Context Sensitive Regression Models with Binary Predictors

Johan Van Horebeek
Centro de Investigación en Matemáticas (CIMAT),
A.P. 402, Guanajuato, Gto., C.P. 36000, Mexico
horebeek@cimat.mx

Hugh Chipman
Department of Mathematics and Statistics, Acadia University,
Wolfville, NS, B4P 2R6, Canada

Summary

This paper considers a linear regression model with binary predictors from the point of view of graphical models. By considering second-order and higher order interactions as well as main effects, one is led naturally to consider the joint effects of predictors on the response. In particular, we introduce the context sensitive regression (CSR) model, in which the effect of a predictor on the response can depend on the level of another predictor. By representing such relations as a directed graph, a compact and interpretable modelling tool is developed. Parallels are identified between a graph-based representation of the linear model and hypothesis tests concerning equality of various coefficients to each other or to zero. For the two-way interaction model, we present a simple method to translate a graph into a set of constraints on a linear model. Extensions and generalizations to higher order models are also considered. Both frequentist and Bayesian methods for identifying CSRs are discussed. Finally, examples using both logistic and least squares regression are used to illustrate the model.
Key Words: Regression Models; Graphical Models; Binary predictors; Higher order interactions.

1 Introduction

The wide-spread use of regression models is partly attributable to their easy interpretation: the parameters provide quantitative information about the way predictors affect the response variable. As the inclusion of higher order interactions tends to obscure interpretability, particular configurations are often excluded (Peixoto (1990)) or penalized thorough the use of an underlying prior (Chipman et al. (1996)).

This is sometimes too restrictive: especially in the case of binary predictors, many models exist that involve higher order terms but still possess straightforward interpretations. As an illustration, consider the following linear regression model:

$$E(Y|X_1 = x_1, X_2 = x_2) = \alpha + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_{1,2} x_1 x_2, \quad x_i \in \{0, 1\}. \tag{1}$$

If $\alpha_{1,2} = -\alpha_1$, we have

$$E(Y|X_1 = x_1, X_2 = x_2) = \alpha + \alpha_1 x_1 (1 - x_2) + \alpha_2 x_2. \tag{2}$$

That is, given $x_2 = 1$, $E(Y|X)$ does not depend on x_1. This corresponds to the case where for one value of a predictor (e.g., for men) a second predictor does not affect the

response but for the other value (e.g., for women) it does. Apart of being an easily interpretable model involving higher order interactions, this example shows also that only testing which parameters might equal 0 overlooks interesting models.

In order to be able to deal with those situations, we introduce in this paper a class of regression models and follow an approach inspired by Graphical Models (Lauritzen (1996)). To this end, we introduce a class of graphs, each one representing a particular family of distributions characterized by a set of easy interpretable regularities in the underlying model. This will lead to a two step data analysis: first at the level of a graph and - only - afterwards at the level of a particular parameterization. In this way, this paper extends the results of Teugels et al. (1998), Højsgaard (2003) and Corander (2003) to regression models.

In the sequel $C(Y|X)$ denotes the property of interest of the set of conditional distributions of Y given the predictors $X = (X_1, \cdots, X_n)$; two important choices are $C(Y|X) = E(Y|X)$ and if Y is a binary variable $C(Y|X) = logit(P(Y = 1|X))$.
As X is a binary vector, without any loss of generality we can suppose that:

$$C(Y|X) = \alpha + \sum_i \alpha_i x_i + \sum_{i,j} \alpha_{i,j} x_i x_j + \cdots + \alpha_{1,\cdots,n} x_1 \cdots x_n. \tag{3}$$

A special case is the so called two-way models:

$$C(Y|X = x) = \alpha + \sum_i \alpha_i x_i + \sum_{i \neq j} \alpha_{i,j} x_i x_j. \tag{4}$$

Denoting by X_{-i} the vector X excluding the i-th entry, we write

$$Y \perp X_i | X_{-i} \tag{5}$$

if the predictor X_i does not *affect* $C(Y|X)$ whatever the values of the remaining predictors are; i.e., $C(Y|X_i = x_i, X_{-i} = x_{-i}) = C(Y|X_i = 1 - x_i, X_{-i} = x_{-i})$, for all x_{-i}.
If the predictor X_i does not affect $C(Y|X)$ only when $X_j = x_j$ (but for all $x_{-i,-j}$), we write:

$$Y \perp X_i | X_j = x_j, X_{-i,-j}. \tag{6}$$

The structure of the paper is as follows. In Section 2 the models are formally introduced. In Section 3 we discuss a case study. As model selection becomes an important issue, we elaborate in the final section, a Bayesian version that illustrates how apriori information can be naturally encoded through a prior on the graphs.

2 Context Sensitive Regression Models

2.1 Model Definition

We define a *context sensitive regression model* (CSR), over $(Y, X_1, \cdots, X_n)$ as the family of models for $C(Y|X)$ satisfying a set of constraints of the form (5) and (6). To avoid redundancy between (5) and (6), by writing $Y \perp X_i | X_j = x_j, X_{-i,-j}$ we implicitly assume that $Y \not\perp X_i | X_j = 1 - x_j, X_{-i,-j}$. When $Y \perp X_i$ given both $X_j = 1$ and $X_j = 0$ we write $Y \perp X_i | X_{-i}$.

With each model, a graph is associated: each node corresponds to a variable and connections are drawn between nodes associated with the predictors and the one associated with the response variable (or more precisely with $C(Y|X)$). The type of connection between the response and predictor X_i identifies the effect of X_i on the response:

1. **No dependence:** the *absence of a connection* encodes a constraint of the form (5).
2. **Partial dependence:** a *dashed connection* encodes a constraint of the form (6); an arrow starts from the predictor X_j and points to the dashed connection between Y and X_i; we call X_j the *controlling* predictor of X_i; we attach the corresponding value of $1 - x_j$ in (6) to the arrow indicating the presence of a relationship (or more precisely: the absence of an independency) between Y and X_i when $X_j = 1 - x_j$.
3. **Full dependence:** a *full connection* encodes the absence of any of the above constraints. In this case Y depends on X_j no matter what levels are assumed by X_{-j}.

We call two predictors *paired* if they control each other.

Example 2.1 *The graph in Figure 1 represents:*
$Y \perp X_1 | X_{-1}$, $Y \perp X_3 | X_4 = 0, X_{-3,-4}$, $Y \perp X_4 | X_3 = 1, X_{-3,-4}$, $Y \perp X_5 | X_6 = 0, X_{-5,-6}$, *and* $Y \perp X_7 | X_6 = 1, X_{-6,-7}$.
The variables X_3, X_4 *and* X_6 *are controlling predictors and* X_3, X_4 *are paired predictors.*

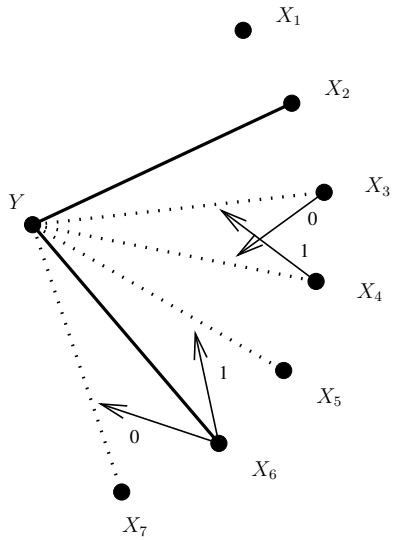


Figure 1

As will be shown in the next section, if we suppose a two-way parameterization of the form (4), the graph translates into:

$$C(Y|X) = \alpha + \alpha_2 x_2 + \alpha_6 x_6 + \alpha_{2,6} x_2 x_6 + \alpha_{3,4}(1 - x_3)x_4 + \alpha_{5,6} x_5 x_6 + \alpha_{6,7}(1 - x_6)x_7 \quad (7)$$

with α*'s free parameters.*

As in the case of Graphical Models, for a given graph those coefficients are set to zero or constrained in (4) so that the independencies reflected in the graph hold. On the other

hand, not every restriction on the coefficients can be reflected in the graph. For this reason we will work in two steps: first at the level of graphs and afterwards at the level of the parameters.

Opposite to a Graphical Model, a CSR defines a *set* of conditional distributions $P(Y|X)$ without revealing the joint distribution $P(Y, X)$. Therefore, questions related to collapsibility can not be answered; e.g., in the above graph: as nothing is specified about the interaction between X_1 and the other predictors, it is not possible to deduce whether $Y \perp X_1$ or not.

As will be explained below, an important subclass of CSR graphs are *regular* graphs.

Definition 2.1 *A graph is regular if it satisfies the following two conditions:*
(A1) at most one arrow arrives at each connection;
(A2) if a predictor X_j controls the connection between Y and X_i, there should be a dashed or full connection between Y and X_j and the node X_i is the only node that might control the connection between Y and X_j.

Observe that as a consequence, in a regular graph a non paired controlling predictor will always be connected to the response predictor with a solid connection.
Each regular graph $\mathcal{M}$ can be easily coded. Define

$$\delta_j(\mathcal{M}) = \begin{cases} i & \text{if } Y \perp X_j | X_i = 1, X_{-i,-j} \quad \text{(partial dependence)} \\ -i & \text{if } Y \perp X_j | X_i = 0, X_{-i,-j} \quad \text{(partial dependence)} \\ * & \text{if } Y \perp X_j | X_{-j} \quad \text{(no dependence)} \\ 0 & \text{otherwise} \quad \text{(full dependence)} \end{cases}$$

As (A1) holds for a regular CSR model, $\mathcal{M}$ will be completely specified by means of the n-tuple $(\delta_1(\mathcal{M}), \cdots, \delta_n(\mathcal{M}))$. E.g., the graph in Example 2.1 is encoded as $(*, 0, -4, 3, -6, 0, 6)$.

Under (A1), (A2) we can group the predictors into sets to obtain a partition with every set belonging to one of the following 4 types:

Type 1: Sets containing one predictor, not connected to the response variable;

Type 2: Sets containing one predictor, connected to the response variable with a full connection and that does not control any other predictor;

Type 3: Sets containing two predictors that are paired;

Type 4: Sets containing two or more predictors: a (unique) controlling predictor and all the predictors it controls.

We call $\mathcal{C}_i$ the set of all sets of type i and denote with c_i the number of sets in $\mathcal{C}_i$.

Example 2.2 *In Example 2.1, $\mathcal{C}_1 = \{\{X_1\}\}$, $\mathcal{C}_2 = \{\{X_2\}\}$, $\mathcal{C}_3 = \{\{X_3, X_4\}\}$ and $\mathcal{C}_4 = \{\{X_5, X_6, X_7\}\}$. Of course, in general several sets of the same type can be present as is the case in Figure 7.*

Consider now the reverse problem: given a graph with full and dashed connections, and arrows as described above, do the depicted independencies in the graph, define a CSR model with no other regularities of the form (5) and (6) ?

Contrary to Graphical Models, this will not always be the case. To this end, consider the graph in Figure 2:

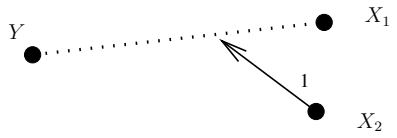


Figure 2

It reflects that:

$$Y \perp X_2 | X_1 \tag{8}$$
$$Y \perp X_1 | X_2 = 0 \tag{9}$$

As,

$$C(Y|X_1 = 0, X_2 = 1) \overset{(8)}{=} C(Y|X_1 = 0, X_2 = 0) \overset{(9)}{=} C(Y|X_1 = 1, X_2 = 0) \overset{(8)}{=} C(Y|X_1 = 1, X_2 = 1)$$

we obtain that $Y \perp X_1 | X_2 = 1$. This contradicts the dashed connection and the arrow of Figure 2. We call such a graph not *consistent* as the underlying models satisfy more independencies of the form (5) and (6) than the depicted ones. This and related aspects will be discussed in the next two subsections: first for the case of a two-way model, and afterwards for the general case.

2.2 The Two-Way Case

In this section we suppose that $C(Y|X)$ is of the form (4). Hence,

$$Y \perp X_i | X_{-i} \quad \text{iff} \quad \alpha_i = \alpha_{i,j} = 0, \ \forall j \neq i \tag{10}$$
$$Y \perp X_i | X_j = x_j, X_{-i,-j} \quad \text{iff} \quad \alpha_i + x_j \alpha_{i,j} = 0, \alpha_{i,k} = 0, \ \forall k \neq i, j. \tag{11}$$

As always, we suppose that all two-way interactions are in the model unless explicitly excluded. Under (4), it is easy to derive that:

$$Y \perp X_i | X_k = x_k, X_{-i,-k} \ \& \ Y \perp X_i | X_l = x_l, X_{-i,-l} \ \& \ k \neq l \ \Rightarrow Y \perp X_i | X_{-i} \tag{12}$$

$$Y \perp X_i | X_k = x_k, X_{-i,-k} \ \& \ Y \perp X_k | X_l = x_l, X_{-k,-l} \ \& \ i \neq l \Rightarrow Y \perp X_i | X_{-i} \tag{13}$$

We observe that (12) is equivalent to condition (A1) and (13) to (A2) in Definition 2.1. A complete characterization is provided by the following property whose proof is included in Appendix 1.

Property 2.1 *A graph represents a consistent two-way CSR model iff conditions (A1) and (A2) are satisfied, i.e., if it is a regular graph.*

A nice feature of these models is that each two-way CSR model - eventually after a suitable transformation of the predictors - is equivalent to a particular regression model without restrictions (like e.g. (11)) between the parameters. Therefore, classical regression estimation procedures can be used. To this end, rewrite (2) (i.e., (1) under $Y \perp X_1 | X_2 = 1$) as:

$$C(Y|X_2 = x_2, Z = z) = \alpha + \alpha_2 x_2 + \alpha_z z \text{ with } z = x_1(1 - x_2). \tag{14}$$

In the same way $Y \perp X_1 | X_2 = 0$ leads to

$$C(Y|X_2 = x_2, Z = z) = \alpha + \alpha_2 x_2 + \alpha_z z \text{ with } z = x_1 x_2. \tag{15}$$

In general, each two-way CSR model can be transformed into a model of the form:

$$C(Y|X_A = x_A, Z_B = Z_B) = \alpha + \sum_{i \in A} \alpha_i x_i + \sum_{(i,j) \in B} \alpha_{i,j} z_{i,j}, \tag{16}$$

with $\alpha_i, \alpha_{i,j}$ free parameters. In Appendix 2, we describe an algorithm to calculate the $z_{i,j}$'s and the sets A and B .

2.3 General Case

In analogy to Graphical Models, we discuss in this section the specification of $C(Y|X)$ of arbitrary form (3) by means of independencies of the form (5) and (6). We refer to them as *general* CSR models.

If we abbreviate $\alpha_{i,i_1,\cdots,i_k}$ by $x_{i,A}$ with $A = \{i_1, \cdots, i_k\}$, after some algebra, one can show that the equivalent to (11) for a general CSR model is:

$$Y \perp X_i | X_j = x_j, X_{-i,-j} \quad \text{iff} \quad \alpha_{i,A} + x_j \alpha_{i,j,A} = 0, \quad \forall A \subset \{1, \cdots, n\} \backslash \{i, j\}. \tag{17}$$

Contrary to the two-way case, a same parameter $\alpha_{i,j,A}$ can appear in several equations so that simultaneously a systems of equations of the form (17) should be solved to obtain an explicit parameterization.

Similar to the previous section, we can ask whether all graphs represent a consistent model. To this end, consider $\{(x_1, \cdots, x_n)\}$ as the corners of a n-dimensional hypercube. For a given independency $Y \perp X_i | X_{-i} = x_{-i}$, denote by $\mathcal{E}(Y \perp X_i | X_{-i} = x_{-i})$ the edge defined by the corners $(x_1, \cdots, x_{i-1}, 0, x_{i+1}, \cdots, x_n)$ and $(x_1, \cdots, x_{i-1}, 1, x_{i+1}, \cdots, x_n)$. See Figure 3 for the case $n = 3$.

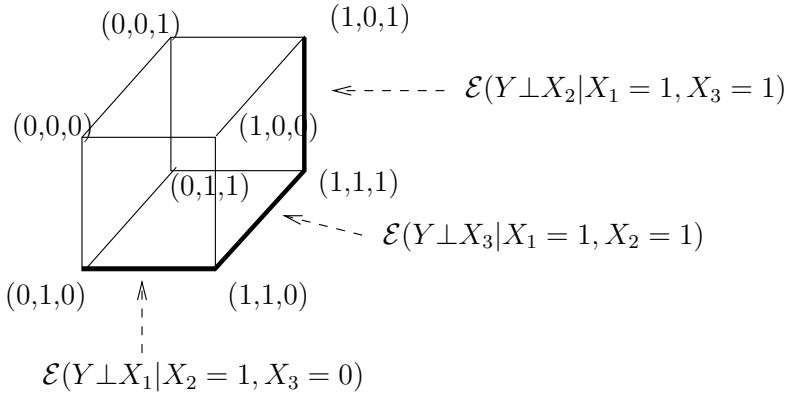


Figure 3

For a given set of independencies S, $\{Y \perp X_i | X_j = x_j, X_{-i,-j} = x_{-i,-j}\}$, we define a *trail* as a sequence of edges $\mathcal{E}(s), s \in S$, such that subsequent edges always have a corner in common. E.g., $Y \perp X_1 | X_2 = 1, X_3 = 0$, $Y \perp X_3 | X_1 = 1, X_2 = 1$, $Y \perp X_2 | X_1 = 1, X_3 = 1$ define a trail as shown in bold lines in Figure 3.

If we associate $C(Y|X = x)$ to each corner $x = (x_1, \cdots, x_n)$, it is easy to see that along a trail $C(Y|X)$ does not change. This leads to the following geometrical characterization:

Lemma 2.1 *A new independency is implied by a given set of independencies S iff the edge corresponding to the new independency is the starting and end point of a trail in S, i.e. they form together a loop.*

Example 2.3 *In Figure 2, $S = \{Y \perp X_2 | X_1 = 0\ , Y \perp X_2 | X_1 = 1\ , Y \perp X_1 | X_2 = 0\}$ implies $Y \perp X_1 | X_2 = 1$ because as shown in Figure 4 there is a trail that starts in $(0, 1)$ and ends in $(1, 1)$ and passes through $\mathcal{E}(Y \perp X_2 | X_1 = 0)$, $\mathcal{E}(Y \perp X_1 | X_2 = 0)$ and $\mathcal{E}(Y \perp X_2 | X_1 = 1)$. This means that the graph does not define a consistent model.*

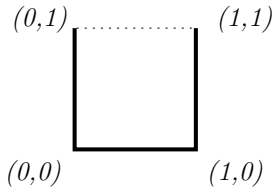


Figure 4

The above gives immediately an algorithm to verify whether a graph is valid or not. In general, it will not always be possible to see this at a first glance. To this end we show that (A2), which is easy to verify visually in the graph, defines a sufficient condition.

Property 2.2 *If a graph satisfies conditions (A2) then it defines a consistent general CSR model.*

Proof:
In the following we denote by $x_{-j}, x^2_{-j}, x^3_{-j}, \cdots$ different values of the vector X_{-j}.
Suppose that

$$Y \perp X_j | X_{-j} = x_{-j} \tag{18}$$

is implied by, but not included in the given set S. By Lemma 2.1, this means that there should exist a trail that starts and ends on the edge $\mathcal{E}(Y \perp X_j | X_{-j} = x_{-j})$. As this forms a loop, there should exist independencies in S that include:

$$Y \perp X_j | X_{-j} = x^2_{-j}. \tag{19}$$

As x^2_{-j} and x_{-j} differ in at least one position, there exists an index $i, i \neq j$ so that $x_i \neq x^2_i$. Hence under S,

$$Y \not\perp X_j | X_i = x_i, X_{-i-j} = x_{-i-j} \quad Y \perp X_j | X_i = 1 - x_i, X_{-i-j} = x^2_{-i-j}. \tag{20}$$

Again, as the trail forms a loop and the only way to *change* through the trail a value of a variable in the conditional part is by specifying an independency between that predictor and the response, for some $x^3_{-i,-j}, x^4_{-i,-j}$ the edges $\mathcal{E}(Y \perp X_i | X_j = 1, X_{-i,-j} = x^3_{-i,-j})$ and $\mathcal{E}(Y \perp X_i | X_j = 0, X_{-i,-j} = x^4_{-i,-j})$ should be part of the trail. This is shown in Figure 5 where the dotted lines refer to omitted parts of the trail.

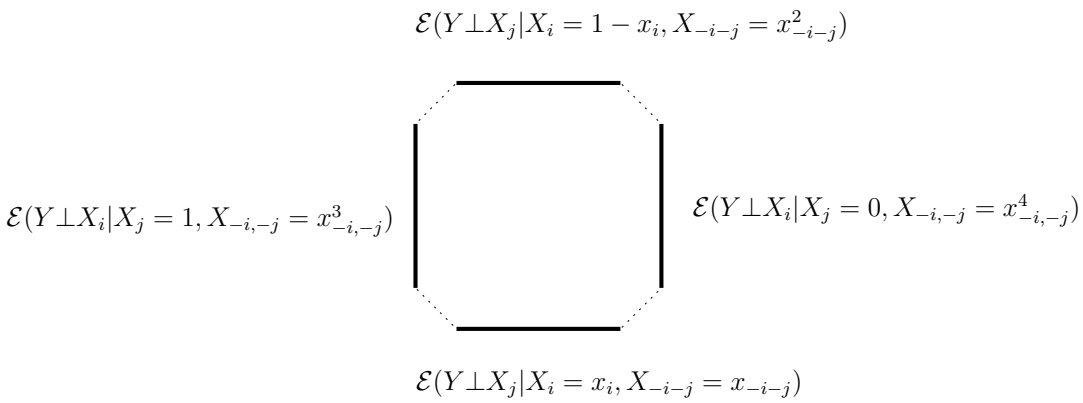


Figure 5

Hence,

$$Y \perp X_i | X_j = 1, X_{-i,-j} = x^3_{-i,-j} \quad Y \perp X_i | X_j = 0, X_{-i,-j} = x^4_{-i,-j}. \tag{21}$$

Given the fact that all independencies are of the form (5) or (6), (20) means that X_i is a controlling predictor for X_j; (21) implies that X_j is not a controlling predictor for X_i and that there is no full connection between Y and X_i. Hence (A2) is not satisfied.

•

3 Example

In this section we illustrate CSR models for the two-way case by means of the *Women and Mathematics* dataset as analyzed extensively in Fowlkes et al. (1988) (henceforward abbreviated as FFL). It concerns a study among 1190 New Jersey high school students about their attitude towards mathematics and the impact of a series of lectures to encourage interest in that area on it. The data are shown in Table 1.

As in FFL, the emphasis will be on methodological issues, without attempting to present a thorough analysis. The main point is to show how context sensitive regression models provide additional information about regularities in the underlying data, complementary to what a *classical* approach as followed in FFL provides.

To facilitate comparisons, we mimic the steps in the analysis of FFL; we also suppose that it is sufficient to look at two-way interaction models of the form (4) for this particular data set as they argue. The way the data were collected and the questions of interest, suggest (as FFL do) the use of an underlying logistic regression model, i.e. $C(Y|X) = logit(P(Y = 1|X))$.

	$X_3=0$				$X_3=1$			
	$X_2=0$		$X_2=1$		$X_2=0$		$X_2=1$	
	$X_1=0$	$X_1=1$	$X_1=0$	$X_1=1$	$X_1=0$	$X_1=1$	$X_1=0$	$X_1=1$
$X_5=0, X_4=0$								
$Y=0$	37	27	51	48	51	55	109	86
$Y=1$	16	11	10	19	24	28	21	25
$X_5=0, X_4=1$								
$Y=0$	16	15	7	6	32	34	30	31
$Y=1$	12	24	13	7	55	39	26	19
$X_5=1, X_4=0$								
$Y=0$	10	8	12	15	2	1	9	5
$Y=1$	9	4	8	9	8	9	4	5
$X_5=1, X_4=1$								
$Y=0$	7	10	7	3	5	2	1	3
$Y=1$	8	4	6	4	10	9	3	6

The variables are: Y whether the student agrees that they will need mathematics in the future (0=agree, 1=disagree); X_1 whether the student attended the lectures (0= yes, 1=no), X_2 his/her sex (0=female, 1=male), X_3 his/her type of school (0=suburban, 1=urban), X_4 his/her course preferences (0=mathematics, 1=liberal arts), X_5 his/her future plans (0=college, 1=job).

Table 1: The Women and Mathematics Dataset.

In Fig. 6 the black circles represent all possible two-way interaction models according to complexity (residual degrees of freedom) on the x-axis and goodness-of-fit ($\mathcal{G}^2$ likelihood-ratio statistic) on the y-axis. We marked explicitly those models identified by FFL as of particular interest, by writing on the left hand side of the corresponding circle, the terms involved in the corresponding two-way interaction model. As only hierarchical models are considered, lower order terms that appear in higher order terms are not mentioned explicitly: e.g., "2,34" means that x_2, x_3, x_4, x_3x_4 are present in the model. The full and (slightly curved) dashed line indicate for each d.f. the mean of $\mathcal{G}^2$ and the 95%-quantile of the distribution of $\mathcal{G}^2$.

The unfilled circles represent the counterpart of the above for all two-way CSR models, with a small shift on the horizontal axes to improve the visualization. Observe that we identify these models by the notation defined in Section 2.1. As there is very strong evidence that x_1 has no effect on the response, we limited ourselves to models without x_1 terms. We identified explicitly the most parsimonious models with residual degrees of freedom between 26 and 29, and showed the corresponding graphs in Fig. 7. One observes clearly some recurrent regularities, several of them are not obvious in the classical approach. E.g., evidence in favor of $Y \perp X_4 | X_5 = 1$ can be read off immediately of the graphs, opposed to the approach in FFL where for a given model, the oddsratio is calculated afterwards to obtain that information (cfr. figure 6 of FFL).

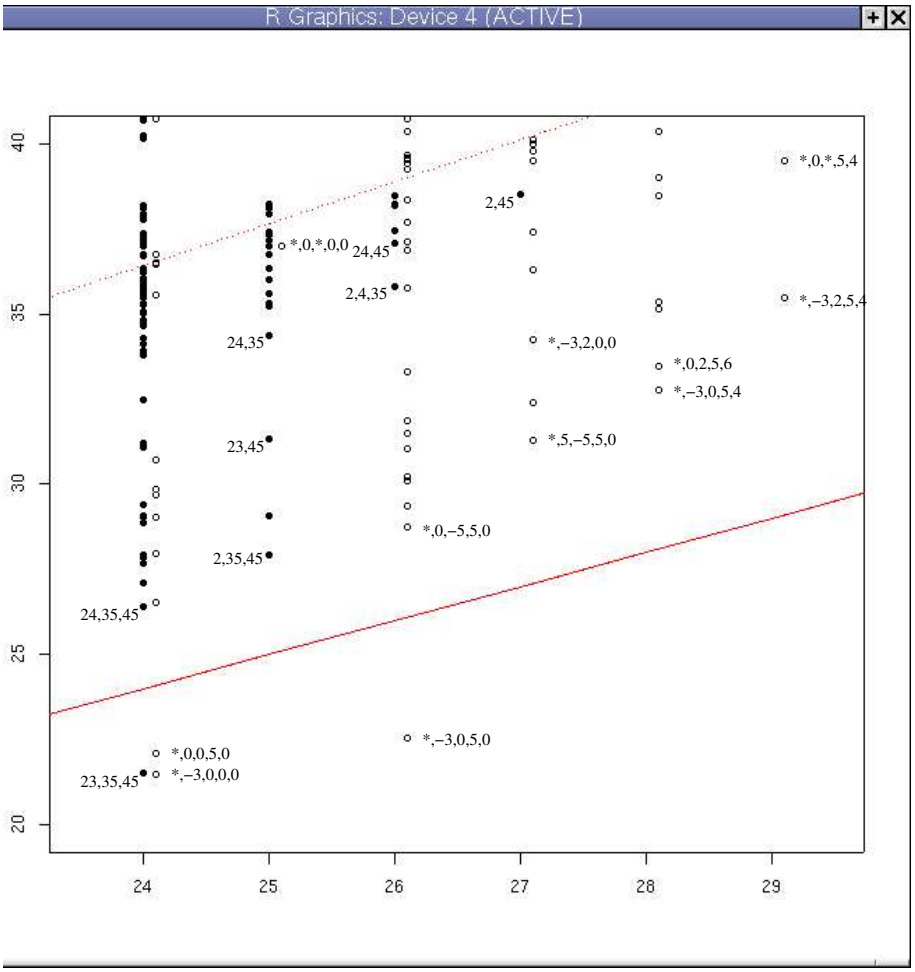


Figure 6

Following FFL, we look next at the residuals. As an illustration, Fig. 8 shows the qq plot of the residuals of the models $\mathcal{M}_7$ and $\mathcal{M}_8$, and -as a reference- the qq plots of the two promising models identified by FFL. All models seem to produce roughly normal residuals. Finally, we look at the (estimates of) the parameters of the models corresponding to $\mathcal{M}_7$ and $\mathcal{M}_8$ to verify whether further simplifications might be made (of course they will be no longer of the form (5) or (6)). Using (16) one gets (the s.e. is given between parentheses):

$$\mathcal{M}_7 : 1.01(0.13) - 0.28(0.16)x_3 - 0.69(0.22)x_5 + 0.76(0.16)z_{2,3} - 1.05(0.33)z_{3,5} - 1.16(0.14)z_{4,5},$$

with $z_{2,3} = x_2x_3$, $z_{3,5} = x_3x_5$ and $z_{4,5} = x_4(1 - x_5)$. As there is light evidence that the coefficient of x_3 is non significant different from 0, one might consider removing this term leading to a model with 27 d.f. and $\mathcal{G}^2 = 25.401$.

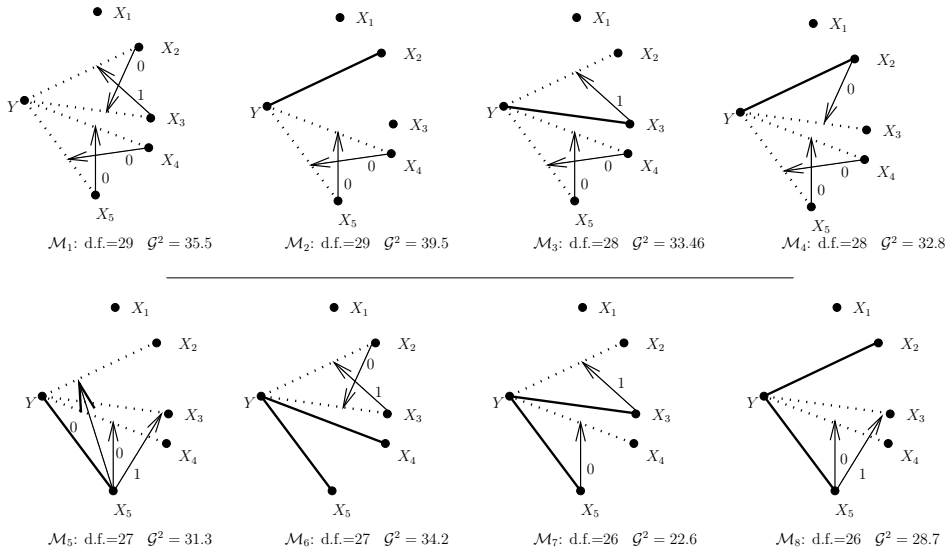
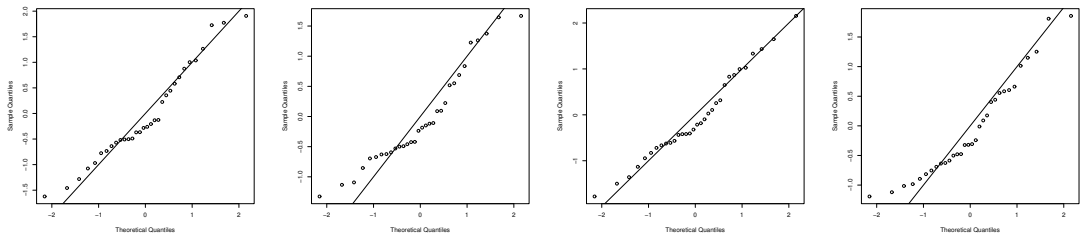


Figure 7

In a similar way:

$$\mathcal{M}_8 : 0.81(0.12)+0.52(0.14)x_2-0.71(0.26)x_5-0.06(0.32)z_{2,5}-0.96(0.30)z_{3,5}-1.13(0.14)z_{4,5},$$

with $z_{2,5} = x_2x_5$, $z_{3,5} = x_3x_5$ and $z_{4,5} = x_4(1-x_5)$. The coefficient of $z_{2,5}$ is not significantly different from 0, hence one might consider removing this term leading to a model with 27 d.f. and $\mathcal{G}^2 = 28.78$. Observe that this is a submodel of "2,35,45" that was identified by FFL as "promising".



(a) (b) (c) (d)

The qq-plots of (a) model 2,35,45 of FFL, (b) model 23,35,45 of FFL, (c) CSR model $\mathcal{M}_8 = (*, 0, -5, 5, 0)$ *and (d) CSR model* $\mathcal{M}_7 = (*, -3, 0, 5, 0)$.

Figure 8

4 A Bayesian Approach

In this section we use Bayesian methods to assign prior probabilities to graphs representing CSR models, and calculate the resultant posterior distributions. One of the main motivations for this approach is the ability to make uncertainty statements about *features* of graphs, rather than only about full graphs. For example, when looking at some of the

most promising CSR models for the *Women in Mathematics* data in Figure 7, one of the first things we try to do is to look for common features across the graphs. Are some predictors usually unconnected to the response? Do certain predictors always control others? Is X_5 a controlling predictor? Of course, some of these questions can be answered via a hypothesis test for the linear model that corresponds to the CSR. But others, such as the last question (is X_5 a controlling predictor) cut across many different CSR models, making traditional frequentist hypothesis testing impossible.

With this motivation, we outline some general issues in the specification of a prior on CSR models, before describing a specific prior and reporting on the Bayesian analysis of a quality-improvement dataset.

A prior distribution for CSR models must be specified for the graph, $\mathcal{M}$, and on the parameters Θ of the linear model associated with the graph. Since the dimensionality of Θ depends on $\mathcal{M}$, the usual factorization of the prior

$$\Pi(\mathcal{M}, \Theta) = \Pi(\mathcal{M})\Pi(\Theta|\mathcal{M})$$

seems especially relevant. In this section, the focus will be on prior $\Pi(\mathcal{M})$, since most of the inferential statements of interest will be in terms of the graph, rather than the associated parameters Θ. The prior distribution $\Pi(\mathcal{M})$ also must be tailored to the introduced class of models.

4.1 Graph Priors

Before outlining our prior of choice, we discuss three general strategies for prior construction: a graph-generating process, uniform distributions within equivalence classes of models and maximum entropy distributions that match desired marginal characteristics of the prior distribution.

In the context of Graphical Models, a popular approach for prior specification is by means of a set of independent Bernoulli variables each one indicating the presence or absence of a particular edge in the graph (Madigan et al. (1994)). As not every graph is a regular CSR graph and various types of edges are possible, the above idea has to be adapted. One way is by means of a graph-generating process: instead of independently deciding whether an edge is present, a sequential approach is followed where one steps through the $X_1, \cdots, X_n$ predictors and in each step one takes a decision about the graph at X_i, given the previous decisions.

So for example in Figure 1, one might first consider X_1: there would be a probability of a connection (in this case X_1 is unconnected). If there had been a connection, there would be a probability on what type (full or partial). If the connection were partial, we know that there would be exactly one controlling predictor. A prior distribution for this controlling predictor would be necessary. Lastly a prior distribution is needed on the level of the controlling predictor (0 or 1) which modifies the effect of X_1. In some ways, this approach of defining a process for constructing the graph is similar to the approach taken by Chipman et al. (1999) for putting a prior on trees. In the case of CSR models, however, the situation is more complicated: not only a natural ordering is missing, also implications of draws already made on the possible outcomes of a draw for characteristics

of the next variable is not always very straightforward. For example, if $X_1, \ldots, X_5$ had all been drawn in Figure 1, then X_6 must have some sort of connection with Y since it has been determined to be a controlling variable for X_5. Such interrelationships complicate both the specification of a prior, and in particular the computation of a prior probability for a specified graph.

A second approach would be to specify a prior distribution on some characteristics of the graph, and assume a uniform prior across all graphs with the same value of this characteristic. For example, one might specify a prior distribution on c_1, the number of connections that are absent between the X's and Y. This prior probability would be divided among all $\mathcal{M}$ with the same value of c_1 (i.e., we consider all models as members of an equivalence class). In order to calculate the probability associated with any particular model $\mathcal{M}_i$, it will be necessary to determine the number of elements in each equivalence class. Combinatorial arguments can be used to show that $N_n(c_1, c_2, c_3, c_4)$, the number of regular graphs with n predictors and with c_i sets of type i, is given by:

$$N_n(c_1, c_2, c_3, c_4) = c_4! 2^{n-c_1-c_2-c_3-c_4} \left(\begin{array}{c} n \\ c_1, c_2, c_3, c_4 \end{array} \right) \left\{ \begin{array}{c} n - c_1 - c_2 - 2c_3 - c_4 \\ c_4 \end{array} \right\} \quad (22)$$

where $\{\cdot\}$ denotes the Stirling number of the second kind (Riordan (1979)), c_1 is the number of unconnected nodes, c_2 is the number of fully connected nodes that do not control another predictor, c_3 is the number of pairs of nodes controlling each other, and c_4 is the number of nodes that have a full connection and control one or more other nodes. By summing out all elements of (22) over c_2, c_3, c_4, the size of an equivalence class determined by a particular c_1 value can be calculated.

A third approach to the specification of $\Pi(\mathcal{M})$ is to only specify some aspects of the prior in terms of specific characteristics of the graph, and then try to identify the maximum entropy prior that most closely matches these conditions (Jaynes (2003)). The computational task of finding a maximum entropy prior is difficult, and so this strategy is not pursued further.

Finally, observe that CSR models are sufficiently complicated objects that it may be difficult to definitively say, at the level of individual models, that one prior distribution better captures expert belief than another. We wish to stress that such priors may best be understood in terms of the prior distributions implied on various marginal characteristics of $\mathcal{M}$, rather than probabilities on individual $\mathcal{M}$. Thus, in the analysis presented below, we emphasize such marginal prior probabilities and the corresponding marginal posterior probabilities.

There is another reason for our reluctance to focus on prior and posterior probabilities at the level of individual $\mathcal{M}$. The problem of dilution (George (1999), Chipman et al. (2002)) may mean that although sensible priors can be specified for components of the model, some models may receive higher or lower mass depending on the number of similar models. With CSR models, for example, there will be more variations on a model with many paired controllers, since each controller could exert control at a 0 or 1 level. This complication suggests that the most important thing is that a prior make sense when collapsed onto marginal priors for quantities such as c_1, or probabilities that variables are controlled or are controllers, etc.

4.2 Example

We illustrate the Bayesian approach to CSR models with the *leafspring dataset* (Pignatiello et al. (1985)). The data are from an experiment on the manufacture of leaf springs for trucks. The response (Y) is the free height of the spring. Five binary predictors studied in the experiment: X_1 = High heat temperature (1840/1880° F), X_2 = Heating time (23/25 seconds), X_3 = Hold down time (2/3 seconds), X_4 = Quench oil temperature (130-150 / 150-170 ° F), and X_5 = Transfer time (10/12 seconds). 48 observations were recorded, in the form of three replicates at each level of a 16-run (2^{5-1}) fractional factorial experiment. The original dataset was aggregated over the non-critical predictor X_5 = *Transfer Time*, after preliminary analysis suggested that this factor did not influence the response. A full least squares model gave (the s.e. is given between parenthesis):

$$C(Y|X) = -0.15(0.06) + 0.15(0.07)x_1 + 0.3(0.07)x_2 + 0.09(0.07)x_3 - 0.2(0.07)x_4 +$$

$$0.034(0.07)x_1x_2 - 0.07(0.07)x_1x_3 + 0.17(0.07)x_1x_4 + 0.04(0.07)x_2x_3 - 0.33(0.07)x_2x_4 + 0.05(0.07)x_3x_4.$$

As discussed above, the equivalence class approach is used to specify $\mathcal{M}$ prior probabilities, in terms of c_1 the number of nodes with no connection to Y. Prior mass is equally redistributed within each equivalence class defined by a c_1 value. A natural choice for the distribution on c_1 is a Binomial distribution where the parameter p indicates the probability that a connection is absent. For $p = 0.4$, 0.5, and 0.6, the prior on $c_5 = 1 - c_1 = \#\{$ fully connected nodes in graph$\}$ is given in Table 2. We see that, intuitively, as the chance of no connection increases, more probability is put on small values of c_5. That is, fewer fully connected nodes are likely apriori. Multiple modes in the prior on c_5 are the result of constraints on the types of models that can possess exactly one connection. Other marginal priors will be mentioned below, after some posterior probabilities are given.

p	$P(C_5 = 0)$	$P(C_5 = 1)$	$P(C_5 = 2)$	$P(C_5 = 3)$	$P(C_5 = 4)$
0.4	0.0256	0.1536	0.3456	0.3456	0.1296
0.5	0.0625	0.25	0.375	0.25	0.0625
0.6	0.1296	0.3456	0.3456	0.1536	0.0256

Table 2: Prior probabilities on C_5 for different values of p.

To facilitate comparisons, we follow an approach similar to George et al. (1993) to define the remaining distributions:

$$Y|\beta_{\mathcal{M}}, \sigma^2, \mathcal{M} \sim \mathcal{N}(X\beta_{\mathcal{M}}, \sigma^2 I) \quad (23)$$

$$\beta_{\mathcal{M}}|\sigma^2, \mathcal{M} \sim \mathcal{N}(\bar{\beta}_{\mathcal{M}}, \alpha\sigma^2 I) \quad (24)$$

$$\sigma^2|\mathcal{M} \sim IG(\kappa/2, \kappa\lambda/2) \quad (25)$$

The prior on β is a mixture of a normal distribution and a point mass at 0, with the former corresponding to a significant coefficient, and the latter corresponding to an insignificant one. We made the neutral choice of $\bar{\beta}_{\mathcal{M}} = 0$. The prior on σ^2 is equivalent to $\kappa\lambda/\sigma^2 \sim \chi^2_\kappa$.

In addition to the specification of $\Pi(\mathcal{M})$, prior hyperparameters κ, α, λ must be specified. We take $\kappa = 5$, representing a prior equivalent to a sample with five residual degrees of freedom. By choosing $\lambda = 0.2$, we ensure that the mle $\hat{\sigma}^2 = 0.013$ under the full model is near the centre of the σ prior, and the upper tail is close to the sample standard deviation of Y. We choose α sufficiently large that the prior for significant effects is more dispersed than the empirical distribution of estimated effects. This leads to a choice of $\alpha = 4.0$.

model	posterior probability
(0, 4, *, 0)	0.41
(0, 4, 1, 0)	0.20
(0, 4, -4, 0)	0.15
(-4, 4, *, 0)	0.07
(0, 4, 0, 0)	0.04

Table 3: Models with the highest posterior probability for the leafspring dataset.

The problem is sufficiently small to explore the model space by complete enumeration. Table 3 gives some of the most probable models according to the posterior. It is presented primarily to motivate the use of marginal priors. This table suggests that X_2 is likely to be controlled by X_4. The posterior probability that X_4 is a controlling variable of any other variable is 0.96, compared to a prior probability of 0.24. The posterior probabilities that X_4 controls each of (X_1, X_2, X_3) are (0.11, 0.94, 0.20) respectively. These probabilities are not mutually exclusive, since in some models X_4 controls more than one of X_1, X_2, X_3. The corresponding prior probability is 0.09. Thus the data strongly support the hypothesis that X_4 controls X_2.

Observe that the fact that X_4 is very likely to control X_2 is also reflected by the almost horizontal line in the $X_2 - X_4$ interaction plot as shown in Figure 9.

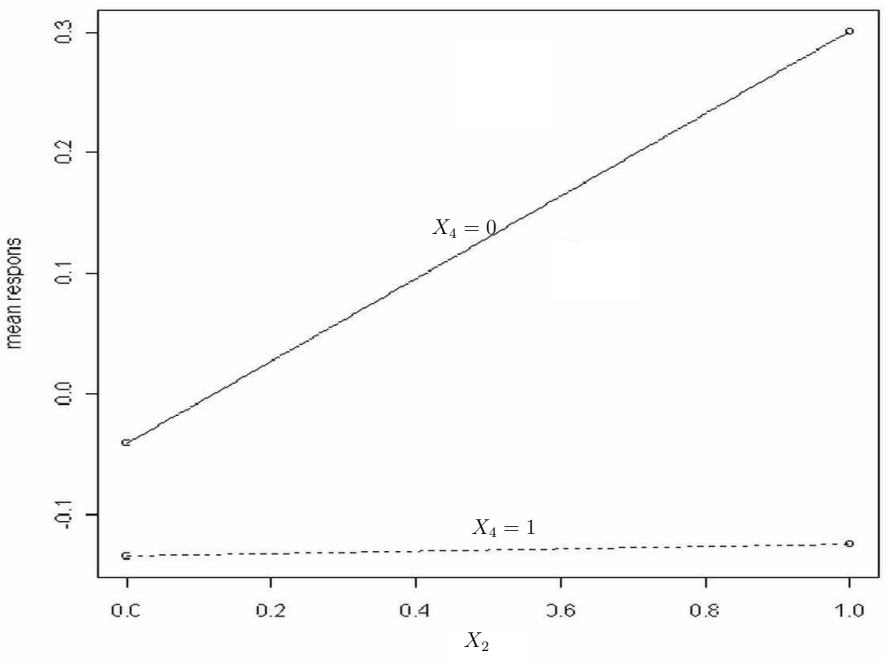


Figure 9

Discussion

In this paper, we described a framework to build regression models with binary predictors by means of easily interpretable (higher order) regularities. Such regularities can be overlooked if one starts with a (generalized) linear model parameterization. We showed how a useful subclass can be constructed for which the consistency and estimation problem can be solved in an efficient way.

Similar to Graphical Models, the proposed approach constitutes a complement to linear models by providing a compact and intuitive clear representation of the global interaction structure between predictors and response variable. From this point of view, several challenging questions come up. First of all about how to make the graphs more informative. For example, the thickness or color of the edges might provide additional degrees of freedom to encode the (kind of) support for a particular hypothesis. Secondly, about how to extend the models to incorporate information on the conditional distribution of the predictors as was done by block or chain models for the case of Graphical Models.

Acknowledgements

Part of the research was done while the first author was visiting the Department of Statistical and Actuarial Sciences of the University of Waterloo.

Appendix 1: Proof of Property 2.1

Proof:
Because of (12) and (13), conditions (A1) and (A2) should be necesarry. To show that they are also sufficient, we show that no independencies of the form (5) nor (6) can be implied by the independencies depicted in a graph satisfying (A1) and (A2),

First, suppose

$$Y \perp X_i | X_{-i} \tag{26}$$

is not depicted in the graph but implied by the independencies depicted in the graph.

Because of (10), in order that (26) holds, the independencies depicted in the graph should put constraints on the values of α_i and $\alpha_{i,\cdot}$. The only independencies which do have this effect on α_i, are of the form $Y \perp X_i | X_j = x_j, X_{-i,-j}$; at least one of them should hold. Suppose (without any loss of generality),

$$Y \perp X_i | X_j = 1, X_{-i,-j}. \tag{27}$$

At the same time, the independencies depicted in the graph, should imply that $\alpha_{i,j} = 0$. The only independencies that will constrain $\alpha_{i,j}$ are:

$$Y \perp X_i | X_k = x_k, X_{-i,-k} \tag{28}$$

or

$$Y \perp X_j | X_l = x_l, X_{-j,-l}. \tag{29}$$

The depicted independencies will satisfy (12); this means that (27) and (28) can only hold if $k = j$; in this case either $x_k = 1$, putting no additional constraint on $\alpha_{i,j}$, or $x_k = 0$, contradicting that (26) itself was not present in the graph.
In the same way, (27) and (29) can only hold if $i = l$, i.e., if i and j are paired predictors. It is easy to show that this does not imply that $\alpha_{i,j} = 0$.

The impossibility of an implied independency of the form (6) can be proven in the same way as (26).

•

Appendix 2: Transformation Algorithm

The following algorithm converts any CSR two-way model into a (classical) model of the form (16).

1. Define $A = \{1, \cdots, n\}$, $B = \{(i,j), i < j\}$ and $z_{i,j} = x_i x_j$, for all $i, j : i < j$.

2. For each set S of the partition corresponding to the given model, and not of type 2:

   2.1 if S is of type 1, hence $S = \{X_i\}$,
   set $A = A \setminus \{i\}$ and $B = B \setminus \{(i,j), \forall j \neq i\}$.

   2.2 if S is of type 3, hence $S = \{X_i, X_k\}$,
   - if $\delta_i(\mathcal{M}) = k$ or $\delta_k(\mathcal{M}) = i$, define:

$$z_{i,k} = I(\delta_i(\mathcal{M}) = k)x_i + I(\delta_k(\mathcal{M}) = i)x_k - x_i x_k \tag{30}$$

   - define $A = A \setminus \{i,k\}$ and $B = B \setminus (\{(i,j), \forall j \neq i,k\} \cup \{(k,j), j \neq i,k\})$.

   2.3 if S is of type 4, hence $S = \{X_{i_1}, \cdots, X_{i_r}, X_k\}$ with X_k the controlling predictor, for all i_l:
   - if $\delta_{i_l}(\mathcal{M}) = k$, define

$$z_{i_l,k} = x_{i_l} - x_{i_l} x_k \tag{31}$$

   - define $A = A \setminus \{i_l\}$ and $B = B \setminus \{(i_l, j), \forall j \neq i_l, k\}$.

It was this algorithm that was used in Example 2.1. As a motivation for the algorithm, consider e.g. step 2.2. Since the effect of each predictor on the response is governed by the other predictor, neither predictor can have a main effect and both main effects involving x_i and x_k are removed. Also, all interactions involving each of the two predictors are removed, with the exception of the $x_i x_k$ interaction. This is because if each predictor controls the other, then no other predictor can control either x_i or x_k. Thus all other interactions with these predictors are removed as well.

To illustrate the application of the algorithm, Table 4 shows the application of the algorithm to Example 2.1. The final model is then

$$C(Y|X) = \alpha + \alpha_2 + \alpha_6 + \alpha_{2,6}x_2x_6 + \alpha_{3,4}(1 - x_3)x_4 + \alpha_{5,6}x_5x_6 + \alpha_{6,7}(1 - x_6)x_7 \tag{32}$$

and coincides with (7).

Step	S	type	A	B
1	–	–	$\{1,\ldots,7\}$	$\{(1,2),\ldots,(6,7)\}$
2.1	$\{X_1\}$	1	$\{2,\ldots,7\}$	$\{(2,3),\ldots,(6,7)\}$
2.2	$\{X_3,X_4\}$	3	$\{2,5,6,7\}$	$\{(2,5),(2,6),(2,7),(3,4),(5,6),(5,7),(6,7)\}$
	$z_{3,4}=1x_4+0x_3-x_3x_4=x_4(1-x_3)$			
2.3	$\{X_5,X_7,X_6\}$	4	$\{2,6\}$	$\{(2,6),(3,4),(5,6),(6,7)\}$
	$z_{6,7}=1x_7-x_7x_6=x_7(1-x_6)$			

Table 4: Application of the two-way algorithm to example 2.1. New variables created by each step are listed in separate rows underneath the corresponding steps.

References

Chipman, H. (1996). Bayesian variable selection with related predictors, *The Canadian Journal of Statistics*, **24**, 17–36.

Chipman, H., George, E. & McCulloch, R. (1999). Bayesian CART Model Search (with discussion),*Journal of the American Statistical Association*, **93**, 935–960.

Chipman, H., George, E. & McCulloch, R. (2001). The Practical Implementation of Bayesian Model Selection, in *Model Selection*, IMS monograph, **38**, 65–116.

Corander, J. (2003). Labelled Graphical Models, *Scandinavian Journal of Statistics*, **30**, 493–508.

Fowlkes, E., Freeny, A. & Landwehr, J. (1988). Evaluating logistic models for large contingency tables, *Journal of the American Statistical Association*, **83**, 611–622.

George, E. & McCulloch, R. (1993). Variable Selection via Gibbs Sampling, *Journal of the American Statistical Association*, **88**, 881–889.

George, E. I. (1999). Discussion of "Bayesian Model Averaging: A Tutorial" by Hoeting, J.A., Madigan, D., Raftery, A. E., and Volinsky, C. T., *Statist. Sci.*, **14**, 401–404.

Højsgaard, S. (2003). Split models for contingency tables, *Computational Statistics and Data Analysis*, **42**, 621–645.

Jaynes, E. (2003). *Probability Theory : The Logic of Science*, Cambridge University Press.

Lauritzen, S. (1996). *Graphical models*, Oxford University Press.

Madigan, D., Raftery (1994). Model Selection and Accounting for Model Uncertainty in Graphical Models using Occam's Window, *Journal of the American Statistical Association*, **89**, 1535–1546.

Peixoto, J.L. (1990). A property of well-formulated polynomial regression models, *American Statistician*, **44**, 26–30.

Pignatiello, J.J. & Ramberg, J.S. (1985). Discussion of Off-Line Quality Control, Parameter Design and the Tagushi Method, *Journal of Quality Technology*, **17**, 198–206.

Riordan, J. (1979). *Combinatorial Identities*, Wiley.

Teugels, J. L. & Van Horebeek, J. (1998). Generalized Graphical Models, *Statistics & Probability Letters*, **38**, 41–47.