Generalized Graphical Models for Discrete Data

JOZEF L. TEUGELS[†] & JOHAN VAN HOREBEEK[†,‡,1,2]

[†] Department of Mathematics, K.U.Leuven,
Celestijnenlaan 200 B, B-3001 Heverlee, Belgium
[‡] Centro de Investigación en Matemáticas,
Apartado Postal 402, 36000 Guanajuato, Gto, Mexico
email: horebeek@fractal.cimat.mx

Abstract *Traditional graphical models are extended by allowing that the presence or absence of a connection between two nodes depends on the values of the remaining variables. We first compare the extended model to the classical log-linear model. After discussing the induced consistency problem we illustrate the corresponding estimation problem by way of an example.*

Keywords: Graphical Model, Conditional Independency, Markov Random Field, Log-linear Model.

[1] This author was partly supported by a "Cátedra Patrimonial" of the Mexican Council of Science and Technology.

[2] To whom reprints should be sent; CIMAT, AP. 402, 36000 Guanajuato, GTO, Mexico

1 Introduction

Given a multivariate discrete distribution, $\mathbf{X} = (X_1, \cdots, X_n)$, *graphical models* are nowadays a popular approach to represent the interaction structure between the components X_i. Such models are characterized by a graph, where each node is associated with a component of $\mathbf{X}$ and where the absence of a connection between nodes i and j means that X_i and X_j are independent given the values of *all* remaining variables. Apart from the compact visual representation, such an approach has the advantage that many properties of the underlying discrete distribution can be immediately formulated in terms of characteristics of the graph. We refer to Ripley (1994) for an overview.

The fundamental assumption for the usual classical graphical models is that the given independency statements should be valid for all values of the variables in the conditional part. In practice this is often a severe restriction. For example, it might be instructivive to know that X_1 is independent of X_3 when $X_2 = 0$, without saying anything about independence when $X_2 = 1$. In this paper we show how to allow such generalizations.

In the sequel we will use the following notation:

$\mathbf{X} \in MD(r_1, \cdots, r_n) \Leftrightarrow X_i \in \{0, \cdots, r_i - 1\}$,
$p_{x_1, \cdots, x_n} = P(X_1 = x_1, \cdots, X_n = x_n)$,
$X_A = \{X_i, i \in A\}$,
N, the total number of observations,
$n^A_{x_A}$, the number of observations equal to x_A.

Furthermore, we will write $X_A \perp X_B | X_C$ iff X_A and X_B are conditionally independent given X_C.

2 Model definition

We define a generalized graphical model to represent a family of random vectors characterized by a set of independency statements of the form:

$$X_i \perp X_j | \{X_k = x_k, k \notin \{i, j\}\}. \tag{1}$$

The graph is constructed by assigning a node to each variable X_i. Two nodes are *not* connected iff the two corresponding variables are independent, given all of the other variables. If the independency only prevails under a specific choice of values for the remaining variables, then a connection is drawn. However, we write a *label* next to the connection, compiling all cases for which nothing is said about potential dependency.

Example 2.1
Given $\mathbf{X} \in MD(2, 2, 2, 2)$:

$$\forall x_1 : X_3 \perp X_4 | X_2 = 1, X_1 = x_1 \tag{2}$$

$$\forall x_3, x_4 : X_1 \perp X_2 | X_3 = x_3, X_4 = x_4 \tag{3}$$

the corresponding graph looks like:

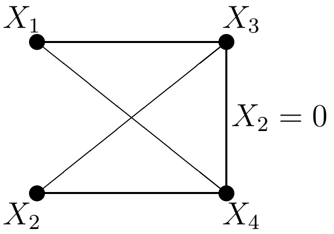

Figure 1

Although the above class of models is defined in a non parametric way, it is instructive to relate them with *log-linear models* (assuming strictly positive distributions).
Recall that any conditional independency can be expressed as a certain factorization property of the joint distribution. There results that (1) can be translated in terms of a restriction on the parameters μ of the log-linear model defined by $\log p_{x_1,\cdots,x_n} = \sum_A \mu^A_{x_A}$ with the restriction: $\forall j, A \ni j$ and x: $\sum_{x_j} \mu^A_{x_A} = 0$ where the sum is over all possible values of the j-th component of x, keeping the remaining components fixed. Specifically,

$$X_i \perp X_j | \{X_k = x'_k, k \notin \{i,j\}\}$$

$$\Leftrightarrow$$

$$\forall x_i, x_j : \sum_{A \supset \{i,j\}} \mu^A_{x_A} = 0 \text{ with } x = (x'_1, \cdots, x'_{i-1}, x_i, x'_{i+1}, \cdots, x'_{j-1}, x_j, x'_{j+1}, \cdots, x'_n).$$

Consequently, in the above example, (2) translates into:

$$\mu^{3,4}_{x_3,x_4} + \mu^{2,3,4}_{1,x_3,x_4} + \mu^{1,3,4}_{x_1,x_3,x_4} + \mu^{1,2,3,4}_{x_1,1,x_3,x_4} = 0.$$

There is also a link between our models and the *splitmodels*, introduced by Hojsgaard et al. (1991). For example, when modeling (2), a splitmodel separates the data into two sets, one with $X_2 = 0$, the other with $X_2 = 1$. Both models are then fitted separately. Such an approach has the drawback that a variable, used in a split of the data, can no longer be used in a subsequent independency statement. For example a splitmodel based on (2) can not include the model (3).

2.1 Consistency problems

In classical graphical models, any set of independency statements of the form (3) specifies a non empty family of distributions. However, in our generalized models, there may appear relationships among the introduced independencies. By way of illustration we formulate the following property.

Property 2.1 *Suppose that* $\mathbf{X} \in MD(2, 2, 2, r_4, \cdots, r_n)$. *If*

$$X_1 \perp X_2 | X_3 = 1 - x_3,\ X_4 = x_4, \cdots, X_n = x_n \tag{4}$$

$$X_1 \perp X_3 | X_2 = x_2,\ X_4 = x_4, \cdots, X_n = x_n \tag{5}$$

$$X_1 \perp X_2 | X_3 = x_3,\ X_4 = x_4, \cdots, X_n = x_n \tag{6}$$

then

$$X_1 \perp X_3 | X_2 = 1 - x_2,\ X_4 = x_4, \cdots, X_n = x_n.$$

Proof: Take $n = 3$ and $x_1 = x_2 = x_3 = 1$. Then

$$P(X_1 = x_1 | X_2 = 0, X_3 = 0) \overset{(4)}{=} P(X_1 = x_1 | X_2 = 1, X_3 = 0) \overset{(5)}{=}$$

$$P(X_1 = x_1 | X_2 = 1, X_3 = 1) \overset{(6)}{=} P(X_1 = x_1 | X_2 = 0, X_3 = 1).$$

□

The above property shows that it is not possible to remove the connection between X_2 and X_3 in Figure 1. It is intuitively clear that if a variable X_i influences the dependency between two variables X_j, X_k, then X_i, X_j and X_k should be somehow dependent. On the other hand, as Figure 5 further on represents a valid model, we don't need to require that the dependency should be valid for *all* values of the remaining variables.

To end up with a more flexible description of a discrete distribution, we can rely on the concept of *blocks* as introduced in Teugels et al. (1995). The characterization of necessary and sufficient conditions so obtained will guide the implementation of a detection algorithm rather than a criterion for visual inspection.
For simplicity we restrict ourselves to binary variables.

Definition 2.1 *For a given variable* $\mathbf{X} \in MD(2, \cdots, 2)$*, we define the block* $\mathcal{B}^{\log p}(X)$ *as the element of* $\mathcal{R}^{2,\cdots,2}$*:*

$$\mathcal{B}^{\log p}(X)_{x_1,\cdots,x_n} = \log P(X_1 = x_1, \cdots, X_n = x_n).$$

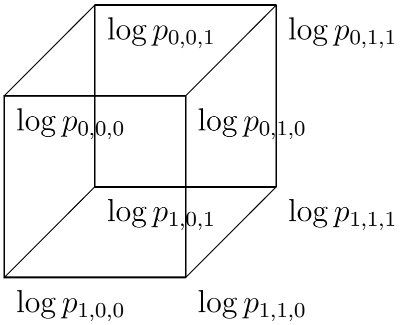

Figure 2: Block representation of a three dimensional binary variable

Addition and scalar multiplication for blocks are defined by the obvious componentwise addition and scalar multiplication. A Cauchy-type inner product is constructed by summing after componentwise multiplication. In the resulting vector space conditional independencies are equivalent to orthogonality constraints.

Example 2.2
Suppose $\mathbf{X} \in MD(2,2,2)$. Then

$$X_1 \perp X_2 | X_3 = 0 \Leftrightarrow \mathcal{B}^{\log p}(X) \perp W;$$

where W represents the following block:

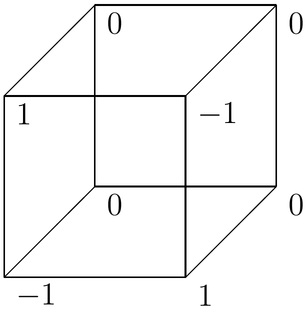

Figure 3

The main characteristic of block W lies in the fact that all of its faces except one contain but zero's. We call this the *marked* face, as it associated with the particular independency statement about X_3..

For a given set of independency statements, we determine all marked faces. We then define a *walk along the marked faces* in the following manner:
a sequence of edges of the block, $(e_1, \cdots, e_k)$, is a walk along marked faces iff e_i and e_{i+1} are parallel edges of the same marked face.

The following result is proven in Teugels et al. (1996):

Property 2.2 *The independency associated with the face defined by the parallel edges* e_1*,* e_k *is implied by a given set of conditional independencies iff one can walk from* e_1 *to* e_k *along the marked faces.*

Example 2.3
Take in Property 2.1, $x_2 = x_3 = 1$ and $n = 3$. We denote an edge by its endpoints and take the following path along the vertical edges of the hypercube to arrive at the requested $X_1 \perp X_3 | X_2 = 0$:

$$\begin{aligned} &e_1 = ((0,0,0);(1,0,0)), \quad e_2 = ((0,1,0);(1,1,0)), \\ &e_3 = ((0,1,1);(1,1,1)), \quad e_4 = ((0,0,1);(1,0,1)). \end{aligned}$$

2.2 Implied independencies

As in the case of classical graphical models, we will use concepts like *paths* and *separability* when characterizing implied independencies.

Definition 2.2 *For a given graph, a given set of variables C and a given set of values x_C for them, the sequence of nodes $\{n_j\}$ is a path under x_C iff for every j, the nodes X_{n_j} and $X_{n_{j+1}}$ are connected given $(X_C = x_C, X_{C^c} = y_{C^c})$ with y arbitrary and (eventually) dependent of j.*

For example, in Figure 4, under $X_3 = 1$, $\{X_1, X_3, X_4, X_2\}$ is a path.

Definition 2.3 *For a given graph, a given set of variables C and a given set of values x_C for them, two sets A and B of nodes are separated by the set C iff every path from an element belonging to A, to an element of B under x_C, contains at least one element of C.*

The following property generalizes the corresponding result for classical graphical models:

Property 2.3 *For a given graph, X_A and X_B are independent given $X_C = x_C$ if C separates A from B under x_C.*

Proof:
It is always allowed to skip some independencies (i.e. to add some connections) in the graph and to prove the requested independency in the new graph. We therefore modify the graph in two steps:

1. First, we adapt the labels reflecting the conditioning on $X_C = x_C$. So, we erase in the labels all occurrences of variables belonging to C and - depending on the label - we add or delete the corresponding connection.

2. Next, we draw an unlabelled connection between the two nodes i and j, if the presence of an (in)dependency between i and j depends on the values of some other variables, not belonging to C.

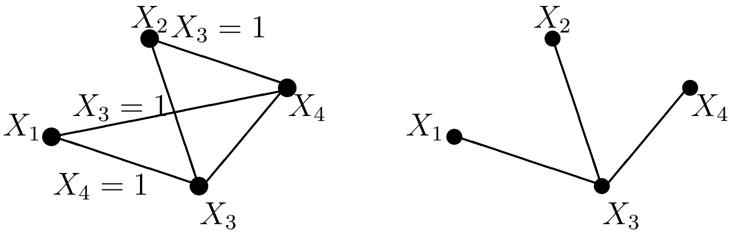

Figure 4:
Illustration of the effect of the above transformation for $C = \{X_3\}, x_3 = 0$.

In this way, given x_C, we obtain a classical graphical model where, by definition of separability, A and B are separated by C. Henceforth, we can apply the results for the latter class of models to show the corresponding independency.

□

2.3 Estimation

Often, the *Maximum Likelihood* estimators can be written down explicitly. Without any loss of generalty, let us take the easiest hypothesis

$$H_0 : \ X_1 \perp X_2 | X_3 = 1$$

which corresponds to $\mu^{1,2}_{0,0} + \mu^{1,2,3}_{0,0,1} = 0$. In an exponential family, and under appropriate regularity conditions, (see Andersen (1991)), the solution of the maximum likelihood equations is obtained by equating the expected values of the sufficient statistics with the observed frequencies. One obtains:

$$EX_1X_2 - EX_1X_2X_3 = n^{1,2}_{1,1}/N - n^{1,2,3}_{1,1,1}/N, \tag{7}$$

and

$$EX_1X_3 = n^{1,3}_{1,1}/N, \quad EX_2X_3 = n^{2,3}_{1,1}/N, \quad EX_i = n^i_1/N. \tag{8}$$

Given that under H_0:

$$EX_1X_2X_3 = \frac{EX_1X_3 \, EX_2X_3}{EX_3} \ , \tag{9}$$

one can solve the above equations immediately for all the maximum likelihood moment estimates. For more complicated hypotheses, we refer to Teugels et al. (1995).

A more general way to obtain a parameter representation for the above hypothesises, is provided by the previously mentioned blockrepresentation. Since independency constraints are equivalent to orthogonality constraints, the formulation of an hypothesis reduces to a classical linear algebra problem of finding a basis for a finite subspace that is orthogonal to given subspaces.

3 Example

Consider the following data shown in Table 1 about the circumstances of accidents with American football players (Buckley (1988)). The variable X_1 indicates whether the accident happened in defense or in an attack , X_2 indicates whether one was throwing the ball or not, while X_3 shows whether the accident happened in a tackle or in a block.

		$X_3 = 0$	$X_3 = 1$
$X_1 = 0$	$X_2 = 0$	125	129
	$X_2 = 1$	85	31
$X_1 = 1$	$X_2 = 0$	216	61
	$X_2 = 1$	62	16

Table 1

1-factor	$\mu_0^1 = -\mu_1^1 = 0.147236(0.04991)$
	$\mu_0^2 = -\mu_1^2 = 0.54974(0.04991)$
	$\mu_0^3 = -\mu_1^3 = 0.44951(0.04991)$
2-factor	$\mu_{0,0}^{1,2} = -\mu_{0,1}^{1,2} = -\mu_{1,0}^{1,2} = \mu_{1,1}^{1,2} = -0.09687(0.04991)$
	$\mu_{0,0}^{1,3} = -\mu_{0,1}^{1,3} = -\mu_{1,0}^{1,3} = \mu_{1,1}^{1,3} = -0.20522(0.04991)$
	$\mu_{0,0}^{2,3} = -\mu_{0,1}^{2,3} = -\mu_{1,0}^{2,3} = \mu_{1,1}^{2,3} = -0.14129(0.04991)$
3-factor	$\mu_{0,0,0}^{1,2,3} = \cdots = -\mu_{1,1,1}^{1,2,3} = -0.11875(0.04991)$

Table 2

Table 2 shows the parameter estimates together with their standard errors using a classical log-linear model.

No classical independency is acceptable because of the highly significant tree-term interaction. On the other hand, the following generalized graphical model is pretty reasonable with a p-value equal to 0.27:

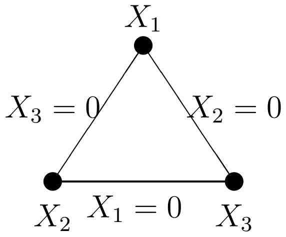

Figure 5

This model corresponds to the hypothesis $\mu_{0,0}^{1,2} + \mu_{0,0,1}^{1,2,3} = \mu_{0,0}^{1,3} + \mu_{0,1,0}^{1,2,3} = \mu_{0,0}^{2,3} + \mu_{1,0,0}^{1,2,3} = 0$ on the parameters of a log-linear model.

4 Final remarks

Going from the classical graphical models to generalized graphical models, there is still one important missing step: given a generalized graphical model, what is the most general representation of the underlying distribution?

For classical models it is well known that one can factorize the distribution in independent functions corresponding to the complete sets within the graph. Therefore, a straightforward adaptation to the generalized model leads to the following form:

$$\log P(X = x) \sim \sum_{\text{complete sets } C} f_C(x_C)\chi(C|x), \tag{10}$$

where $\chi(C|x)$ is a $0-1$ function that indicates whether C is under x a complete set.

In this respect it is interesting to mention the contributions of Baddeley et al. (1989) on expressions like (10) for Markov Point Processes within the context of image analysis. Such results can be translated to discrete graphs as shown in Van Horebeek (1994). However in order to guarantee representations like in (10), one has to make additional restrictions on the graphs. For example, if a variable influences the presence or absence of a complete set, then that variable should be connected with all members of the complete set. In image analysis one can usually completely control the structure of the neighborhood. However, in statistical data analysis, the restriction to a group of graphs without an intuitive meaning, does not seem recommendable.

References

Andersen, E. (1991), *The Statistical Analysis of Categorical Data* (Springer-Verlag).

Buckley, W. (1988), Concussions in football: a multivariate analysis, *American Journal of Sport Medicine*, **16**, 609–617.

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

Ripley, B. (1994) Networks Methods in Statistics, in: F.P. Kelly ed., *Probability, Statistics and Optimization* (John Wiley) pp. 241–253.

Teugels, J.L & Van Horebeek J. (1995), Algebraic descriptions of nominal discrete data. Submitted Paper.

Teugels, J.L & Van Horebeek J. (1996), *Generalized graphical models for discrete data.* Technical report I-96-16 (CIMAT, Guanajuato).

Van Horebeek J. (1994), *Het modelleren van nominale discrete data* (Doctoral dissertion, K.U. Leuven, Belgium).

Tables and Legends

		$X_3 = 0$	$X_3 = 1$
$X_1 = 0$	$X_2 = 0$	125	129
	$X_2 = 1$	85	31
$X_1 = 1$	$X_2 = 0$	216	61
	$X_2 = 1$	62	16

Table 1

1-factor	$\mu_0^1 = -\mu_1^1 = 0.147236(0.04991)$
	$\mu_0^2 = -\mu_1^2 = 0.54974(0.04991)$
	$\mu_0^3 = -\mu_1^3 = 0.44951(0.04991)$
2-factor	$\mu_{0,0}^{1,2} = -\mu_{0,1}^{1,2} = -\mu_{1,0}^{1,2} = \mu_{1,1}^{1,2} = -0.09687(0.04991)$
	$\mu_{0,0}^{1,3} = -\mu_{0,1}^{1,3} = -\mu_{1,0}^{1,3} = \mu_{1,1}^{1,3} = -0.20522(0.04991)$
	$\mu_{0,0}^{2,3} = -\mu_{0,1}^{2,3} = -\mu_{1,0}^{2,3} = \mu_{1,1}^{2,3} = -0.14129(0.04991)$
3-factor	$\mu_{0,0,0}^{1,2,3} = \cdots = -\mu_{1,1,1}^{1,2,3} = -0.11875(0.04991)$

Table 2

Figure 2: Block representation of a three dimensional binary variable

Figure 4:
Illustration of the effect of the above transformation for $C = \{X_3\}, x_3 = 0$.