Exploiting Monotonicity via Logistic Regression in Bayesian Network Learning

Angelo C. Restificar Thomas G. Dietterich
School of Electrical Engineering and Computer Science
Oregon State University
Corvallis, OR 97331

Abstract

An important challenge in machine learning is to find ways of learning quickly from very small amounts of training data. The only way to learn from small data samples is to constrain the learning process by exploiting background knowledge. In this report, we present a theoretical analysis on the use of constrained logistic regression for estimating conditional probability distribution in Bayesian Networks (BN) by using background knowledge in the form of qualitative monotonicity statements. Such background knowledge is treated as a set of constraints on the parameters of a logistic function during training. Our goal of finding the appropriate BN model is two-fold: (a) we want to exploit any monotonic relationship between random variables that may generally exist as domain knowledge and (b) we want to be able to address the problem of estimating the conditional distribution of a random variable with a large number of parents. We discuss variants of the logistic regression model and present an analysis on the corresponding constraints required to implement monotonicity. More importantly, we outline the problem in some of these variants in terms of the number of parameters and constraints which, in some cases, can grow exponentially with the number of parent variables. To address this problem, we present two variants of the constrained logistic regression model, M^{2b}_{CLR} and M^{3}_{CLR}, in which the number of constraints required to implement monotonicity does not grow exponentially with the number of parents hence providing a practicable method for estimating conditional probabilities with very sparse data.

1 Introduction

Learning quickly from very small amounts of observed data is an important challenge in machine learning. Such need can be seen in many applications where the ratio of the number of observations to the number of variables is very low, for instance in modeling the transmission of new diseases (*e.g.*, the West Nile Virus) and in understanding the propagation of new computer worms. The only way to learn from small data samples is to constrain the learning process by exploiting background knowledge. The key is to identify kinds of background knowledge that are easy for

experts to specify and easy for algorithms to exploit. One example of such background knowledge is the causal relationships in a domain, which can be encoded in the graph structure of a Bayesian network. While causal relationships help constrain Bayesian network learning, they do not always provide enough constraint. The number of parameters that must be learned grows exponentially with the number of parents, and this means that very large amounts of data are needed.

In this report we focus on one kind of background knowledge, qualitative monotonicities, that is easy to specify and that can help address this problem. A qualitative monotonicity statement says that one variable increases (or decreases) monotonically as a function of another variable. Examples are "warmer temperatures increase the size of the mosquito population" and "the risk of having elevated blood pressure increases with a person's body mass index". The use of qualitative monotonicities in machine learning has been investigated in the past by several authors from different vantage points and incorporating them into the learning algorithm have been found to be useful [6, 7, 4, 5, 9, 10, 8, 14, 12]. The work that we describe here focuses on algorithms for learning from very small amounts of samples constrained by qualitative monotonicity knowledge formalized in terms of first-order stochastic dominance [16, 1].

Previous work by Altendorf, Restificar, and Dietterich [2] showed significant improvements in classification accuracy with very small amounts of training data (less than 10 examples) by exploiting qualitative monotonicities. However, when the number of parents N increases (*e.g.*, $N > 7$) the approach they reported suffers from two distinct disadvantages. The number of parameters that need to be estimated and the number of constraints on the parameters required to implement monotonicity both increase exponentially with N. The former could lead to underfitting hence producing models that give oversimplified hypotheses, while the latter could indirectly impose practical limits on the size of the problem that can be solved by a computer due to huge memory requirements. We propose to address these limitations by formulating the problem of exploiting qualitative monotonicities in Bayesian network parameter learning as a constrained logistic regression problem. This novel formulation affords us two major advantages. Parameter estimation can now be viewed as a regression problem over a set of parameters whose size only grows linearly with the number of parents. Moreover, the number of constraints required to exploit qualitative monotonicity can be shown to only grow linearly with the number of parents as well.

Clearly, not all constrained logistic regression models address the problems outlined above. This report provides an analysis on the kinds of practicable models that can be used. More specifically, we define and analyze variants of the constrained logistic regression model and demonstrate how qualitative monotonicities can be implemented in these models. We show that, in general, some of these models suffer from the need to estimate and solve an exponential number of parameters and constraints. Finally, we show that there exists variants of the logistic regression model in which the number of parameters needed to learn the conditional probability distributions in a Bayesian network as well as the number of constraints required to exploit qualitative monotonicities only grow linearly with the number of parents.

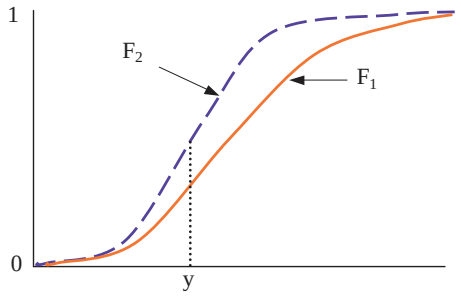

Figure 1: F_1 first-order stochastically dominates another cumulative disribution F_2.

2 Preliminaries

Given two random variables X and Y, we shall use the notation $X\ Q+\ Y$ (resp. $X\ Q-\ Y$) to mean 'there is a good chance that higher values of X result in higher (resp. lower) values of Y'. From time to time, we refer to $Q+$ (resp. $Q-$) as the qualitative influence of X on Y. It is also convenient to express monotonicity in terms of first-order stochastic dominance. If we have two cumulative distribution functions F_1 and F_2, the distribution F_1 first-order stochastically dominates F_2 whenever $F_1(y) \leq F_2(y)$ for any value y of a random variable Y. If we plot F_1 and F_2, it is the case that F_2 will be above F_1 for all values of y (see Figure 1). The reason why F_1 (the curve below) 'dominates' F_2 (the curve above) is because for all values of y, the probability of getting *at least* y is always greater under F_1 than under F_2. Note that in Figure 1, $F_1(y) < F_2(y)$ and so $1-F_1(y) \geq 1-F_2(y)$, thus F_1 'dominates' F_2.

Another way to characterize first-order stochastic dominance is through counts obtained from training examples. Suppose that we are given two random variables BP, which represents a person's risk of having elevated blood pressure, and BM, which represents a person's body mass index. Also, assume that from background knowledge BP is affected by BM and that both are binary random variables that can take on one of either values $high$ or low. The statement "the risk of having elevated blood pressure increases with a person's body mass index" translates to an expectation in our observation where the number of people expected to have low risk of elevated blood pressure would be *greater* than those expected to have $high$ risk just in the case where $BM = low$. Moreover, the number of people expected to have $high$ risk of elevated blood pressure would be *greater* than those expected to have low risk in the other case where $BM = high$. Figure 2 shows a typical scenario. It shows the counts (left table) that are associated with the conditional probability table or CPT of a Bayesian network (right table) with a target variable BP having a single parent BM. Note that in the case where $BM = low$ the bulk of the observed counts is on the upper left cell (100). However, when $BM = high$ the bulk of the observed counts shifts from the upper left cell to the lower right cell (105). The corresponding CPT has two conditional probability distributions, one for each row: $P_{high} = P(BP|BM = high)$ and $P_{low} = P(BP|BM = low)$. Recall that the conditional cumulative distribution can be written as $F_i(y) = \sum_{y'=low}^{y} P_i(BP = y'|BM = i)$, for $i = low, high$. The idea of first-order stochastic dominance simply states that for F_{high} to stochastically dominate F_{low}, the conditions in Figure 2 needs to be true, *i.e.*, $F_{low}(y) \geq F_{high}(y)\ \forall y$.

$Counts$	$BP = low$	$BP = high$
$BM = low$	100	10
$BM = high$	15	105

CPT	$BP = low$	$BP = high$
$BM = low$	0.91	0.09
$BM = high$	0.125	0.875

Figure 2: Example of a monotonic relationship between random variables BM and BP.

From Figure 2, $F_{low}(low) = 0.91$ and $F_{high}(low) = 0.125$. In addition, $F_{low}(high) = 1.0$ and $F_{high}(high) = 1.0$. Hence, it is clear that the conditional cumulative distributions satisfy first-order stochastic dominance. Let us denote first-order stochastic dominance by **FSD**. Formally,

Definition 1 (First-Order Stochastic Dominance (FSD)) *Given two cumulative distributions F_1 and F_2,*

$$F_1 \textbf{ FSD } F_2 \quad \textit{iff} \quad \forall y \; F_1(y) \leq F_2(y) \tag{1}$$

For a multi-valued parent X, we define monotonicity as satisfying a set of constraints on the cumulative distributions for any pair of configurations x_i, x_j of X such that $x_i \geq x_j$.

Definition 2 (FSD Monotonicity) *Let X and Y be random variables in a Bayesian network where Y is the child (target) of a single parent variable X. Y is* **FSD** *monotonic in X if*

$$F_{x_i}(Y \mid X = x_i) \textbf{ FSD } F_{x_j}(Y \mid X = x_j) \;\; \forall \, x_i, x_j, \; x_i \geq x_j \tag{2}$$

In the case where the number of parents is more than one we define FSD monotonicity in the *ceteris paribus* sense, *i.e.*, all other variable assignments being equal. Suppose Y has multiple parents $X_1, X_2, \ldots, X_q$. Y is FSD monotonic in X_i if and only if Eq. (2) holds when all the variable assignments for $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_q$ are held fixed. The term *isotonic* refers to a positive monotonic relationship, *i.e.*, Y increases as X increases, and the term *anti-tonic* refers to a negative monotonic relationship, *i.e.*, Y decreases as X increases. When no confusion will arise we will use the term monotonic to refer to either relationship.

3 Constrained Logistic Regression Models

In this section, we present variants of constrained regression models with varying degrees of complexity. Here we hope to show that not all variants and models of logistic regression are ideally suited in addressing the problem of estimating conditional probabilities of nodes with a large number of parents especially in the context of very small training set sizes. Such remark is based on the number of constraints that each model requires to exploit monotonicity between random variables. Figure 3 shows a specific example of the set of constraints needed to implement monotonicity for a Bayesian network with a ternary target Y and a ternary parent X. It is clear that as the number of parents N increases the number of configurations and the number of constraints required to exploit monotonocity also increase exponentially. For instance, without finding a way to reduce

$P(Y \mid X)$	$Y=0$	$Y=1$	$Y=2$
$X=0$	θ_0	θ_3	$1-\theta_0-\theta_3$
$X=1$	θ_1	θ_4	$1-\theta_1-\theta_4$
$X=2$	θ_2	θ_5	$1-\theta_2-\theta_5$

$$\begin{array}{c}\text{Constraints :}\\ \theta_0 \geq \theta_1\\ \theta_1 \geq \theta_2\\ \theta_0+\theta_3 \geq \theta_1+\theta_4\\ \theta_1+\theta_4 \geq \theta_2+\theta_5\end{array}$$

Figure 3: CPT parameters (left table) for ternary variables X (parent) and Y (target). The set of constraints (right table) that are required to implement monotonicity between X and Y.

the number of parameters, the total number of parameters to be fitted is exponential in N, *i.e.*, $(|Y|-1)\prod_{i=1}^{N}|X_i|$. In the simplest case, if there are n binary parents of a binary target the set of constraints can be depicted as a hypercube whose nodes are parent configurations and in which any pair of configurations that differ only in one position is connected by an edge. The number of edges in the hypercube *i.e.,* the number of constraints, is $n2^{n-1}$. We now present a way to reduce the number of parameters while at the same time respecting the required monotonicity constraints. We propose to estimate the parameters using a more compact representation, a logistic function.

3.1 Constrained Logistic Regression Model 1 (M^1_{CLR})

In the first model, the conditional probability distribution is estimated by using the parents (as opposed to parent levels) of the target random variable. In this model, k_Y (equal to the number of levels of the target variable) logistic functions are used and normalized exponentiation is applied to calculate the conditional probability distribution given a parent configuration.

Definition 3 (M^1_{CLR}) *Let Y be a child variable with k_Y levels*[1]*, $k_Y > 2$, in a Bayesian network BN with parents $X_1, \ldots, X_n$. Assume that k_Y is indexed as $0, \ldots, k_Y-1$. Given some configuration $c = \langle x_1, \ldots, x_n \rangle$, define*

$$P(Y=j|c) = \frac{\exp\{\beta_{0j} + \beta_{1j}X_1 + \cdots + \beta_{nj}X_n\}}{\sum_{j'=0}^{k_Y-1}\exp\{\beta_{0j'} + \beta_{1j'}X_1 + \cdots + \beta_{nj'}X_n\}} \tag{3}$$

Also, let the following set of constraints hold:

1. *For $i = 1, \ldots, n$ and $j = 0, \ldots, k_Y-1$*

$$\beta_{ij}^{X_i} \geq 0 \tag{4}$$

2. *For each pair of configuration $c_1 = \langle x_1, \ldots, x_q, \ldots, x_n \rangle$ and $c_2 = \langle x_1, \ldots, x'_q, \ldots, x_n \rangle$ such*

[1]The simpler binary cases can be handled with just one logistic function.

that $x'_q > x_q$*, where* $1 \leq q \leq n$ *and for* $\forall j, j = 0, \ldots, k_Y - 1$

$$\frac{\sum_{j'=0}^{j} \exp\{\beta_{0j'} + \beta_{1j'}x_1 + \cdots + \beta_{qj'}x_q + \cdots + \beta_{nj'}x_n\}}{\sum_{j''=0}^{k_Y-1} \exp\{\beta_{0j''} + \beta_{1j''}x_1 + \cdots + \beta_{qj''}x_q + \cdots + \beta_{nj''}x_n\}} \geq \frac{\sum_{j'=0}^{j} \exp\{\beta_{0j'} + \beta_{1j'}x_1 + \cdots + \beta_{qj'}x'_q + \cdots + \beta_{nj'}x_n\}}{\sum_{j''=0}^{k_Y-1} \exp\{\beta_{0j''} + \beta_{1j''}x_1 + \cdots + \beta_{qj''}x'_q + \cdots + \beta_{nj''}x_n\}} \tag{5}$$

Theorem 1 *Let* Y *be a child variable with* k_Y *levels,* $k_Y > 2$*, in a Bayesian network BN with parents* $X_1, \ldots, X_n$*. If* M^1_{CLR} *is the constrained logistic regression model,* $Q+$ *the qualitative influence of each* X_i *on* Y*,* $i = 1, \ldots, n$ *then* Y *is* **FSD** *monotonic in* X_i*.*

To prove the theorem, we need to show the following:

(a) Let $\theta_{c_m j}$ denote a CPT parameter for configuration c_m and for class j of Y*, i.e.,* $Y = j$ in the CPT table. Then

$$0 \leq \theta_{c_m j} \leq 1 \tag{6}$$

(b) For any configuration c_m*,*

$$\sum_{j=0}^{k_Y-1} \theta_{c_m j} \;\; = \;\; 1 \tag{7}$$

(c) For any pair of configurations c_1, c_2 such that $c_2 > c_1$ *ceteris paribus,*

$$F_{c_2}(j) \leq F_{c_1}(j) \tag{8}$$

for $j = 0, \ldots, k_Y - 1$.

The proof is trivial. Condition **(a)** above follows immediately from the definition of sigmoid logistic functions and Condition **(b)** follows from the definition of normalized exponentiation. Condition **(c)** follows immediately from Eq. (5).

The number of constraints N^1_C for this model is exponential and is equal to $(k_Y - 1)$ times the number of edges in a lattice formed by mapping a configuration to a node in the lattice where nodes that differ only in one position are connected by an edge. In the simplest case, if there are n parents and each parent is a binary random variable, the lattice is an n-hypercube where the number of edges is $n2^{n-1}$. The number of constraints therefore is $(k_Y - 1)n2^{n-1}$ and hence exponential in the number of parents. However, the number of β parameters for model M^1_{CLR}, N^1_β, is $k_Y(1 + n)$. Hence, the number of parameters we need to fit only grows linearly in the number of parents.

3.2 Constrained Logistic Regression Model 2 (M^2_{CLR})

Unlike the first model, the second model offers more representational flexibility by estimating the conditional probability distribution using the levels of the parents of the target random variable instead of associating only a single parameter for each parent. Here, we estimate the conditional distribution by using $k_Y - 1$ logistic functions. The conditional probabilities of the child random variable Y are obtained by subtracting adjacent levels of Y.

Definition 4 (M^2_{CLR}) *Let Y be a child variable with k_Y levels, $k_Y \geq 2$, in a Bayesian network BN with parents $X_1, \ldots, X_n$ where each X_i has k_{X_i} levels, $k_{X_i} \geq 2$ for $i = 1, \ldots, n$. Assume that k_Y is indexed as $0, \ldots, k_Y - 1$. Given some configuration $c = \langle x_1, \ldots, x_n \rangle$, define*

$$\begin{aligned}\log \frac{P(Y > j|c)}{P(Y \leq j|c)} = \beta_{0j} + \beta^{X_1}_{1j} I[X_1 \geq 1] + \cdots + \beta^{X_1}_{k_{X_1} j} I[X_1 \geq k_{X_1}] + \cdots + \\ \beta^{X_n}_{1j} I[X_n \geq 1] + \cdots + \beta^{X_n}_{k_{X_n} j} I[X_n \geq k_{X_n}] \text{ for } j = 0, \ldots, k_Y - 2 \end{aligned} \tag{9}$$

Also, let the following set of constraints hold

1. *For $i = 1, \ldots, n$, $r = 1, \ldots, k_{X_i}$, and $j = 0, \ldots, k_Y - 2$*

$$\beta^{X_i}_{rj} \geq 0 \tag{10}$$

2. *For each configuration $\langle x_1, \ldots, x_n \rangle$ and j, $j = 1, \ldots, k_Y - 2$*

$$\begin{aligned}\beta_{0(j-1)} + \beta^{X_1}_{1(j-1)} I[X_1 \geq 1] + \cdots + \beta^{X_1}_{k_{X_1}(j-1)} I[X_1 \geq k_{X_1}] + \ldots + \\ \beta^{X_n}_{1(j-1)} I[X_n \geq 1] + \cdots + \beta^{X_n}_{k_{X_n}(j-1)} I[X_n \geq k_{X_n}] \geq \\ \beta_{0j} + \beta^{X_1}_{1j} I[X_1 \geq 1] + \cdots + \beta^{X_1}_{k_{X_i} j} I[X_1 \geq k_{X_1}] + \ldots + \\ \beta^{X_n}_{1j} I[X_n \geq 1] + \cdots + \beta^{X_n}_{k_{X_n} j} I[X_n \geq k_{X_n}] \end{aligned} \tag{11}$$

Definition 4 expresses the log odds ratio between the cumulative conditional probabilities $P(Y > j|c)$ and $P(Y \leq j|c)$ as a linear function of the levels of the parent variables $X_1, \ldots, X_n$. Such formulation allows one to model the contribution of each level of X_i to the log odds ratio via the β parameters. As will be shown below, domain knowledge about monotonic qualitative influences can be exploited to speed up Bayesian network learning by imposing constraints on these parameters and then solving the corresponding optimization problem. In addition, synergistic and anti-synergistic influences between different parents can very well be modeled by adding extra terms. For instance, if we want to model a synergistic interaction in a Bayesian network with two ternary-valued parents X_1 and X_2 whenever $X_1 \geq 1$ and $X_2 \geq 2$, then such can be expressed as

$$\begin{aligned}\log \frac{P(Y > j|c)}{P(Y \leq j|c)} = \beta_{0j} + \beta^{X_1}_{1j} I[X_1 \geq 1] + \beta^{X_1}_{2j} I[X_1 \geq 2] + \beta^{X_2}_{1j} I[X_2 \geq 1] + \beta^{X_2}_{2j} I[X_2 \geq 2] + \\ \beta^{X_1 X_2}_{1j} I[X_1 \geq 1] I[X_2 \geq 2] \end{aligned} \tag{12}$$

For a child variable with k_Y levels, we only need to fit $k_Y - 1$ logistic functions, *i.e.,* one less than the number of k_Y levels since the last CPT column can be obtained by subtracting the cumulative probability at $j = (k_Y - 2)$ from 1. The number of parameters needed to be fitted in this model is linear in the total number of levels of the parents. In particular, if N^2_β denote the number of parameters then $N^2_\beta = (k_Y - 1)[1 + \sum_{i=1}^{n}(k_{X_i} - 1)]$. However, as in M^1_{CLR} the number of constraints is still exponential in the number of parents. If N^2_C denote the number of constraints for model M^2_{CLR} then $N^2_C = (k_Y - 1)\sum_{i=1}^{n}(k_{X_i} - 1) + (k_Y - 2)\prod_{i=1}^{n} k_{X_i}$. The first term is the total number of constraints from Eq. (10). The second term sums up all the constraints for all configurations as stated in Eq. (11).

Theorem 2 *Let Y be a child variable with k_Y levels, $k_Y > 2$, in a Bayesian network BN with parents $X_1, \ldots, X_n$ where each X_i has k_{X_i} levels, $i = 1, \ldots, n$. If M^2_{CLR} is the constrained logistic regression model, $Q+$ the qualitative influence of each X_i on Y, then Y is* **FSD** *monotonic in X_i.*

To prove the theorem, we need to show the following:

(a) Let $\theta_{c_m j}$ denote a CPT parameter for configuration c_m and for class j of Y, *i.e.,* $Y = j$ in the CPT table. Then

$$0 \leq \theta_{c_m j} \leq 1 \tag{13}$$

(b) For any configuration c_m,

$$\sum_{j=0}^{k_Y - 1} \theta_{c_m j} \;\; = \;\; 1 \tag{14}$$

(c) For any pair of configurations c_1, c_2 such that $c_2 > c_1$ *ceteris paribus,*

$$F_{c_2}(j) \leq F_{c_1}(j) \tag{15}$$

for $j = 0, \ldots, k_Y - 1$.

Given a configuration $c_m = \langle x_{i1}, \ldots, x_{in} \rangle$, the log odds ratio as expressed in our model is

$$\begin{aligned} \log \frac{P(Y > j|c_m)}{P(Y \leq j|c_m)} = {} & \beta_{0j} + \beta^{X_1}_{1j} I[X_1 \geq 1] + \cdots + \beta^{X_1}_{k_{X_1} j} I[X_1 \geq k_{X_1}] + \cdots + \\ & \beta^{X_n}_{1j} I[X_n \geq 1] + \cdots + \beta^{X_n}_{k_{X_n} j} I[X_n \geq k_{X_n}] \text{ for } j = 0, \ldots, k_Y - 2 \end{aligned} \tag{16}$$

Let $c_1 = \langle x_{11}, \ldots, x_{1q}, \ldots, x_{1n} \rangle$ and $c_2 = \langle x_{21}, \ldots, x_{2q}, \ldots, x_{2n} \rangle$ be a pair of configurations such that $c_1 < c_2$ *ceteris paribus, i.e.,* they differ only at a single position q so that $x_{1q} < x_{2q}$. Let

$P(Y=j|c_m)=\theta_{c_m j}$. We want to show that for all values j, $j=0,\ldots,k_Y-2$, it is the case that $F_{c_2}(j)\leq F_{c_1}(j)$. Assuming $j=0$, from Eq. (16) we have for $i=1,2$

$$\log\frac{P(Y>0|c_m)}{P(Y\leq 0|c_m)}=\beta_{00}+\beta_{10}^{X_1}I[X_1\geq 1]+\cdots+\beta_{k_{X_1}0}^{X_1}I[X_1\geq k_{X_1}]+\cdots+ \beta_{10}^{X_n}I[X_n\geq 1]+\cdots+\beta_{k_{X_n}0}^{X_n}I[X_n\geq k_{X_n}] \tag{17}$$

$$\frac{1-\theta_{c_m 0}}{\theta_{c_m 0}}=\exp\{\beta_{00}+\beta_{10}^{X_1}I[X_1\geq 1]+\cdots+\beta_{k_{X_1}0}^{X_1}I[X_1\geq k_{X_1}]+\cdots+ \beta_{10}^{X_n}I[X_n\geq 1]+\cdots+\beta_{k_{X_n}0}^{X_n}I[X_n\geq k_{X_n}]\} \tag{18}$$

Since $c_2>c_1$, Eq. (18) implies that

$$\theta_{c_1 0}=\frac{1}{1+\exp\{\beta_{00}+\beta_{10}^{X_1}+\cdots+\beta_{k_1 0}^{X_1}+\cdots+\beta_{10}^{X_q}+\cdots+\beta_{k_q 0}^{X_q}+\cdots+\beta_{10}^{X_n}+\cdots+\beta_{k_n 0}^{X_n}\}} \tag{19}$$

$$\theta_{c_2 0}=\frac{1}{1+\exp\{\beta_{00}+\beta_{10}^{X_1}+\cdots+\beta_{k_1 0}^{X_1}+\cdots+\beta_{10}^{X_q}+\cdots+\beta_{k_q 0}^{X_q}+\beta_{(k+1)_q 0}^{X_q}+\cdots+\beta_{10}^{X_n}+\cdots+\beta_{k_n 0}^{X_n}\}} \tag{20}$$

for some $(k_1\leq k_{X_1}),\ldots,(k_q\leq k_{X_q}),((k+1)_q\leq k_{X_q}),\ldots,(k_n\leq k_{X_n})$. Since Eq. (10) holds, it is clear from Eqs. (19) and (20) that $\theta_{c_1 0}\geq\theta_{c_2 0}$.

Now, suppose that $0<j\leq(k_Y-2)$. From Eq. (16),

$$\frac{1-\sum_{j'=0}^{j}\theta_{c_m j'}}{\sum_{j'=0}^{j}\theta_{c_m j'}}=\exp\{\beta_{0j}+\beta_{1j}^{X_1}I[X_1\geq 1]+\cdots+\beta_{k_{X_1}j}^{X_1}I[X_1\geq k_{X_1}]+\cdots+ \beta_{1j}^{X_n}I[X_n\geq 1]+\cdots+\beta_{k_{X_n}j}^{X_n}I[X_n\geq k_{X_n}]\} \tag{21}$$

Since $c_2>c_1$,

$$\sum_{j'=0}^{j}\theta_{c_1 j'}=\frac{1}{1+\exp\{\beta_{0j}+\beta_{1j}^{X_1}+\cdots+\beta_{k_1 j}^{X_1}+\cdots+\beta_{1j}^{X_q}+\cdots+\beta_{k_q j}^{X_q}+\cdots+\beta_{1j}^{X_n}+\cdots+\beta_{k_n j}^{X_n}\}} \tag{22}$$

and

$$\sum_{j'=0}^{j}\theta_{c_2 j'}=\frac{1}{1+\exp\{\beta_{0j}+\beta_{1j}^{X_1}+\cdots+\beta_{k_1 j}^{X_1}+\cdots+\beta_{1j}^{X_q}+\cdots+\beta_{k_q j}^{X_q}+\beta_{(k+1)_q j}^{X_q}+\cdots+\beta_{1j}^{X_n}+\cdots+\beta_{k_n j}^{X_n}\}} \tag{23}$$

Again, from Eqs. (22) and (23)

$$\sum_{j'=0}^{j}\theta_{c_1 j'}\geq\sum_{j'=0}^{j}\theta_{c_2 j'}$$

since Eq. (10) holds. By definition,

$$F_{c_2}(j) \leq F_{c_1}(j)$$

holds for $j = 0, \ldots, k_Y - 2$. Note, however, that once we have shown Eq. (14) to hold, we will have shown that Eq. (15) also holds, *i.e.,* for all levels of Y.

Observe that

$$\sum_{j'=0}^{j} \theta_{c_m j'} = \frac{1}{1+\exp\{\beta_{0j}+\beta_{1j}^{X_1}+\cdots+\beta_{k_1 j}^{X_1}+\cdots+\beta_{1j}^{X_q}+\cdots+\beta_{k_q j}^{X_q}+\cdots+\beta_{1j}^{X_n}+\cdots+\beta_{k_n j}^{X_n}\}} \tag{24}$$

and that

$$\theta_{c_m j} = \sum_{j'=0}^{j+1} \theta_{c_m j'} - \sum_{j''=0}^{j} \theta_{c_m j''} \tag{25}$$

Hence,

$$\begin{aligned}\sum_{j=0}^{k_Y-1} \theta_{c_m j} &= \theta_{c_m 0} + [(\theta_{c_m 0} + \theta_{c_m 1}) - \theta_{c_m 0}] + \ldots + \left[\sum_{j'=0}^{j} \theta_{c_m j'} - \sum_{j'=0}^{j-1} \theta_{c_m j'}\right] \\ &\quad + \left[\sum_{j'=0}^{j+1} \theta_{c_m j'} - \sum_{j'=0}^{j} \theta_{c_m j'}\right] + \ldots + \left[\sum_{j'=0}^{k_Y-2} \theta_{c_m j'} - \sum_{j'=0}^{k_Y-3} \theta_{c_m j'}\right] + \left[1 - \sum_{j'=0}^{k_Y-2} \theta_{c_m j'}\right] \qquad (26) \\ &= 1 \qquad (27)\end{aligned}$$

Since every term in the RHS of Eq. (26) cancels out except the penultimate term 1, Eq. (14) holds, and therefore Eq. (15) follows.

Now, we need to show that each $\theta_{c_m j} \in [0, 1]$. Suppose $j = 0$ then

$$\sum_{j'=0}^{j} \theta_{c_m j'} = \theta_{c_m 0} = \frac{1}{1+\exp\{\beta_{0j}+\beta_{1j}^{X_1}+\cdots+\beta_{k_1 j}^{X_1}+\cdots+\beta_{1j}^{X_n}+\cdots+\beta_{k_n j}^{X_n}\}} \tag{28}$$

Since Eq. (28) is a logistic sigmoid function, $\theta_{c_m 0} \in [0, 1]$. Now, suppose that $0 < j \leq k_Y - 1$, then

$$\begin{aligned}\theta_{c_m j} &= \sum_{j'=0}^{j+1} \theta_{c_m j'} - \sum_{j'=0}^{j} \theta_{c_m j'} \qquad (29) \\ &= \frac{1}{1+\exp\{\beta_{0(j+1)}+\beta_{1(j+1)}^{X_1}+\cdots+\beta_{k_1(j+1)}^{X_1}+\cdots+\beta_{1(j+1)}^{X_n}+\cdots+\beta_{k_n(j+1)}^{X_n}\}} \\ &\quad - \frac{1}{1+\exp\{\beta_{0j}+\beta_{1j}^{X_1}+\cdots+\beta_{k_1 j}^{X_1}+\cdots+\beta_{1j}^{X_n}+\cdots+\beta_{k_n j)}^{X_n}\}} \qquad (30)\end{aligned}$$

However, from the constraint in Eq. (11),

$$\exp\{\beta_{0j}+\beta_{1j}^{X_1}+\cdots+\beta_{k_1j}^{X_1}+\cdots+\beta_{1j}^{X_n}+\cdots+\beta_{k_nj)}^{X_n}\} \geq \exp\{\beta_{0(j+1)}+\beta_{1(j+1)}^{X_1}+\cdots+\beta_{k_1(j+1)}^{X_1}+\cdots+\beta_{1(j+1)}^{X_n}+\cdots+\beta_{k_n(j+1)}^{X_n}\} \quad (31)$$

Eqs. (30) and (31) together imply that

$$\theta_{c_m j} \;\; \geq \;\; 0 \quad (32)$$

Since the two terms in the RHS of Eq. (30) are logistic sigmoid functions their difference can not exceed 1. Hence $0 \leq \theta_{c_m j} \leq 1$. **QED**.

In the case of binary target variables the required constraint is simpler, *i.e.*, the constraints on the $\beta_{rj}^{X_i}$ parameters suffice.

Corollary 3 (M_{CLR}^{2a}) *Let Y be a child variable with k_Y levels, $k_Y = 2$, in a Bayesian network BN with parents $X_1, \ldots, X_n$ where each X_i has k_{X_i} levels, $i = 1, \ldots, n$. Suppose that $Q+$ is the qualitative influence of each X_i on Y and that M_{CLR}^2 is the constrained logistic regression model with the set of constraints replaced by*

$$\beta_{rj}^{X_i} \;\; \geq \;\; 0 \quad (33)$$

Then Y is **FSD** *isotonic in X_i.*

We need to show that

(a) For any configuration c_m and for class $j = 0, 1$ of Y

$$0 \leq \theta_{c_m j} \leq 1 \quad (34)$$

(b) For any configuration c_m,

$$\theta_{c_m 0} + \theta_{c_m 1} \;\; = \;\; 1 \quad (35)$$

(c) For any pair of configurations c_1, c_2 such that $c_2 > c_1$ *ceteris paribus*,

$$F_{c_2}(j) \leq F_{c_1}(j) \quad (36)$$

for $j = 0, 1$.

Since there are only two CPT cells in each row corresponding to a configuration c_m, we need only solve the value for one of the cells. Let us denote this as $\theta_{c_m 0}$. So $\theta_{c_m 1} = 1 - \theta_{c_m 0}$. Hence we only need perform one logistic regression. From our definition, the log odds ratio is

$$\log \frac{P(Y > 0|c_m)}{P(Y \leq 0|c_m)} = \beta_{00} + \beta_{10}^{X_1} I[X_1 \geq 1] + \cdots + \beta_{k_{X_i}0}^{X_1} I[X_1 \geq k_{X_1}] + \cdots + \beta_{10}^{X_n} I[X_n \geq 1] + \cdots + \beta_{k_{X_n}0}^{X_n} I[X_n \geq k_{X_n}] \quad (37)$$

$$\frac{1-\theta_{c_m 0}}{\theta_{c_m 0}} = \exp\{\beta_{00} + \beta_{10}^{X_1} I[X_1 \geq 1] + \cdots + \beta_{k_{X_1} 0}^{X_1} I[X_1 \geq k_{X_1}] + \cdots + \beta_{10}^{X_n} I[X_n \geq 1] + \cdots + \beta_{k_{X_n} 0}^{X_n} I[X_n \geq k_{X_n}]\} \quad (38)$$

Suppose we choose two configurations c_1 and c_2 so that $c_2 > c_1$ *ceteris paribus*, Eq. (38) implies that

$$\theta_{c_1 0} = \frac{1}{1+\exp\{\beta_{00}+\beta_{10}^{X_1}+\cdots+\beta_{k_1 0}^{X_1}+\cdots+\beta_{10}^{X_q}+\cdots+\beta_{k_q 0}^{X_q}+\cdots+\beta_{10}^{X_n}+\cdots+\beta_{k_n 0}^{X_n}\}} \quad (39)$$

$$\theta_{c_2 0} = \frac{1}{1+\exp\{\beta_{00}+\beta_{10}^{X_1}+\cdots+\beta_{k_1 0}^{X_1}+\cdots+\beta_{10}^{X_q}+\cdots+\beta_{k_q 0}^{X_q}+\beta_{(k+1)_q 0}^{X_q}+\cdots+\beta_{10}^{X_n}+\cdots+\beta_{k_n 0}^{X_n}\}} \quad (40)$$

for some $(k_1 \leq k_{X_1}), \ldots, (k_q \leq k_{X_q}), ((k+1)_q \leq k_{X_q}), \ldots, (k_n \leq k_{X_n})$.

Since Eq. (33) holds, it is clear from Eqs. (39) and (40) that $\theta_{c_1 0} \geq \theta_{c_2 0}$. The condition in Eq. (35) is trivial since for any c_m, $\theta_{c_m 0} + (1 - \theta_{c_m 0}) = 1$. So, when $j = 0$, $F_{c_2}(0) \leq F_{c_1}(0)$ because $\theta_{c_2 0} \leq \theta_{c_1 0}$ from Eqs. (39) and (40). In addition, when $j = 1$, it is the case that $F_{c_2}(1) = F_{c_1}(1) = 1$. Therefore, the condition for stochastic dominance in (36) is also satisfied. Now, we need to show that $\theta_{c_m 0}$ and $\theta_{c_m 1}$ are points in the interval $[0, 1]$. This immediately follows since $\theta_{c_m 0}$ is a logistic sigmoid function and since $\theta_{c_m 1} = 1 - \theta_{c_m 0}$, it follows that $0 \leq \theta_{c_m j} \leq 1$, $j = 0, 1$. **QED**

Although the number of constraints N_C^2 as shown in Theorem 2 is exponential in the number of parents, $N_C^2 = (k_Y - 1) \sum_{i=1}^{n} k_{X_i} + (k_Y - 2) \prod_{i=1}^{n} k_{X_i}$, it is possible to impose constraints on the β parameters so that the resulting number of constraints do not depend on the number of parent configurations. We can do this by replacing the constraints in Eq. 11 with a set of stronger but simpler constraints.

Corollary 4 (M_{CLR}^{2b}) *Let Y be a child variable with k_Y levels, $k_Y > 2$, in a Bayesian network BN with parents $X_1, \ldots, X_n$ where each X_i has k_{X_i} levels, $i = 1, \ldots, n$. Suppose that $Q+$ is the qualitative influence of each X_i on Y, $i = 1, \ldots, n$ and that M_{CLR}^2 is the constrained logistic regression model with the set of constraints replaced by*

1. *For $i = 1, \ldots, n$, $r = 1, \ldots, k_{X_i}$, and $j = 0, \ldots, k_Y - 1$*

$$\beta_{rj}^{X_i} \geq 0 \quad (41)$$

2. For each pair $(j-1)$,j*, where* $j = 1, \ldots, k_Y - 2$

$$\begin{aligned}
\beta_{0(j-1)} &\geq \beta_{0j} \\
\beta^{X_1}_{1(j-1)} &\geq \beta^{X_1}_{1j} \\
&\vdots \\
\beta^{X_1}_{k_{X_1}(j-1)} &\geq \beta^{X_1}_{k_{X_1}j} \\
&\vdots \\
\beta^{X_n}_{1(j-1)} &\geq \beta^{X_n}_{1j} \\
&\vdots \\
\beta^{X_n}_{k_{X_n}(j-1)} &\geq \beta^{X_n}_{k_{X_n}j}
\end{aligned} \tag{42}$$

Then Y *is* **FSD** *monotonic in* X_i*.*

It is obvious that if each component in the LHS of Eq. (11) is at least as large as its corresponding component in the RHS then the sum of the LHS is at least as large as the sum of the RHS. Hence, Eq. (42) implies Eq. (11). **QED**.

Here, the number of β parameters is $(k_Y - 1)[1 + \sum_{i=1}^{n}(k_{X_i} - 1)]$. The advantage of using Corollary 4 is that the number of constraints no longer depend on the number of configurations. In fact, the number of constraints, like the number of β parameters, is also linear in the number of parent levels. In particular, if $N_C^{2'}$ denote the number of constraints then $N_C^{2'} = (k_Y - 1)\sum_{i=1}^{n} k_{X_i} + (k_Y - 2)[1 + \sum_{i=1}^{n} k_{X_i}]$.

3.3 Constrained Logistic Regression Model 3 (M^3_{CLR})

As in M^2_{CLR}, the conditional distribution in the third model is estimated by using the parent levels of the random variable. Here, we also estimate the conditional distribution by estimating $k_Y - 1$ logistic functions and then subtract the values of the estimated adjacent logistic functions to compute the conditional distribution of Y given a parent configuration. The main difference between M^2_{CLR} and M^3_{CLR} is that the β parameters in M^3_{CLR} are constrained to increase as the parent level is increased. In addition, the inequality $\geq$ in the indicator function I for a parent level is replaced by the equality operator. We show that the constraints we impose on the β parameters of M^3_{CLR} are sufficient for first-order stochastic dominance.

Definition 5 (M^3_{CLR}**)** *Let* Y *be a child variable with* k_Y *levels,* $k_Y \geq 2$*, in a Bayesian network BN with parents* $X_1, \ldots, X_n$ *where each* X_i *has* k_{X_i} *levels,* $k_{X_i} \geq 2$ *for* $i = 1, \ldots, n$*. Given some configuration* $c = \langle x_1, \ldots, x_n \rangle$*, define*

$$\begin{aligned}
\log \frac{P(Y > j|c)}{P(Y \leq j|c)} = \beta_{0j} + \beta^{X_1}_{1j} I[X_1 = 1] + \cdots + \beta^{X_1}_{k_{X_1}j} I[X_1 = k_{X_1}] + \cdots + \\
\beta^{X_n}_{1j} I[X_n = 1] + \cdots + \beta^{X_n}_{k_{X_n}j} I[X_n = k_{X_n}] \text{ for } j = 0, \ldots, k_Y - 2
\end{aligned} \tag{43}$$

Also, let the following set of constraints hold

1. *For each parent* X_i, $i = 1, \ldots, n$ *and* $j = 0, \ldots, k_Y - 1$

$$\beta^{X_i}_{k_{X_i} j} \geq \beta^{X_i}_{(k_{X_i}-1)j} \geq \cdots \geq \beta^{X_i}_{2j} \geq \beta^{X_i}_{1j} \geq 0 \tag{44}$$

2. *For each pair* $(j-1)$,j, *where* $j = 1, \ldots, k_Y - 2$

$$\begin{aligned}
\beta_{0(j-1)} &\geq \beta_{0j} \\
\beta^{X_1}_{1(j-1)} &\geq \beta^{X_1}_{1j} \\
&\vdots \\
\beta^{X_1}_{k_{X_1}(j-1)} &\geq \beta^{X_1}_{k_{X_1} j} \\
&\vdots \\
\beta^{X_n}_{1(j-1)} &\geq \beta^{X_n}_{1j} \\
&\vdots \\
\beta^{X_n}_{k_{X_n}(j-1)} &\geq \beta^{X_n}_{k_{X_n} j}
\end{aligned} \tag{45}$$

Definition 5 expresses the log odds ratio between the cumulative conditional probabilities $P(Y > j|c)$ and $P(Y \leq j|c)$ as a linear function of the levels of the parent variables $X_1, \ldots, X_n$. Such formulation allows one to model the contribution of each level of X_i to the log odds ratio via the β parameters. The main difference between Definition 5 and Definition 4 is that in the latter the contribution of the parent levels are arranged so that they become cumulative (similar to a thermometer bar) where if $X = k$ for some k then all terms $X = k'$, $k' < k$ also need to be true. In the former case the contribution of each parent level is independent of the contribution of the other parent levels. M^3_{CLR} also allows the expression of synergistic and anti-synergistic influences where terms that interact are simply added to the linear function, similar to M^2_{CLR} (see Eq. (12) for an example under M^2_{CLR}). In addition, the number of constraints N^3_C for model M^3_{CLR} is linear in the number of parent levels since $N^3_C = n + \sum_{i=1}^{n}(k_{X_i} - 1) + (k_Y - 1)\sum_{i=1}^{n}(k_{X_i} - 1) = n + k_Y \sum_{i=1}^{n}(k_{X_i} - 1)$. The number of β parameters, N^3_β, is also linear in the number of parent levels, $N^3_\beta = (k_Y - 1)[1 + \sum_{i=1}^{n}(k_{X_i} - 1)]$.

Theorem 5 *Let* Y *be a child variable with* k_Y *levels,* $k_Y > 2$*, in a Bayesian network BN with parents* $X_1, \ldots, X_n$ *where each* X_i *has* k_{X_i} *levels,* $i = 1, \ldots, n$*. If* M^3_{CLR} *is the constrained logistic regression model,* $Q+$ *the qualitative influence of each* X_i *on* Y*, then* Y *is* **FSD** *monotonic in* X_i*.*

To prove the theorem, again we need to show the following:

(a) Let $\theta_{c_m j}$ denote a CPT parameter for configuration c_m and for class j of Y, *i.e.,* $Y = j$ in the CPT table. Then

$$0 \leq \theta_{c_m j} \leq 1 \tag{46}$$

(b) For any configuration c_m,

$$\sum_{j=0}^{k_Y-1} \theta_{c_m j} \;=\; 1 \tag{47}$$

(c) For any pair of configurations c_1, c_2 such that $c_2 > c_1$ *ceteris paribus*,

$$F_{c_2}(j) \leq F_{c_1}(j) \tag{48}$$

for $j = 0, \ldots, k_Y - 1$.

Given a configuration $c_m = \langle x_{i1}, \ldots, x_{in} \rangle$, the log odds ratio as expressed using M^3_{CLR} is

$$\begin{aligned}\log \frac{P(Y > j|c_m)}{P(Y \leq j|c_m)} = \beta_{0j} + \beta_{1j}^{X_1} I[X_1 = 1] + \cdots + \beta_{k_{X_i} j}^{X_1} I[X_1 = k_{X_1}] + \cdots + \\ \beta_{1j}^{X_n} I[X_n = 1] + \cdots + \beta_{k_{X_n} j}^{X_n} I[X_n = k_{X_n}] \text{ for } j = 0, \ldots, k_Y - 2\end{aligned} \tag{49}$$

Let $c_1 = \langle x_{11}, \ldots, x_{1q}, \ldots, x_{1n} \rangle$ and $c_2 = \langle x_{21}, \ldots, x_{2q}, \ldots, x_{2n} \rangle$ be a pair of configurations such that $c_1 < c_2$ *ceteris paribus, i.e.,* they differ only at a single position q so that $x_{1q} < x_{2q}$. Also, let $P(Y = j|c_m) = \theta_{c_m j}$. As in previous sections, we want to show that for all values j, $j = 0, \ldots, k_Y - 2$, it is the case that $F_{c_2}(j) \leq F_{c_1}(j)$. Assuming $j = 0$, from Eq. (49) we have for $i = 1, 2$

$$\begin{aligned}\log \frac{P(Y > 0|c_m)}{P(Y \leq 0|c_m)} = \beta_{00} + \beta_{10}^{X_1} I[X_1 = 1] + \cdots + \beta_{k_{X_1} 0}^{X_1} I[X_1 = k_{X_1}] + \cdots + \\ \beta_{10}^{X_n} I[X_n = 1] + \cdots + \beta_{k_{X_n} 0}^{X_n} I[X_n = k_{X_n}]\end{aligned} \tag{50}$$

$$\begin{aligned}\frac{1 - \theta_{c_m 0}}{\theta_{c_m 0}} = \exp\{\beta_{00} + \beta_{10}^{X_1} I[X_1 = 1] + \cdots + \beta_{k_{X_1} 0}^{X_1} I[X_1 = k_{X_1}] + \cdots + \\ \beta_{10}^{X_n} I[X_n = 1] + \cdots + \beta_{k_{X_n} 0}^{X_n} I[X_n = k_{X_n}]\}\end{aligned} \tag{51}$$

Eq. (51) implies that

$$\theta_{c_1 0} = \frac{1}{1 + \exp\{\beta_{00} + \cdots + \beta_{k_q 0}^{X_q} + \cdots + \beta_{k_p 0}^{X_p}\}} \tag{52}$$

$$\theta_{c_2 0} = \frac{1}{1 + \exp\{\beta_{00} + \cdots + \beta_{k'_q 0}^{X_q} + \cdots + \beta_{k_p 0}^{X_p}\}} \tag{53}$$

Since $c_2 > c_1$, there exists indices k_q and k'_q such that $k_q \leq k'_q \leq k_{X_q}$ for some parent X_q of Y where $q < p \leq n$. By Eq. (44), $\beta_{k'_q 0}^{X_q} \geq \beta_{k_q 0}^{X_q}$. Hence, Eqs. (52) and (53) imply that $\theta_{c_1 0} \geq \theta_{c_2 0}$.

Now, suppose that $0 < j \leq (k_Y - 2)$. From Eq. (49),

$$\frac{1-\sum_{j'=0}^{j}\theta_{c_m j'}}{\sum_{j'=0}^{j}\theta_{c_m j'}} = \exp\{\beta_{0j} + \beta_{1j}^{X_1} I[X_1 = 1] + \cdots + \beta_{k_{X_1} j}^{X_1} I[X_1 = k_{X_1}] + \cdots + \beta_{1j}^{X_n} I[X_n = 1] + \cdots + \beta_{k_{X_n} j}^{X_n} I[X_n = k_{X_n}]\} \quad (54)$$

since $c_2 > c_1$, Eq. (54) implies

$$\sum_{j'=0}^{j}\theta_{c_1 j'} = \frac{1}{1+\exp\{\beta_{0j} + \cdots \beta_{k_q j}^{X_q} + \cdots + \beta_{k_p j}^{X_p}\}} \quad (55)$$

and

$$\sum_{j'=0}^{j}\theta_{c_2 j'} = \frac{1}{1+\exp\{\beta_{0j} + \cdots \beta_{k'_q j}^{X_q} + \cdots + \beta_{k_p j}^{X_p}\}} \quad (56)$$

where $k'_q > k_q$ and $q < p \leq n$. Again, from Eqs. (55) and (56)

$$\sum_{j'=0}^{j}\theta_{c_1 j'} \geq \sum_{j'=0}^{j}\theta_{c_2 j'}$$

since according to Eq. (44), $\beta_{k'_q j}^{X_q} \geq \beta_{k_q j}^{X_q}$. By definition,

$$F_{c_2}(j) \leq F_{c_1}(j)$$

holds for $j = 0, \ldots, k_Y - 2$. Again, note that once we have shown Eq. (47) to hold, we will have shown that Eq. (48) also holds, *i.e.*, for all levels of Y.

Observe that

$$\sum_{j'=0}^{j}\theta_{c_m j'} = \frac{1}{1+\exp\{\beta_{0j} + \cdots \beta_{k_q j}^{X_q} + \cdots + \beta_{k_p j}^{X_p}\}} \quad (57)$$

and that

$$\theta_{c_m j} = \sum_{j'=0}^{j+1}\theta_{c_m j'} - \sum_{j''=0}^{j}\theta_{c_m j''} \quad (58)$$

Hence,

$$\sum_{j=0}^{k_Y-1}\theta_{c_m j} = \theta_{c_m 0} + [(\theta_{c_m 0} + \theta_{c_m 1}) - \theta_{c_m 0}] + \ldots + \left[\sum_{j'=0}^{j}\theta_{c_m j'} - \sum_{j'=0}^{j-1}\theta_{c_m j'}\right] + \left[\sum_{j'=0}^{j+1}\theta_{c_m j'} - \sum_{j'=0}^{j}\theta_{c_m j'}\right] + \ldots + \left[\sum_{j'=0}^{k_Y-2}\theta_{c_m j'} - \sum_{j'=0}^{k_Y-3}\theta_{c_m j'}\right] + \left[1 - \sum_{j'=0}^{k_Y-2}\theta_{c_m j'}\right] \quad (59)$$

$$= 1 \quad (60)$$

Since every term in the RHS of Eq. (59) cancels out except the penultimate term 1, Eq. (47) holds, and therefore Eq. (48) follows.

Now, we need to show that each $\theta_{c_m j} \in [0, 1]$. Suppose $j = 0$ then

$$\sum_{j'=0}^{j} \theta_{c_m j'} = \theta_{c_m 0} = \frac{1}{1+\exp\{\beta_{0j}+\cdots+\beta_{k_q j}^{X_q}+\cdots+\beta_{k_p j}^{X_p}\}} \tag{61}$$

for $q < p \leq n$. Since Eq. (61) is a logistic sigmoid function, $\theta_{c_m 0} \in [0, 1]$. Suppose that $0 < j \leq k_Y$–1, then

$$\theta_{c_m j} = \sum_{j'=0}^{j+1} \theta_{c_m j'} - \sum_{j'=0}^{j} \theta_{c_m j'} \tag{62}$$

$$= \frac{1}{1+\exp\{\beta_{0(j+1)}+\cdots+\beta_{k_q(j+1)}^{X_q}+\cdots+\beta_{k_p(j+1)}^{X_p}\}} - \frac{1}{1+\exp\{\beta_{0j}+\cdots+\beta_{k_q j}^{X_q}+\cdots+\beta_{k_p j)}^{X_p}\}} \tag{63}$$

However, from the constraint in Eq. (45),

$$\exp\{\beta_{0j}+\cdots+\beta_{k_q j}^{X_q}+\cdots+\beta_{k_p j)}^{X_p}\} \geq \exp\{\beta_{0(j+1)}+\cdots+\beta_{k_q(j+1)}^{X_q}+\cdots+\beta_{k_p(j+1)}^{X_p}\} \tag{64}$$

Eqs. (63) and (64) together imply that

$$\theta_{c_m j} \geq 0 \tag{65}$$

Since the two terms in the RHS of Eq. (63) are logistic sigmoid functions their difference can not exceed 1. Hence $0 \leq \theta_{c_m j} \leq 1$. **QED**.

4 Enforcing margins

Given two parent configurations c_1 and c_2 where $c_1 < c_2$, the idea of enforcing a non-negative margin $\epsilon > 0$ between the cumulative distribution F_{c_1} and F_{c_2} has appeared to help improve classification accuracy as reported in previous work [2]. In this section, we discuss how margins could be enforced in the constrained logistic regression setting using M^3_{CLR} as an example to demonstrate the idea. The margins, which can also be elicited from a domain expert, can help strengthen monotonicity assumptions and could be considered beneficial in the design of classifiers especially in cases where there are very small amounts of training data. Enforcing margins has the effect of reducing the hypothesis search space (see Figure 4) during parameter fitting[2]. After fitting

[2]Parameter fitting can be done via maximum likelihood estimation (see [11, 15, 2]) for more detailed discussion.

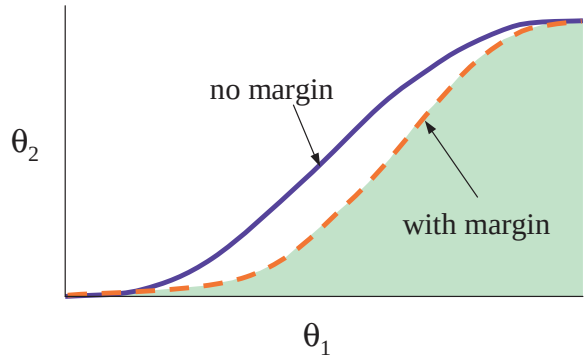

Figure 4: Effect of the margin on the parameter space in the logistic regression setting

the β parameters one could easily get, if needed, the corresponding θ parameters by plugging in the β parameters to the corresponding logistic functions, for example by using Eq. (52) and (53). The solid curve in Figure 4 divides the space of possible parameter values into two regions: $\theta_1 \geq \theta_2$ which is the region to the right of the solid curve and $\theta_1 \leq \theta_2$ which is the region to the left of the solid curve. When a margin is enforced, the region in which $\theta_1 \geq \theta_2$ becomes smaller (confined to the shaded region under the dashed curve). This focuses the search for parameter values toward a much smaller region where correct parameter values are likely to be in, potentially aiding in faster learning with sparse data sets. The idea simply means that instead of the inequality in Eq. (48) we now require that their difference be at least some margin ϵ.

$$F_{c_2}(j) + \epsilon \leq F_{c_1}(j) \tag{66}$$

There are some issues involving how the margin ϵ should be chosen and how it should be enforced. For example, it is possible that ϵ could be chosen based on some probability distribution and as more evidence becomes available a posterior distribution could be obtained from the initial prior distribution. Also, the margins between cumulative distributions need not be equal, *i.e.*, for some cumulative distributions $F_{c_m}(j), F_{c_{m+1}}(j), F_{c_{m+2}}(j)$ margins can be enforced such that $F_{c_m}(j) - F_{c_{m+1}}(j) \geq \epsilon_m$ and $F_{c_{m+1}}(j) - F_{c_{m+2}}(j) \geq \epsilon_{m+1}$, $\epsilon_m \neq \epsilon_{m+1}$. Many issues still remain unclear as to what the best approach for enforcing margins should be and we hope to address them in future work. Meantime, we are interested in finding a way to adjust the margin without solving an exponential number of constraints.

Let us state what the problem is in the context of M^3_{CLR}. Given a pair of configuration c_1 and c_2, where $c_2 > c_1$ *ceteris paribus* we want

$$F_{c_1}(j) - F_{c_2}(j) \geq \epsilon \tag{67}$$

which implies

$$\sum_{j'=0}^{j} \theta_{c_1 j'} - \sum_{j'=0}^{j} \theta_{c_2 j'} \geq \epsilon \tag{68}$$

for $j = 0, \ldots, k_Y - 2$. From Eq. (43),

$$\frac{1}{1+\exp\{\beta_{0j} + \cdots \beta^{X_q}_{k_q j} + \cdots + \beta^{X_m}_{k_m j}\}} - \frac{1}{1+\exp\{\beta_{0j} + \cdots \beta^{X_q}_{k'_q j} + \cdots + \beta^{X_m}_{k_m j}\}} \geq \epsilon \tag{69}$$

where $k'_q = k_q + 1$. Note that by Eq. (44), $\beta^{X_q}_{k'_q j} \geq \beta^{X_q}_{k_q j}$. If we group the β terms that are equal and assign their sum as some constant κ then we can rearrange the above inequality into

$$\frac{1}{1+\exp\{\kappa + \beta^{X_q}_{k_q j}\}} - \frac{1}{1+\exp\{\kappa + \beta^{X_q}_{k'_q j}\}} \geq \epsilon \tag{70}$$

It is not desirable to use Eq. (70) for each pair of configuration c_1 and c_2 since this will generate an exponential number of constraints. Our goal then is to find a way to control ϵ by adjusting the difference between the parameters $\beta^{X_q}_{k'_q j}$ and $\beta^{X_q}_{k_q j}$. From the LHS of Eq. (70) we have

$$\frac{1}{1+\exp\{\kappa + \beta^{X_q}_{k_q j}\}} - \frac{1}{1+\exp\{\kappa + \beta^{X_q}_{k'_q j}\}} \tag{71}$$

Given that $\beta^{X_q}_{k'_q j} \geq \beta^{X_q}_{k_q j}$ it can be shown from Eq. (71) that by adjusting $(\beta^{X_q}_{k'_q j} - \beta^{X_q}_{k_q j})$ the resulting margin is bounded. Define

$$M^{X_q}_{k_q k'_q j} \;\; = \;\; (\beta^{X_q}_{k'_q j} - \beta^{X_q}_{k_q j}) \tag{72}$$

If we let $\beta^{X_q}_{k'_q j} \rightarrow +\infty$ in Eq. (71) then $[F_{c_1}(j) - F_{c_2}(j)]$ tends toward the limit

$$\frac{1}{1+\exp\{\kappa + \beta^{X_q}_{k_q j}\}} \tag{73}$$

and if $\beta^{X_q}_{k_q j} \rightarrow \beta^{X_q}_{k'_q j}$, then $[F_{c_1}(j) - F_{c_2}(j)] \rightarrow 0$. Hence the bounds are

$$0 \leq [F_{c_1}(j) - F_{c_2}(j)] \leq \frac{1}{1+\exp\{\kappa + \beta^{X_q}_{k_q j}\}} \tag{74}$$

The obvious implication here is that one can control the margin by adjusting the difference between $\beta^{X_q}_{k'_q j}$ and $\beta^{X_q}_{k_q j}$ and one simple strategy which has been adopted by Altendorf et al. [2] is to assign a uniform margin for each pair of conditional cumulative distributions. However, note that equal difference between any pair of $\beta^{X_q}_{k'_q j}$ and $\beta^{X_q}_{k_q j}$ does not necessarily translate to equal margins between $\sum_{j'=0}^{j} \theta_{c_m j}$ in Eq. (68). In other words, given a specified ϵ the difference between $\beta^{X_q}_{k'_q j}$ and $\beta^{X_q}_{k_q j}$ varies from one pair to another. As a simple example, consider a Bayesian network whose ternary target random variable Y is influenced by a single ternary parent X.

$$P(Y = 0|X) = \frac{1}{1 + \exp\{\beta_{00} + \beta^X_{10} I[X = 1] + \beta^X_{20} I[X = 2]\}} \tag{75}$$

Suppose that $\beta_{00} = 0.2$, $\beta^X_{10} = 0.4$, and $\beta^X_{20} = 0.6$. If the β parameters have these values (note that their consecutive differences is 0.2) then the margins between the corresponding $\theta_{c_m 0}$, $c_m = 0, 1, 2$

are

$$\begin{aligned}
\theta_{00} - \theta_{10} &= \frac{1}{1+\exp\{\beta_{00}\}} - \frac{1}{1+\exp\{\beta_{00}+\beta_{10}^{X}\}} \\
&= \frac{1}{1+\exp\{0.2\}} - \frac{1}{1+\exp\{0.2+0.4\}} \\
&= 0.0958 \qquad (76) \\
\theta_{10} - \theta_{20} &= \frac{1}{1+\exp\{\beta_{00}+\beta_{10}^{X}\}} - \frac{1}{1+\exp\{\beta_{00}+\beta_{20}^{X}\}} \\
&= \frac{1}{1+\exp\{0.2+0.4\}} - \frac{1}{1+\exp\{0.2+0.6\}} \\
&= 0.0443 \qquad (77)
\end{aligned}$$

From the above example, it is clear that unless we know the values of the β parameters, it is not possible to uniformly[3] set the margin ϵ over a set of constraints using the difference between two consecutive β parameters. Moreover, since the β parameters are not known at the time the margin is specified the upper bound in Eq. (74) also can not be computed. In other words, given that we want to enforce a margin of $\epsilon > 0$ it is not possible to specify what the value of $M_{k_q k'_q j}^{X_q}$ should be.

Another simple approach to adjusting the margin ϵ is to choose a quantity whose range is in $[0, 1]$. Define the ratio

$$R_{k_q k'_q j}^{X_q} = \frac{\beta_{k_q j}^{X_q}}{\beta_{k'_q j}^{X_q}} \qquad (78)$$

Although this makes the specification of a margin by a domain expert more intuitive in the sense that the expert need only think of a quantity between 0 and 1, note that solving the exact ratio for a given ϵ is still untenable since the β parameters are not known at the time when ϵ is specified. Like Eq.(72), this disadvantage is simply the result of the tradeoff between expressing the margin in terms of the β parameters and expressing it as a constraint for every configurations c_m and c_{m+1} *ceteris paribus* via the θ parameters. As we have mentioned, our goal is to avoid the exponential number of constraints implicit in the latter method so a feasible initial strategy is to specify the margins in terms of the β parameters, instead of the original θ parameters.

If $\beta_{k_q j}^{X_q} \to 0$ then $R_{k_q k'_q j}^{X_q} \to 0$ while if $\beta_{k_q j}^{X_q} \to \beta_{k'_q j}^{X_q}$ then $R_{k_q k'_q j}^{X_q} \to 1$. So the bound for $R_{k_q k'_q j}^{X_q}$ is,

$$0 \leq R_{k_q k'_q j}^{X_q} \leq 1 \qquad (79)$$

In addition, we can define a term $\alpha_{k_q k'_q j}^{X_q}$ such that $1 - \alpha_{k_q k'_q j}^{X_q} = R_{k_q k'_q j}^{X_q}$. If we want to increase the difference between $\beta_{k_q j}^{X_q}$ and $\beta_{k'_q j}^{X_q}$ then we simply increase $\alpha_{k_q k'_q j}^{X_q}$. Setting $\alpha_{k_q k'_q j}^{X_q} = 0$ reduces

[3]It is still an open question whether a uniform margin is desirable. Previous work by Altendorf, Restificar, and Dietterich [2] using a uniform amount of margin for each constraint appears to be a good initial strategy.

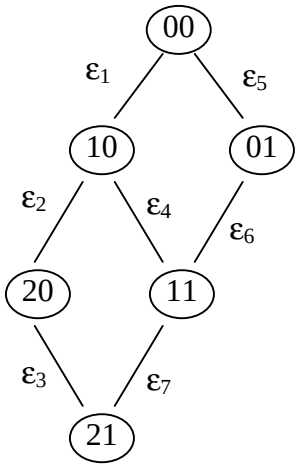

Figure 5: Constraint lattice

this difference to 0. As $\alpha^{X_q}_{k_q k'_q j} \rightarrow 1$, the resulting difference in the cumulative distribution of configurations c_1 and c_2 can not exceed 1. The argument is the same as that of $M^{X_q}_{k_q k'_q j}$ since both terms in Eq. (71) are logistic functions and $\alpha^{X_q}_{k_q k'_q j} \rightarrow 0$ can mean letting either $\beta^{X_q}_{k_q j} \rightarrow 0$ from the right or $\beta^{X_q}_{k'_q j} \rightarrow +\infty$. The result is exactly that in Eq. (73). From above, we can now state the following formally,

Theorem 6 *Suppose that* $M^{X_q}_{k_q k'_q j} = \beta^{X_q}_{k_{q'} j} - \beta^{X_q}_{k_q j}$ *and* $\alpha^{X_q}_{k_q k'_q j} = 1 - \frac{\beta^{X_q}_{k_q j}}{\beta^{X_q}_{k_{q'} j}}$. *Then*

$$0 \leq M^{X_q}_{k_q k'_q j} < \infty \implies 0 \leq [F_{c_1}(j) - F_{c_2}(j)] \leq \frac{1}{1 + \exp\{\kappa + \beta^{X_q}_{k_q j}\}} \tag{80}$$

$$0 \leq \alpha^{X_q}_{k_q k'_q j} \leq 1 \implies 0 \leq [F_{c_1}(j) - F_{c_2}(j)] \leq \frac{1}{1 + \exp\{\kappa + \beta^{X_q}_{k_q j}\}} \tag{81}$$

Now that it is clear that the additional constraints imposed on the β parameters to enforce margins result in cumulative distributions that are bounded, let us consider all the cumulative distributions that are part of a chain of inequality constraints whose cumulative difference along the chain must not exceed 1 lest the feasible region becomes empty. To illustrate the problem, let us assume another Bayesian network with two parents X_1 and X_2. X_1 is a ternary random variable and X_2 is a binary random variable. The target, Y, is a ternary variable. Figure 5 shows as edges the corresponding constraints between cumulative distributions for a specified $Y = j$. For example, the node 00 and 01 represents $F_{00}(j)$ and $F_{01}(j)$ if we denote $c_1 = 00$ and $c_2 = 01$. The constraint $F_{00}(j) - F_{01}(j) \geq \epsilon_1$ is represented by the edge labeled ϵ_1. Note that the longest chain in the lattice has 3 edges. One example is the chain $\epsilon_1 : \epsilon_2 : \epsilon_3$. The corresponding label of the nodes of this chain satisfy $c_m > c_{m+1}$ *ceteris paribus*. It is also clear that the sum $\epsilon_1 + \epsilon_2 + \epsilon_3$ can not exceed 1 lest the feasible region becomes empty. For example, it is not possible to satisfy $\epsilon_1 + \epsilon_2 > 1$. Suppose

that $\epsilon_1 = k_{\epsilon_1}$ where $k_{\epsilon_1} \in [0, 1]$. Suppose that $F_{00}(j) = 1$, the highest probability value possible, then $F_{10}(j) = 1 - k_{\epsilon_1}$. Since the smallest value $F_{20}(j)$ can take is 0, the maximum value that ϵ_2 can have is also $1 - k_{\epsilon_1}$. So $\epsilon_1 + \epsilon_2 = k_{\epsilon_1} + (1 - k_{\epsilon_1}) = 1$. Clearly, the sum can not be greater than 1 unless $F_{00}(j) > 1$.

So, if we hold either $M^{X_q}_{k_q k'_q j}$ or $\alpha^{X_q}_{k_q k'_q j}$ constant, how does $[F_{c_1}(j) - F_{c_2}(j)]$ behave across the chain of inequalities in the lattice? Do we have guarantees of monotonicity on the values of the margins? Unfortunately, for both $M^{X_q}_{k_q k'_q j}$ and $\alpha^{X_q}_{k_q k'_q j}$ it is possible to find examples where for a set of β parameters $[F_{c_1}(j) - F_{c_2}(j)]$ is either decreasing, increasing or even nonmonotonic. As an example, consider again our previous Bayesian network with one parent. Suppose that this time the parent variable, X, has four levels ($\{0, 1, 2, 3\}$) and that we are only interested in $P(Y = 0|c_m)$, $c_m = 0, 1, 2, 3$. If we let β_{00}, β^X_{10}, β^X_{20} and β^X_{30} have as values -1, 0.1, 2.1, and 4.1, respectively, where $M^X_{k_q k'_q 0} = 2$ and $\kappa = 1$ then $[F_0(0) - F_1(0)] = 0.02$, $[F_1(0) - F_2(0)] = 0.20$, and $[F_2(0) - F_3(0)] = 0.04$. Note that $[F_{c_m}(j) - F_{c_{m+1}}(j)]$ is nonmonotonic. Hence, in general there is no way to tell in advance[4] how the various ϵ_m in the constraint lattice of a given problem will behave given a fixed $M^{X_q}_{k_q k'_q j}$ or $\alpha^{X_q}_{k_q k'_q j}$.

5 Related Work

Ng and Jordan [13] analyzed and compared both Naive Bayes and logistic regression classifiers. Their analyses show that asymptotically, the error of the generative Naive Bayes classifier is higher than that of the discriminative logistic regression classifier. In addition, however, the parameters of the Naive Bayes classifier, need only a number of samples that is logarithmic in n, the dimension of the input space, to be uniformly close to their asymptotic values while that of the logistic regression classifier need an order of n examples. Their results indicate that even though the asymptotic error for Naive Bayes is higher, it could converge more quickly than the logistic regression classifier. This implies, that given a small number of training instances the use of Naive Bayes classifier is advantageous and experimental results by the authors support their analysis.

Altendorf et al. [2] addressed the problem of learning from very sparse data, *e.g.*, with number of training instances ranging only between 1 and 10, by exploiting qualitative monotonicities, a specific form of background knowledge. Qualitative monotonicities which can be formalized as stochastic dominance [16, 1] are statements that describe monotonic relationship between two random variables such as "warm temperature increases increases the size of the mosquito population" or "increase in body mass index leads to an increase risk of elevated blood pressure". The general idea behind Altendorf et al.'s work is to focus the search for parameters during parameter-fitting to a smaller region of the hypothesis space that satisfies the constraints implied by the monotonocity constraints. In addition to the use of margins that reflect the domain expert's confidence of the background knowledge the technique significantly outperformed all the other algorithms tested, including Naive Bayes, in the case where the training data is sparse. This approach, however, is prone to problems related to parameter estimation especially on Bayesian networks where the

[4] A possible method might be to perform interleaved optimization for β and $M^X_{k_q k'_q 0}$ parameters until convergence is achieved (see *e.g.*, Ando and Zhang [3]) but as of now we will leave this problem for future work.

number of parents is large, *e.g.,* those that exceed six. The difficulty is largely attributed to an exponential number of parameters that need to be fitted.

Greiner et al. [11] provide formal analysis and experimental results on learning the parameters of an arbitrary belief network using normalized exponentiation, *i.e.,* using a simple logistic regression formulation where the linear discriminant of each logistic function only consists of the β constant. A simple gradient-descent algorithm is given that optimizes the log conditional likelihood for a given set of training instances instead of the joint likelihood. The same approach using normalized exponentiation is also employed independently by Altendorf et al. [2] to exploit monotonicity constraints, although the focus of the latter is orthogonal to that of the former. Greiner et al. work's emphasis is on finding a method of improving classification accuracy given a sufficient (but possibly incomplete) amount of training instances while Altendorf et al. focused on addressing the problem of learning with very sparse data. In the context of learning from very sparse data, both approaches, however, could easily lead to an exponential number of parameters especially with belief networks with large number of parents.

Roos et al. [15] present a study on the equivalence of conditional probability models that can be represented by Bayesian networks and those that can be represented by logistic regression. Roos et al. show that this equivalence holds whenever a given Bayesian network B has a canonical version B^* that is perfect, *i.e.,* all nodes with a common child are connected. The canonical version B^* is constructed by restricting B to the Markov blanket of the target class Y and adding arcs so that the parents of Y are fully connected. The authors, however, formulate the logistic regression problem in Bayesian networks differently from our model. While our model only associates β parameters to parent levels which without parent level interaction is linear in the number of parents, Roos et al. presented a model which associates β parameters to all parent configurations of a node. This implies that the number of β parameters in their model is exponential in the number of parents. In the context of learning from sparse data, especially when the traning set size is between 1 and 10, choosing a model with an exponential number of parameters will likely lead to poor parameter fitting. Evidence of this can be found in our experiments on Bayesian networks with large number of parents. In addition, while our model exploits background knowledge from domain experts, the inherent assumption for our approach to become effective in the sparse data setting, is that there is only a significantly small number of parent level interactions needed to be explicitly represented. From the practical standpoint, it is important to note that while our model can scale to a full model *i.e.,* with all the parent level interactions explicitly represented, there is no requirement imposed on the modeler to represent any parent level interaction unless such interaction is deemed necessary.

6 Summary

In this report, we have presented and provided theoretical analyses on three logistic regression models and their variants in terms of their suitability for Bayesian network learning with very sparse data. We presented at least two logistic regression models, M^{2b}_{CLR} and M^{3}_{CLR}, and have provided analyses on their suitability as models for Bayesian network learning with very sparse data. In particular, the number of parameters to be estimated and the number of constraints needed

to implement stochastic dominance for these models do not grow exponentially in the number of parents of the target variable. In addition, we have also discussed how margins can be enforced in the logistic regression setting and have pointed out some of the challenging issues related to its use, after having demonstrated the advantage of using them to enhance monotonicity in previous work.

References

[1] Alan Agresti and Christy Chuang. Bayesian and maximum likelihood approaches to order-restricted inference for models for ordinal categorical data. In *Proc. Advances in Order Restricted Statistical Inference*, pages 6–27. Springer-Verlag, 1985.

[2] Eric E. Altendorf, Angelo C. Restificar, and Thomas G. Dietterich. Learning from sparse data by exploiting monotonicity constraints. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI 2005)*, Edinburgh, Scotland, July 2005.

[3] Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, November 2005.

[4] Norman P. Archer and Shouhong Wang. Application of the back propagation neural network algorithm with monotonicity constraints for two-group classification problems. *Decision Sciences*, 24(1):60–75, 1993.

[5] Arie Ben-David. Monotonicity maintenance in information-theoretic machine learning algorithms. *Machine Learning*, 19(1):29–43, 1995.

[6] P. Clark and S. Matwin. Using qualitative models to guide inductive learning. In *Proc. of the International Conference on Machine Learning*, pages 49–56, 1993.

[7] Peter Clark and Stan Matwin. Using qualitative models to guide inductive learning. In *Proc. International Conference on Machine Learning*, pages 49–56, 1993.

[8] H. Daniels, A. Feelders, and M. Velikova. Integrating economic knowledge in data mining algorithms. In *Intl. Conf. Soc. Comp. Economics*, 2002.

[9] H. Daniels and B. Kamp. Application of MLP networks to bond rating and house pricing. *Neural Computing & Applications*, 8:226–234, 1999.

[10] A. J. Feelders. Prior knowledge in economic applications of data mining. In Djamel A. Zighed, Henryk Jan Komorowski, and Jan M. Zytkow, editors, *PKDD*, volume 1910 of *Lecture Notes in Computer Science*, pages 395–400. Springer, 2000.

[11] Russell Greiner, Xiaoyuan Su, Bin Shen, and Wei Zhou. Structural extension to logistic regression: Discriminative parameter learning of belief net classifiers. *Machine Learning*, 59:297–322, 2005.

[12] Herbert Kay and Lyle Ungar. Deriving monotonic function envelopes from observations. In *Proc. Qualitative Reasoning about Physical Systems*, 1993.

[13] Andrew Y. Ng and Michael I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in Neural Information Processing Systems 14)*, Cambridge, MA, 2002.

[14] Rob Potharst and A. J. Feelders. Classification trees for problems with monotonicity constraints. *SIGKDD Explorations*, 4(1):1–10, 2002.

[15] Teemu Roos, Hannes Wettig, Peter Grunwald, Petri Myllymaki, and Henry Tirri. On discriminative bayesian network classifiers and logistic regression. *Machine Learning*, 59:267–296, 2005.

[16] M. P. Wellman. Fundamental concepts of qualitative probabilistic networks. *Artif. Intell.*, 44(3):257–303, 1990.