & 1 & \rho \\ . & . & . & 1 \end{pmatrix}, \quad \rho = \frac{\theta}{1-2\theta^2}, \quad \rho^* = \frac{2\theta^2}{1-2\theta^2}, \quad -\tfrac{1}{2} < \theta < \tfrac{1}{2}. \tag{7}$$

In symbolic form, the correlations in Σ result generally best from the partial correlations in Θ by computing $M = (2I - \Theta)^{-1}|\Theta|$, where I denotes the identity matrix, and standardizing the elements of M to give $\rho_{st} = m_{st}/\sqrt{m_{ss}m_{tt}}$.

By using Σ of equation (7), one obtains for the two terms in the log-likelihood function of equation (4):

$$|\Sigma^{-1}| = \frac{(1-2\theta^2)^4}{(1-4\theta^2)^3}, \quad \mathrm{tr}(\Sigma^{-1}R) = \frac{4(1-2\theta^2)(1-2\bar{r}\theta)}{1-4\theta^2},$$

where $\bar{r}$ is the average of the observed correlations at edges present in the labelled 4-cycle of Figure 1:

$$\bar{r} = (r_{13} + r_{14} + r_{23} + r_{24})/4. \tag{8}$$

The log-likelihood function of equation (4) is then

$$\ell(\theta) = \frac{4\left(2\theta^2-1\right)\left(2\bar{r}\theta-1\right)}{4\theta^2-1} - 3\log\left(1-4\theta^2\right) + 4\log\left(1-2\theta^2\right).$$

The likelihood equations for $\hat{\theta}$ result by setting the derivative of $\ell(\theta)$ with respect to θ to zero:

$$\frac{\left(8\theta^4-2\theta^2+1\right)\left(2\bar{r}\theta^2+\theta-\bar{r}\right)}{\left(2\theta^2-1\right)\left(4\theta^2-1\right)^2} = 0.$$

For $-\frac{1}{2} < \theta < \frac{1}{2}$, the equation $8\theta^4 - 2\theta^2 + 1 = 0$ has no real solution, so that the MLE $\hat{\theta}$ solves $2\bar{r}\theta^2 + \theta - \bar{r} = 0$. Because the average correlation for the edges present must satisfy $-1 < \bar{r} < 1$, there is a single solution:

$$\hat{\theta} = \frac{\sqrt{8\bar{r}^2+1}-1}{4\bar{r}}. \tag{9}$$

3.2 The palindromic cyclic Ising model of equal edge strength

For a palindromic Ising model, we assume that the sample data are arranged in a table of counts $n(x)$, where x denotes a cell and $N = \sum_x n(x)$. Let the parameters λ_{st} be arranged as elements in a symmetric matrix Λ with zeros along the diagonal. Then, under multinomial sampling and no further constraints, the log-likelihood function for equation (2) is

$$\ell(\Lambda) = \tfrac{1}{2}N^{-1}\textstyle\sum_x n(x)x^T\Lambda x - \log\{\sum_x \exp(\tfrac{1}{2}x^T\Lambda x)\}, \tag{10}$$

where the second term on the right-hand side is $\log Z(\Lambda)$. For any cyclic model in d variables, maximization is for real symmetric matrices Λ such that $\lambda_{st} = 0$ for edges (s,t) missing in the graph.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.155

With the additional assumption of equal edge strength, $\lambda_{rs} = \lambda$ for $\{r,s\} \in E$, so that $\Lambda = \lambda A$, where A is the adjacency matrix of the cycle, the log-likelihood function reduces to

$$\ell(\lambda) = \tfrac{1}{2}\lambda \mathrm{tr}\{AB\} - \log Z(\lambda)\,, \tag{11}$$

where $B = N^{-1}\sum_x n(x)xx^T$. Equation (11) uses $\sum_x n(x)x^T\Lambda x = \lambda\sum_x n(x)\mathrm{tr}(x^TAx) = \lambda\mathrm{tr}\{A\sum_x n(x)xx^T\}$.

The matrix B contains uncorrected sums of squares and cross products, but it may also be interpreted as an empirical correlation matrix R for data corresponding to the symmetrized counts, $\bar{n}(x) = \{n(x) + n(-x)\}/2$. One obtains

$$R = N^{-1}\sum_x \bar{n}(x)xx^T = N^{-1}\tfrac{1}{2}\sum_x\{n(x) + n(-x)\}xx^T = B.$$

Thus, the log-likelihood function in equation (11) can be written as

$$\ell(\lambda) = d\,\bar{r}\,\lambda - \log Z(\lambda)\,, \tag{12}$$

where $\bar{r}$ is the average of elements of R at edges present in the cycle, like in equation (8) for the Gaussian case.

In the special case of $d = 4$, the term from the normalizing constant, $\log Z(\lambda)$, simplifies as well, giving

$$\ell(\lambda) = d\bar{r}\lambda - \log\{2\exp(-4\,\lambda) + 2\exp(4\,\lambda) + 12\}.$$

We use next a simple relation between partial correlations and log-linear interactions, derived in Section 4, $\theta = \frac{1}{2}\tanh(2\lambda)$, to reparametrize this log-likelihood function as

$$\ell(\theta) = 2\bar{r}\tanh^{-1}(2\theta) - \log\left(\frac{4(4 - 8\theta^2)}{1 - 4\theta^2}\right).$$

After setting the derivative of $\ell(\theta)$ with respect to θ to zero, the likelihood equations for $\hat{\theta}$ are

$$-\frac{4\left(2\bar{r}\theta^2 + \theta - \bar{r}\right)}{8\theta^4 - 6\theta^2 + 1} = 0\,.$$

The single solution is identical to the one in equation (9). Thus, provided that the sample correlation matrices R coincide for the Gaussian variables and for the symmetrized binary variables, there is the same explicit MLE of Θ for the Gaussian and for the palindromic Ising model generated over 4-cycles of equal edge strength.

4 Some more explanations and some extensions

In palindromic Ising models generated over d-cycles, the conditional correlations are constant at all level combinations of the remaining variables so that each of them coincides with a corresponding single partial correlation coefficient (Wermuth & Marchetti, 2017). For the 4-cycle of equal edge strength in Figure 1, the partial and marginal correlations are identical to those in equation (7) for the corresponding Gaussian model. In addition, the joint probabilities and the dependence structure may be expressed in terms of the single conditional odds ratio, $a > 1$, as in the next table, where the lower level is written as zero to save space and $c(a) = 2(a^2 + 6a + 1)$ is the sum of the table entries.

$$\begin{bmatrix} x: & 0000 & 1000 & 0100 & 1100 & 0010 & 1010 & 0110 & 1110 & 0001 & 1001 & 0101 & 1101 & 0011 & 1011 & 0111 & 1111 \\ c(a)\,p(x): & a^2 & a & a & a & a & 1 & a & a & a & a & 1 & a & a & a & a & a^2 \end{bmatrix}.$$

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.155

The ISI's Journal for the Rapid Dissemination of Statistics Research

Thus, the partial correlations are zero for pairs $(1,3)$ and $(2,4)$, and the relations between θ and $a > 1$ are

$$\theta = \tfrac{1}{2}(a-1)/(a+1), \qquad a = (1+2\theta)/(1-2\theta).$$

With effect coding, ± 1, of all variable levels, $\lambda = \log(a)/4$ and $a = e^{4\lambda}$ so that one obtains directly:

$$\theta = \tfrac{1}{2}(e^{4\lambda}-1)/(e^{4\lambda}+1) = \tfrac{1}{2}\tanh(2\lambda), \qquad \lambda = \tfrac{1}{2}\tanh^{-1}(2\theta). \tag{13}$$

These relations hold in all equal edge-strength d-cycles, and extensions will be explored in a forthcoming paper.

There is an extremely attractive property of maximum-likelihood estimation: the MLE $\hat{\theta}$ relates in the same ways to the MLEs $\hat{a}$ and $\hat{\lambda}$ that hold for the corresponding parameters (Fisher, 1922). This implies here in particular that also the MLE of the joint probability vector is available in closed form.

With all odds ratios a in $0 < a < 1$, the sign of all non-zero partial correlation θ changes, as well as the order of terms in a^2, a and 1, but the independence structure remains unchanged. The contingency table, having as smallest entry a one, is then as shown in the second row of the next table, where also $c(a)$ is unchanged.

The structure in the third row of the next table has $d(a) = 8(1+a)$ and $\Theta = \Sigma$; the edge strength is equal to θ for pairs $(1,2),(2,3),(3,4)$, while for pair $(1,4)$, it is $-\theta$.

$$\begin{bmatrix} x: & 0000 & 1000 & 0100 & 1100 & 0010 & 1010 & 0110 & 1110 & 0001 & 1001 & 0101 & 1101 & 0011 & 1011 & 0111 & 1111 \\ c(a)\,p(x): & 1 & a & a & a & a & a^2 & a & a & a & a & a^2 & a & a & a & a & 1 \\ d(a)\,p(x): & a & a & 1 & a & 1 & 1 & 1 & a & a & 1 & 1 & 1 & a & 1 & a & a \end{bmatrix}.$$

Thus, there is the same 4-cycle in both the partial and marginal correlations, and the MLE of θ turns into

$$\hat{\theta} = \hat{\rho} = \tilde{r} \quad \text{with} \quad \tilde{r} = (r_{12} + r_{23} + r_{34} - r_{14})/4 \ \text{ for } -1/2 < \tilde{r} < 1/2\,,$$

where $\tilde{r}$ is a signed average of observed correlations at edges present in the cycle for the symmetrized counts in the palindromic Ising model or, as can also be shown, for the observed correlations in a corresponding Gaussian model.

Going back to the equal edge-strength model for the 5-cycle, the MLE $\hat{\theta}$ is also identical for the Gaussian and for the palindromic Ising model: it solves a different quadratic equation, which uses again an average of correlations for the edges present, $\bar{r}$, which is the solution of

$$(\bar{r}-1)\theta^2 + (\bar{r}+1)\theta - \bar{r} = 0 \text{ for } -3/5 < \bar{r} < 1 \text{ and } -1/2 < \theta < 1/2. \tag{14}$$

For every positive definite observed 5×5 correlation matrix, $\bar{r} > -1/4$ so that a single solution exists. For the equal edge-strength 6-cycle, a cubic equation is to be solved for the MLE $\hat{\theta}$. We have not explored the details of the solutions for $d > 6$.

5 Discussion

Cyclic models are probability distributions generated over graphs in the simplest subclass of non-chordal graphs, that is over chordless cycles. In general, iterative procedures are required for obtaining an MLE in all chordless cycles.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.155

Mean-zero Ising models, also known as palindromic Ising models, are joint distributions of symmetric binary variables within the quadratic exponential family. Recent results prove, surprisingly, that pairwise conditional independences given all remaining variables show as zeros in their overall partial correlation matrices just as for joint Gaussian distributions. This holds not only for simple models like Markov chains but also for cyclic models.

For chordless 4-cycles of the binary variables and dependences of equal strength at edges present, we derived a closed-form MLE for the correlation matrix of symmetrized counts. We could show that an identical MLE arises when starting with an observed correlation matrix of the same form for a standardized joint Gaussian distribution, that is, in spite of the quite different likelihood functions. There is an even simpler, identical form MLE when there is a chordless 4-cycle in both partial and marginal correlations because the dependence at one edge agrees only in absolute value with the others: the MLEs of the partial correlations are then signed averages of the simple correlations corresponding to edges present.

For 4-cycle models of equal edge strength, one can show also that simple correlations induced for the missing edges are positive, no matter whether the starting non-zero partial correlations are all positive or all negative. We could however not find general conditions assuring this if the given negative partial correlations are not of the same size.

A most attractive feature of the studied palindromic Ising models is the closed-form relation between a non-zero partial correlation and the constant conditional log-linear, two-factor interaction because this leads directly to the full contingency table from the matrix of partial correlation, for both parameters and their maximum likelihood estimates.

Acknowledgements

We used Matlab for the computations. We thank David Cox for stimulating discussions, and him, Rolf Sundberg as well as the referees for their constructive comments.

References

Baba, K, Shibata, R & Sibuya, M (2004), 'Partial correlation and conditional correlation as measures of conditional independence', *Australian & New Zealand Journal of Statistics*, **46**, 657–664.

Bhattacharya, BB & Mukherjee, S (2018), 'Inference in Ising models', *Bernoulli*, **24**, 493–525.

Bishop, YMM, Fienberg, SF & Holland, PW (1975), *Discrete Multivariate Analysis*, *MIT Press*, Cambridge.

Bresler, G (2015), *Efficiently learning Ising models on arbitrary graphs*, Proceedings 47th ACM Symposium on Theory of Computing, STOC 15, Portland, OR, 771–782.

Dahl, J, Vandenberghe, L & Roychowdhury, V (2008), 'Covariance selection for non-chordal graphs via chordal embedding', *Optimization Methods Software*, **23**, 501–520.

Darroch, JN, Lauritzen, SL & Speed, TP (1980), 'Markov fields and log-linear models for contingency tables', *Annals of Statistics*, **8**, 522–539.

Dempster, AP (1972), 'Covariance selection', *Biometrics*, **28**, 157–175.

Fallat, S, Lauritzen, S, Sadeghi, K, Uhler, C, Wermuth, N & Zwiernik, P (2017), 'Total positivity in Markov structures', *Annals of Statistics*, **45**, 1152–1184, also on ArXiv: 1510.01290.

Fisher, RA (1922), 'On the mathematical foundations of theoretical statistics', *Philosophical Transactions of the Royal Society of London Series A*, **222**, 309–368.

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.155

The ISI's Journal for the Rapid Dissemination of Statistics Research

Foygel Barber, R & Drton, M (2015), 'High-dimensional Ising model selection with Bayesian information criteria', *Electronic Journal of Statistics*, **9**, 567–607.

Globerson, A & Jaakkola, T (2007), *Approximate Inference in using planar graph decompositions*, Advances in Neural Information Processing (NIPS 2006), **19**, Vancouver, Canada.

Højsgaard, S & Lauritzen, SL (2007), 'Inference in graphical Gaussian models with edge and vertex symmetries with the gRc package for R', *Journal of Statistical Software*, **23**, 6.

Johnson, JK, Oyen, D, Cherkov, M & Netrapalli, P (2016), 'Learning planar Ising models', *Journal of Machine Learning Research*, **17**, 1–26.

Lauritzen, SL, Uhler, C & Zwiernik, P (2017), *'Maximum likelihood estimation in Gaussian models under total positivity',* submitted, also on ArXiv: 1702.04031.

Marchetti, GM & Wermuth, N (2016), 'Palindromic Bernoulli distributions', *Electronic Journal of Statistics*, **10**, 2435–2460. also on ArXiv: 1510.09072.

Martín del Campo, A, Cepeda S & Uhler, C (2017), 'Exact goodness-of-fit testing for the Ising model', *Scandinavian Journal of Statistics*, **44**, 285–306.

Murphy, KP (2012), *Machine Learning: A Probabilistic Perspective*, *MIT Press*, Cambridge.

Niss, M (2009), 'History of the Lenz–Ising model 1950–1965: from irrelevance to relevance', *Archive for History of Exact Sciences*, **63**, 243–287.

Sadeghi, K (2016), 'Marginalization and conditioning for LWF chain graphs', *Annals of Statistics*, **44**, 1792–1816.

Sadeghi, K & Lauritzen, SL (2014), 'Markov properties for mixed graphs', *Bernoulli*, **20**, 676–696.

Sadeghi, K & Marchetti, GM (2012), 'Graphical Markov models with mixed graphs in R', *The R Journal*, **4**, 65–73.

Sadeghi, K & Wermuth, N (2016), 'Pairwise Markov properties for regression graphs', *Stat*, **5**, 286–294; also on ArXiv: 1512.09016.

San Martín, E, Mouchart, M & Rolin, JM (2005), 'Ignorable common information, null sets and Basu's first theorem', *Sankhya*, **67**, 674–698.

Scutari, M & Strimmer, K (2011), *Introduction to graphical modelling*, Handbook of Statistical Systems Biology in Balding, DJ, Stumpf, M & Girolami, M (eds), Chapter 11, Wiley Online Library, John Wiley & Sons, Chichester, UK.

Speed, TP & Kiiveri, HT (1986), 'Gaussian Markov distributions over finite graphs', *Annals of Statistics*, **14**, 138–150.

Studený, M (2005), *Probabilistic Conditional Independence Structures*, *Springer*, London.

Studený, M & Cussens, J (2016), 'The chordal graph polytope for learning decomposable models', *JMLR Workshop and Conference Proceedings*, **52**, 499–510.

Tarjan, RE & Yannakakis, M (1984), 'Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs', *SIAM Journal on Computing*, **13**, 566–579.

Thomas, A & Green, PJ (2009), 'Enumerating the junction trees of a decomposable graph', *Journal of Computational and Graphical Statistics*, **18**, 930–940.

Wermuth, N (1976), 'Analogies between multiplicative models for contingency tables and covariance selection', *Biometrics*, **32**, 95–108.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.155

Wermuth, N (2011), 'Probability models with summary graph structure', *Bernoulli*, **17**, 845–879.

Wermuth, N (2015), *Graphical Markov models, unifying results and their interpretation*, *Wiley Statsref: Statistics Reference Online*, John Wiley & Sons, Hoboken, NJ,. also on ArXiv: 1505.02456.

Wermuth, N, Cox, DR & Marchetti, GM (2006), 'Covariance chains', *Bernoulli*, **12**, 841–862.

Wermuth, N & Sadeghi, K (2012), 'Sequences of regressions and their independences (with discussion)', *TEST*, **21**, 215–279, also on ArXiv: 2523.1103.

Wermuth, N & Marchetti, GM (2017), 'Generating large Ising models with Markov structure via simple linear relations', *submitted*. also on ArXiv: 1704.01649.