Learning Bayesian Network Parameters Under Order Constraints

Ad Feelders, Linda C. van der Gaag
Institute of Information and Computing Sciences, Utrecht University,
PO Box 80.089, 3508TB Utrecht, The Netherlands

Abstract

We consider the problem of learning the parameters of a Bayesian network from data, while taking into account prior knowledge about the signs of influences between variables. Such prior knowledge can be readily obtained from domain experts. We show that this problem of parameter learning is a special case of isotonic regression and provide a simple algorithm for computing isotonic estimates. Our experimental results for a small Bayesian network in the medical domain show that taking prior knowledge about the signs of influences into account leads to an improved fit of the true distribution, especially when only a small sample of data is available. More importantly, however, the isotonic estimator provides parameter estimates that are consistent with the specified prior knowledge, thereby resulting in a network that is more likely to be accepted by experts in its domain of application.

1 Introduction

Bayesian networks by now are widely accepted as powerful tools for representing and reasoning with uncertainty in decision-support systems. A Bayesian network is a concise model of a joint probability distribution over a set of stochastic variables [1]; it consists of a directed acyclic graph that captures the qualitative dependence structure of the distribution and a numerical part that specifies conditional probability distributions for each variable given its parents in the graph. Since a Bayesian network defines a unique distribution, it provides for computing any probability of interest over its variables.

For constructing a Bayesian network, often knowledge is acquired from experts in the domain of application. Experience shows that domain experts can quite easily and reliably specify the graphical structure of a network [2], yet tend to have more problems in coming up with the probabilities for its numerical part [3]. If data from every-day problem solving in the domain are available therefore, one would like to use these data for estimating the probabilities that are required for the graphical structure to arrive at a fully specified network. Often, unfortunately, the available data sample is quite small, giving rise to inaccurate estimates. These inaccuracies may then lead to a reasoning behaviour of the resulting network that violates common domain knowledge, and the network will not easily be accepted by experts in the domain.

While domain experts often are found to have difficulties in coming up with probability assessments, evidence is building up that they feel more comfortable with providing qualitative knowledge about the probabilistic influences between the variables concerned [2, 4]. The qualitative knowledge provided by the experts, moreover, tends to be more robust than their numerical assessments. We demonstrate in this paper that expert knowledge about the signs of the influences between the variables in a Bayesian network can be used to improve the probability estimates obtained from small data samples. We show how these signs impose order constraints on the probabilities required for the network. We then show that the problem of estimating probabilities under these order constraints is a special case of isotonic regression. Building upon this property, we present an estimator that is guaranteed to produce probability estimates that reflect the qualitative knowledge that has been specified by the experts. The resulting network as a consequence is less likely to exhibit counterintuitive reasoning behaviour and is more likely to be accepted than a network with unconstrained estimates.

The paper is organised as follows. In the next section, we briefly review Bayesian networks and qualitative influences. In section 3, a small network in medicine is introduced; we show that this network may reveal highly counterintuitive reasoning behaviour when quantified with probability estimates from a small data sample that do not adhere to the specified domain knowledge. In section 4, we discuss isotonic regression and provide two algorithms for its computation; in section 5 then, we show that the problem of learning constrained network parameters is a special case of isotonic regression. We study the complexity of one of the algorithms in section 6. In section 7, we report on experiments that we performed on our small example network. In the final section, we draw a number of conclusions from our work and indicate interesting directions for further research.

2 Preliminaries

We briefly review a number of concepts from the field Bayesian networks that we will use in the sequel.

2.1 Bayesian networks

A *Bayesian network* is a concise representation of a joint probability distribution over a set of stochastic variables $\mathbf{X} = (X_1, \ldots, X_m)$. In the sequel, we assume all variables to be binary, adopting one of the values 0 and 1; slightly abusing terminology, we will sometimes say that X_i *occurs* or *is present* if it has the value 1. The network consists of a directed acyclic graph in which each node corresponds with a variable and the arcs capture the qualitative dependence structure of the distribution. The network further includes a number of conditional probabilities, or *parameters*, $p(X_i \mid \mathbf{X}_{\pi(i)})$ for each variable X_i given its parents $\mathbf{X}_{\pi(i)}$ in the graph. The graphical structure and associated probabilities with each other represent a unique joint probability distribution $\Pr(\mathbf{X})$ over the variables involved, which is factorised according to

$$\Pr(\mathbf{X}) = \prod_{i=1}^{m} p(X_i \mid \mathbf{X}_{\pi(i)})$$

2.2 Parameter estimation

The parameters of a Bayesian network can be estimated from a data sample $\mathcal{D}$; in this paper we assume that the sample does not include any missing values. Let $n(x_i, \mathbf{x}_{\pi(i)})$ denote the number of observations in $\mathcal{D}$ in which the variable X_i has adopted the value x_i and its parents $\mathbf{X}_{\pi(i)}$ have taken for their values the configuration $\mathbf{x}_{\pi(i)}$. Under the assumption that the various different parameters of a network are unrelated, their estimation decomposes into a number of independent estimation problems, one for each variable and possible parent configuration. The standard estimate for a parameter $p(x_i \mid \mathbf{x}_{\pi(i)})$ is

$$\hat{p}(x_i \mid \mathbf{x}_{\pi(i)}) = \frac{n(x_i, \mathbf{x}_{\pi(i)})}{n(\mathbf{x}_{\pi(i)})}$$

This estimate has been shown to maximise the log-likelihood $\ell(\mathbf{p} \mid \mathcal{D})$ of the network's parameters $\mathbf{p}$ given the available data, where

$$\ell(\mathbf{p} \mid \mathcal{D}) = \sum_{i=1}^{m} \sum_{x_i, \mathbf{x}_{\pi(i)}} n(x_i, \mathbf{x}_{\pi(i)}) \cdot \log p(x_i \mid \mathbf{x}_{\pi(i)})$$

2.3 Qualitative influences

A Bayesian network in essence models the probabilistic influences between its variables. The concept of qualitative influence has been designed to describe these influences in a qualitative way [5]. A *qualitative influence* between two variables expresses how observing a value for the one variable affects the probability distribution for the other variable. A positive qualitative influence of a variable X on a variable Y along an arc $X \rightarrow Y$ in the network, then means that the occurrence of X increases the probability of Y occurring, assuming that the values of the other parents of Y remain the same, that is,

$$p(Y = 1 \mid X = 1, \mathbf{s}) \geq p(Y = 1 \mid X = 0, \mathbf{s})$$

for any combination of values $\mathbf{s}$ for the set of parents of Y other than X. Similarly, there is a negative influence between X and Y along the arc $X \rightarrow Y$ if the occurrence of X decreases the probability of Y occurring, that is, if

$$p(Y = 1 \mid X = 1, \mathbf{s}) \leq p(Y = 1 \mid X = 0, \mathbf{s})$$

for any combination $\mathbf{s}$. A positive qualitative influence of X on Y is denoted by $X \xrightarrow{+} Y$ and a negative influence by $X \xrightarrow{-} Y$.

3 A motivating example

As an example of the effect of parameter estimates from small data samples, we consider a small Bayesian network in the medical domain. We take the following fragment of medical knowledge adapted from [6]:

> Consider a primary tumour with an uncertain prognosis in an arbitrary patient. The cancer can metastasize to the brain and to other sites. Metastatic cancer (MC) may be detected by an increased level of serum calcium (ISC). The presence of a brain tumour (B) may be established from a CT scan (CT). Severe headaches (SH) are indicative of the presence of a brain tumour. Both a brain tumour and an increased level of serum calcium are likely to cause a patient to fall into a coma (C) in due course.

From the domain knowledge, the graphical structure depicted in figure 1 has been configured. A domain expert in addition has provided the signs of the various qualitative influences in the network; these signs are shown in the figure over the graph's arcs. The positive sign of the influence of the variable B on the variable C for example expresses that the presence of a brain tumour increases the probability of the patient falling into a coma, regardless of his level of serum calcium.

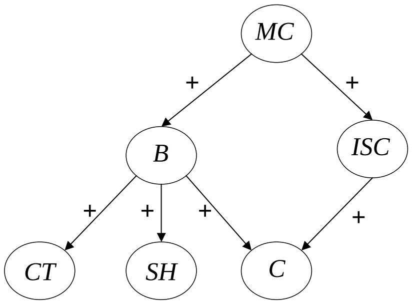

Figure 1: The *Brain Tumour* network

Now suppose that just a small data sample, of 50 patient cases, is available for quantification purposes, from which the following maximum-likelihood estimates for the parameters $p(C \mid B, ISC)$ are obtained:

$$\begin{array}{ll} \hat{p}_{C|B,ISC}(1 \mid 1,1) = \frac{1}{2} = 0.5 & \qquad \hat{p}_{C|B,ISC}(1 \mid 0,1) = \frac{13}{14} = 0.929 \\ \hat{p}_{C|B,ISC}(1 \mid 1,0) = \frac{1}{2} = 0.5 & \qquad \hat{p}_{C|B,ISC}(1 \mid 0,0) = \frac{1}{32} = 0.031 \end{array}$$

In the network supplemented with these estimates, a patient with an increased calcium level and no brain tumour would have a 93% chance of falling into a coma, whereas a patient with an increased calcium level as well as a brain tumour would see this probability drop to 50%. These inference results clearly violate the qualitative knowledge provided by the domain expert. It will further be evident that a network from which such counterintuitive inferences are made, would not be easily accepted by any physician.

4 Isotonic Regression

Our approach to obtaining parameter estimates for a Bayesian network that satisfy the qualitative influences specified by experts, is a special case of isotonic regression [7]. In this section we review isotonic regression in general; in the next section we discuss its application to parameter estimation for Bayesian networks.

Let $Z = \{z_1, z_2, \ldots, z_n\}$ be a nonempty finite set of constants and let $\preceq$ be a partial order on Z. Any real-valued function f on Z is called *isotonic* with respect to $\preceq$ if, for any $z, z' \in Z$, $z \preceq z'$ implies $f(z) \leq f(z')$. We assume that each element z_i of Z is associated with a real number $g(z_i)$; these numbers typically are estimates of the function values of an unknown isotonic function on Z. Each element of Z further has associated a positive weight $w(z_i)$ that indicates the precision of this estimate. An isotonic function g^* on Z now is an *isotonic regression* of g with respect to the weight function w and the partial order $\preceq$ if and only if it minimises the sum

$$\sum_{i=1}^{n} w(z_i) \cdot [f(z_i) - g(z_i)]^2$$

within the class of isotonic functions f on Z. The existence of a unique g^* has been proved by Brunk [8].

Isotonic regression provides a solution to many estimation problems in which we have prior knowledge about the order of the parameters to be estimated. As an example, we suppose that we would like to estimate a vector of means

$$\boldsymbol{\mu} = (\mu(z_1), \mu(z_2), \ldots, \mu(z_n))$$

where $\mu(z_i)$ denotes the mean in population z_i. We assume that an expert has provided prior knowledge about the order of these means, which amounts to $\mu(z_i)$ being isotonic with respect to some partial order $\preceq$ on the collection of populations Z. Let n_i denote the number of observations sampled from population z_i, and let the observations Y_{ij} be normally distributed with $Y_{ij} \sim N(\mu(z_i), \sigma^2)$, $j = 1, \ldots, n_i$. Then, the isotonic regression of the estimates $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$ with weights $w(z_i) = n_i$ provides the order-constrained maximum-likelihood estimate of $\boldsymbol{\mu}$. As another example, we suppose that we want to estimate binomial parameters

$$\mathbf{p} = (p(z_1), p(z_2), \ldots, p(z_n))$$

where $p(z_i)$ denotes the probability of success in population z_i. Again, we assume that an expert has provided prior knowledge about the order of these probabilities, which amounts to $p(z_i)$ being isotonic with respect to some partial order $\preceq$ on Z. Let n_i denote the number of observations sampled from population z_i, and let the number of successes Y_i be binomially distributed with $Y_i \sim B(n_i, p(z_i))$. Then, the isotonic regression of the estimates $\bar{Y}_i = Y_i/n_i$ with weights $w(z_i) = n_i$ provides the order-constrained maximum-likelihood estimate of $\mathbf{p}$. Note that the examples suggest that the order-constrained estimates are obtained by first computing the unconstrained estimates and then performing the isotonic regression of these estimates with appropriate weights.

The problem of computing the isotonic regression can be solved by quadratic programming methods. Various dedicated algorithms, often restricted to a particular type of order, have been proposed as well. For Z linearly ordered, for example, the *pool adjacent violators* (PAV) algorithm was developed [9]. This algorithm considers a nonempty finite set of constants $Z = \{z_1, z_2, \ldots, z_n\}$ with the *total* order $z_1 \preceq z_2 \preceq \ldots \preceq z_n$. Associated with the set are vectors $\mathbf{g} = (g(z_1), \ldots, g(z_n))$ and $\mathbf{w} = (w(z_1), \ldots, w(z_n))$ with $g(z_i)$ and $w(z_i)$ as before. Within the algorithm, the *weighted average* of the function g over the subinterval, or *block*, $[p, q]$ of $[1, n]$ with $p \leq q$, is defined as

$$\mathrm{Av}(p, q) = \frac{\sum_{j=p}^{q} w(z_j) \cdot g(z_j)}{\sum_{j=p}^{q} w(z_j)}$$

The PAV algorithm now computes the isotonic regression $\mathbf{g}^*$ of $\mathbf{g}$ with respect to $\mathbf{w}$ and $\preceq$ as follows:

```
**PoolAdjacentViolators** ($\mathbf{g}$, $\mathbf{w}$)
Blocks = $[[1,1], [2,2], \ldots, [n,n]]$
**For** each block $[i,i]$ **do**
    $w[i,i] = w(z_i)$
    $\mathrm{Av}[i,i] = g(z_i)$
**od**
**While** there are blocks $[p,q], [q+1,r]$ with $\mathrm{Av}[p,q] > \mathrm{Av}[q+1,r]$ **do**
    Replace blocks $[p,q]$ and $[q+1,r]$ by block $[p,r]$ with
    weight $w[p,r] = w[p,q] + w[q+1,r]$ and
    weighted average $\mathrm{Av}[p,r] = \frac{w[p,q] \cdot \mathrm{Av}[p,q] + w[q+1,r] \cdot \mathrm{Av}[q+1,r]}{w[p,q] + w[q+1,r]}$
**od**
**For** each block $[p,q]$ **do**
    $g^*(z_j) = \mathrm{Av}[p,q]$ for all $j \in [p,q]$
**od**
**Return** $\mathbf{g}^*$
```

The algorithm does exactly what its name suggests: as long as there are two adjacent blocks whose weighted averages of the function g violate the imposed order constraints, these two blocks are *pooled*, that is, they are merged into a single block with a new weight and a new average. The PAV algorithm thus resolves violations of the order constraints by averaging the function values of g over consecutive elements of Z. For the final solution, the set Z is partitioned into sets of consecutive elements, called *solution blocks*, on which the isotonic regression $\mathbf{g}^*$ is constant and equal to the weighted average of the original values of g within that block. Note that if $\mathbf{g}$ already satisfies the imposed order constraints, then the PAV algorithm simply returns $\mathbf{g}^* = \mathbf{g}$. The algorithm is readily shown to have a time complexity that is linear in the size of the set of constants [10].

The PAV algorithm requires a total order on the set of constants for which an isotonic regression is to be computed. For our application, however, we require an algorithm that is applicable to sets of constants with arbitrary *partial* orders. For this purpose we will use the *minimum lower sets* (MLS) algorithm proposed by Brunk [11]. This algorithm builds upon the concept of lower set. A subset L of Z is a *lower set* of Z if $z \in L$, $z' \in Z$, and $z' \preceq z$ imply $z' \in L$. Within the algorithm, the *weighted average* of a function g on Z for a nonempty subset A of Z is defined as

$$\mathrm{Av}(A) = \frac{\sum_{z \in A} w(z) g(z)}{\sum_{z \in A} w(z)}$$

The algorithm takes for its input the set of constants $Z = \{z_1, z_2, \ldots, z_n\}$ with an arbitrary partial order $\preceq$. With the set Z again are associated a vector $\mathbf{w}$ of weights and a vector $\mathbf{g}$ of real numbers. The algorithm returns the isotonic regression $\mathbf{g}^*$ of $\mathbf{g}$ with respect to $\mathbf{w}$ and $\preceq$.

Just like the PAV algorithm, the MLS algorithm resolves violations of the order constraints by averaging over suitably chosen subsets of Z. For the final solution, it partitions the set Z into a number of subsets on which the isotonic regression is constant. These subsets are no longer blocks of consecutive elements of Z however, but lower sets. The first subset B_1 in the final solution is a lower set of $(Z, \preceq)$. The second subset is a lower set of $(Z \setminus B_1, \preceq_2)$, where the partial order $\preceq_2$ is obtained from $\preceq$ by removing all order relations involving elements of B_1. This process is continued until the set Z is exhausted. In each iteration the lower set with minimum weighted average is selected for the solution; in case multiple lower sets attain the same minimum, their union is taken.

```
MinimumLowerSets($Z$, $\preceq$, $\mathbf{g}$, $\mathbf{w}$)
$\mathcal{L}$ = Collection of all lower sets of $Z$ wrt $\preceq$
Repeat
    $B = \bigcup\{A \in \mathcal{L} \mid \mathrm{Av}(A) = \min_{L \in \mathcal{L}} \mathrm{Av}(L)\}$
    For each $z \in B$ do
        $g^*(z) = \mathrm{Av}(B)$
    For each $L \in \mathcal{L}$ do
        $L = L \setminus B$
    $Z = Z \setminus B$
Until $Z = \varnothing$
Return $\mathbf{g}^*$
```

The bottleneck of the algorithm from a computational point of view clearly is the generation of the lower sets, which is exponential in the size of the set of constants. We will return to this issue presently.

5 Learning network parameters with order constraints

We address the maximum-likelihood estimation of parameters for a Bayesian network subject to the constraints that are imposed by the signs that have been provided by experts for the network's qualitative influences. We show more specifically that this constrained estimation can be viewed as a special case of isotonic regression. Before doing so, we would like to note that in the presence of qualitative influences the parameters associated with the different parent configurations for a variable are no longer unrelated; in fact, only those combinations of parameter values are feasible that are isotonic with respect to the order imposed by the signs of the influences. The parameters associated with different variables are still unrelated however, because the sign of an influence imposes constraints on the parameters for a single variable only. We restrict our presentation therefore to a single variable and its parents.

5.1 Order constraints imposed by the signs of qualitative influences

We consider a variable Y with the parents $X_1, \ldots, X_k$. Let $\Omega(X_i) = \{0, 1\}$ denote the domain of values for the parent X_i and let $\Omega(\mathbf{X}) = \Omega(X_1) \times \ldots \times \Omega(X_k) = \{0, 1\}^k$ consist of vectors $\mathbf{x} = (x_1, \ldots, x_k)$ of values for all k parents, that is, $\Omega(\mathbf{X})$ is the set of all parent configurations for Y. We now use the signs that have been specified for the qualitative influences on Y to define an order on its parent configurations. We take a positive qualitative influence of X_i on Y to impose the order $\leq$ with $0 \leq 1$ on $\Omega(X_i)$; a negative influence of X_i is taken to impose $1 \leq 0$ on $\Omega(X_i)$. If no sign for the influence of X_i on Y has been specified, we have that neither $0 \leq 1$ nor $1 \leq 0$, that is, the values 0 and 1 are taken to be incomparable. We then say that the influence is *unsigned*; positive and negative influences are called *signed* influences. The signs of the separate qualitative influences of $X_1, \ldots, X_k$ on Y now impose a partial order $\preceq$ on $\Omega(\mathbf{X})$ where for any $\mathbf{x}, \mathbf{x}' \in \Omega(\mathbf{X})$ we have that

$$\mathbf{x} = (x_1, \ldots, x_k) \preceq \mathbf{x}' = (x'_1, \ldots, x'_k)$$

if and only if $x_i \leq x'_i$ for all i. In the sequel, we will use the signs that have been specified for the various influences to constrain the parameters $p(y = 1 \mid \mathbf{x})$ to be non-decreasing on $(\Omega(\mathbf{X}), \preceq)$.

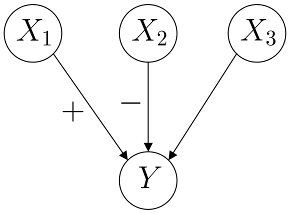

Figure 2: A Bayesian network fragment with qualitative influences

Throughout the paper, we will assume that a domain expert specifies the signs of the qualitative influences between the variables in a network. We would like to mention that for real-life applications these signs are quite readily obtained from experts by using a special-purpose elicitation technique tailored to the acquisition of probability orders [4].

5.2 Parameter estimation with order constraints

In the previous section, we have established the partial order $\preceq$ that is imposed by the signs of the qualitative influences on a variable Y, on its parent configurations. We now exploit this partial order to obtain parameter estimates that reflect the qualitative knowledge that has been specified through these signs. We recall that the unconstrained maximum-likelihood estimate of a parameter $p(y = 1 \mid \mathbf{x})$ equals

$$\hat{p}(y = 1 \mid \mathbf{x}) = \frac{n(Y = 1, \mathbf{x})}{n(\mathbf{x})}$$

The isotonic regression $\mathbf{p}^*$ of the estimates $\hat{p}$ with weights $w(\mathbf{x}) = n(\mathbf{x})$ now provides the maximum-likelihood estimates of the parameters $p(y = 1 \mid \mathbf{x})$, subject to the specified order constraints [12,7(page 32)].

To illustrate the construction of the partial order on $\Omega(\mathbf{X})$ and the computation of the isotonic estimates, we consider a small fragment of a Bayesian network. The network includes a variable Y and its three parents X_1, X_2 and X_3, as depicted in figure 2; the parent X_1 has a positive qualitative influence on Y, X_2 has a negative influence on Y, and the influence of X_3 on Y is unsigned. Figure 3 now shows the partial order that is imposed, by the specified signs, on the various parent configurations for Y, where an arrow from a configuration $\mathbf{x}$ to a configuration $\mathbf{x}'$ indicates that $\mathbf{x}$ immediately precedes $\mathbf{x}'$ in the ordering. Note that since the influence of the parent X_3 on Y is unsigned, any two parent configurations that differ in their value for X_3 are incomparable. As a consequence, the set of configurations $\Omega(\mathbf{X})$ is partitioned into two disjoint subsets, one for $X_3 = 0$ and one for $X_3 = 1$, such that no element of the first subset is order

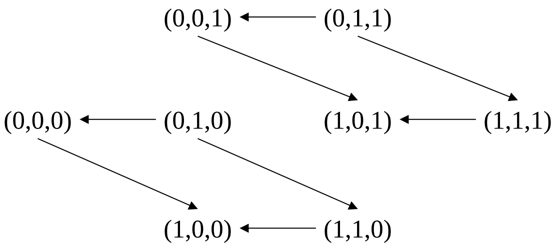

Figure 3: The partial order on the parent configurations for Y

Table 1: The counts and unconstrained ML estimates for $y = 1$

$X_3 = 0$	$X_2 = 0$	$X_2 = 1$	$X_3 = 1$	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	$4/10 = 0.4$	$6/20 = 0.3$	$X_1 = 0$	$4/20 = 0.2$	$3/5 = 0.6$
$X_1 = 1$	$12/15 = 0.8$	$24/40 = 0.6$	$X_1 = 1$	$9/10 = 0.9$	$20/40 = 0.5$

related to any element of the second subset. Constrained estimates may therefore be computed for the two subsets separately, that is, as if they were distinct spaces [12].

Now suppose that we have a small sample of data available with respect to the four variables. Table 1 shows the counts, per parent configuration, that are obtained from the sample as well as the associated unconstrained maximum-likelihood estimates $\hat{p}(y = 1 \mid \mathbf{x})$. We recall that the estimates are constrained to be non-increasing in each row and non-decreasing within each column for both $X_3 = 0$ and $X_3 = 1$. From the table, we observe that, for $X_3 = 0$, the obtained estimates $\hat{p}$ satisfy the order constraints: $\hat{p}$ is decreasing within each row and increasing within each column of the table. So, for $X_3 = 0$, the isotonic regression $\mathbf{p}^*$ equals the basic estimate $\hat{\mathbf{p}}$. For $X_3 = 1$ however, the maximum-likelihood estimates obtained from the sample do not satisfy the constraints imposed by the signs of the influences: we observe that $\hat{p}$ increases within the first row and decreases within the second column of the table.

We now compute the isotonic regression $\mathbf{p}^*$ of the parameter estimates given $X_3 = 1$, by applying the minimum lower sets algorithm. The algorithm starts with the computation of the weighted average of the unconstrained estimates $\hat{p}$ for all lower sets. Table 2 reports the various lower sets and their associated averages in the subsequent iterations of the algorithm. In the first iteration, the lower set with minimum weighted average is $\{(0, 1, 1), (0, 0, 1)\}$, and the algorithm sets $p^*(y = 1 \mid 0, 1, 1) = p^*(y = 1 \mid 0, 0, 1) = 0.28$. After removal of the two elements $(0, 1, 1)$ and $(0, 0, 1)$ from the other sets, the lower sets $\{(1, 1, 1)\}$ and $\{(1, 1, 1), (1, 0, 1)\}$ remain, with 0.5 and 0.58 respectively, for their weighted averages. The algorithm sets $p^*(y = 1 \mid 1, 1, 1) = 0.5$. After removal of $(1, 1, 1)$ from the single remaining set, only the lower set $\{(1, 0, 1)\}$ is left. The algorithm sets $p^*(y = 1 \mid 1, 0, 1) = 0.9$ and halts since it has exhausted $\Omega(\mathbf{X})$. The thus computed isotonic estimates are summarised in table 3. Note that the computed estimates satisfy the constraints imposed by the signs of the qualitative influences involved: p^* does no longer increase within the first row nor decrease in the second column of the table.

5.3 Zero counts

In practice, a sample of data may include no observations at all for a particular parent configuration $\mathbf{x}$ for the variable Y of interest. In fact, the smaller the data sample, the more likely this situation is to occur. Disregarding any constraints there may be, the lack of observations means that there are no data to estimate the parameter $p(y = 1 \mid \mathbf{x})$ from. In line with common practice, we then set $\hat{p}(y = 1 \mid \mathbf{x}) = 0.5$, that is, we assume a uniform distribution for the variable Y given

Table 2: The weighted averages of the unconstrained estimates for the lower sets given $X_3 = 1$

Lower Set	Av1	Av2	Av3
$\{(0, 1, 1)\}$	$3/5 = 0.6$	–	–
$\{(0, 1, 1), (0, 0, 1)\}$	$7/25 = 0.28$	–	–
$\{(0, 1, 1), (1, 1, 1)\}$	$23/45 = 0.51$	$20/40 = 0.5$	–
$\{(0, 1, 1), (0, 0, 1), (1, 1, 1)\}$	$27/65 = 0.42$	$20/40 = 0.5$	–
$\{(0, 1, 1), (0, 0, 1), (1, 1, 1), (1, 0, 1)\}$	$36/75 = 0.48$	$29/50 = 0.58$	$9/10 = 0.9$

Table 3: The isotonic estimates p^* for $y = 1$, given the parent configurations with $X_3 = 1$

$X_3 = 1$	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	$7/25 = 0.28$	$7/25 = 0.28$
$X_1 = 1$	$9/10 = 0.9$	$20/40 = 0.5$

x. This estimate may however lead to a violation of the constraints imposed by the signs of the qualitative influences involved. Whenever it does so, it should be adjusted by the minimum lower sets algorithm, just like any other estimate that results in such a violation. We can easily accommodate for a lack of observations for a parent configuration by using a slightly modified weighted average function within the algorithm:

$$\mathrm{Av}(A) = \begin{cases} \dfrac{\sum_{z\in A} w(z)\cdot g(z)}{\sum_{z\in A} w(z)} & \text{if } \sum_{z\in A} w(z) > 0 \\ 0.5 & \text{if } \sum_{z\in A} w(z) = 0 \end{cases}$$

As an example we consider the data and associated parameter estimates from table 4. Note that since there are no observations for the parent configuration $(0, 0)$, we have set $\hat{p}(y = 1 \mid 0, 0) = 0.5$. We assume that the expert has specified positive influences of both X_1 and X_2 on Y. We now find that the value of $\hat{p}(y = 1 \mid 0, 0)$ violates the order constraints. The lower set with minimum weighted average is

$$\mathrm{Av}(\{(0, 0), (0, 1)\}) = \frac{0 \cdot 0.5 + 10 \cdot 0.4}{0 + 10} = 0.4$$

The minimum lower sets algorithm thus yields $p^*(y = 1 \mid 0, 0) = p^*(y = 1 \mid 0, 1) = 0.4$. Note that any value for $p^*(y = 1 \mid 0, 0)$ in the interval $[0, 0.4]$ is equally good as far as the data and specified constraints are concerned.

5.4 Estimation with complete order

So far we have assumed that the signs that are specified by experts for the qualitative influences on a specific variable, impose a partial order on its parent configurations. In some cases, however, an expert can indicate a total order. In these cases, the pool adjacent violators algorithm can be used to obtain order-constrained maximum-likelihood parameter estimates. As an example, we consider a variable with three parents. We suppose that the expert has specified the following total order on the parent configurations:

$$(0, 0, 0) \preceq (0, 0, 1) \preceq (0, 1, 0) \preceq (1, 0, 0) \preceq (0, 1, 1) \preceq (1, 0, 1) \preceq (1, 1, 0) \preceq (1, 1, 1)$$

Table 4: An example with no observations for the parent configuration $(0, 0)$

X_1	X_2	$n(X_1, X_2)$	$n(Y = 1, X_1, X_2)$	$\hat{p}(y = 1 \mid X_1, X_2)$
0	0	0	0	0.5
0	1	10	4	0.4
1	0	20	9	0.45
1	1	20	16	0.8

We further suppose that we have the following data:

$$\hat{\mathbf{p}} = (0.1, 0.05, 0.2, 0.4, 0.35, 0.5, 0.7, 0.8)$$
$$\mathbf{n} = (10, 20, 20, 10, 20, 30, 30, 10)$$

The vector of estimates includes two violations of the specified order; these are between elements one and two, and between elements four and five. These violations are resolved by the PAV algorithm by averaging the elements:

$$\mathbf{p}^* = (0.067, 0.067, 0.2, 0.37, 0.37, 0.5, 0.7, 0.8)$$

Since it requires much less computational effort than the minimum lower sets algorithm, exploiting the PAV algorithm whenever a total order is available, may drastically reduce the overall runtime of learning a network's parameters.

5.5 Bayesian estimation

The parameter-learning method described above does not require that an expert specifies numerical values for the parameters concerned: he only has to provide signs for the various influences. If the expert would be able to provide uncertain prior knowledge about the numerical values in addition to these signs, then this knowledge is readily accommodated in a Bayesian approach to constrained parameter estimation.

Suppose that the expert has specified Beta priors for the parameters $p(y = 1 \mid \mathbf{x})$. We assume that he has chosen the hyperparameters $a(\mathbf{x})$ and $b(\mathbf{x})$ for these priors given $\mathbf{x}$ in such a way that his experience is equivalent to having seen the value $y = 1$ a total of $a(\mathbf{x}) - 1$ times in $h(\mathbf{x}) = a(\mathbf{x}) + b(\mathbf{x}) - 2$ observations of $\mathbf{x}$; h is called the precision of the specified prior given $\mathbf{x}$. Now, let $p_0(y = 1 \mid \mathbf{x})$ denote the modal value of the Beta prior for $p(y = 1 \mid \mathbf{x})$, that is, $p_0(y = 1 \mid \mathbf{x})$ is a priori considered the most likely value of the parameter $p(y = 1 \mid \mathbf{x})$. We suppose that the expert's values for $a(\mathbf{x})$ and $b(\mathbf{x})$ are such that the modes $p_0(y = 1 \mid \mathbf{x}) = (a(\mathbf{x}) - 1)/h(\mathbf{x})$ are isotonic with respect to the order imposed by the signs that he has specified. In forming the joint prior for the parameters, we now assume local independence [13], except for the constraint that the parameter values must be isotonic. The joint prior thus equals 0 for non-isotonic value combinations for the parameters and is proportional to the product Beta distribution for isotonic combinations. The constrained MAP estimates given the data $\mathcal{D}$ then are given by the isotonic regression of

$$p_0(y = 1 \mid \mathbf{x}, \mathcal{D}) = \frac{n(\mathbf{x}) \cdot \hat{p}(y = 1 \mid \mathbf{x}) + h(\mathbf{x}) \cdot p_0(y = 1 \mid \mathbf{x})}{n(\mathbf{x}) + h(\mathbf{x})}$$

with weights $n(\mathbf{x}) + h(\mathbf{x})$ [14]. Order-constrained estimation thus again amounts to performing isotonic regression on basic estimates with appropriately chosen weights. The unconstrained MAP estimates $p_0(y = 1 \mid \mathbf{x}, \mathcal{D})$ for the various parameters are taken for the basic estimates. The weights for these estimates are $n(\mathbf{x}) + h(\mathbf{x})$, that is, for a parent configuration $\mathbf{x}$, the weight is taken to be the sum of the number of actual observations for $\mathbf{x}$ and the precision $h(\mathbf{x})$ specified for his prior estimate by the expert. Note that if the expert has specified a flat prior Beta(1,1) with $h = 0$, then the order-constrained maximum-likelihood estimates are returned.

As an example, we consider again the network fragment from figure 2 and the data from table 1. We suppose that the expert has specified the following Beta priors for the various parameters given $X_3 = 1$: $p(y = 1 \mid 0, 0, 1) \sim \text{Beta}(5, 7)$, $p(y = 1 \mid 1, 0, 1) \sim \text{Beta}(9, 3)$, $p(y = 1 \mid 0, 1, 1) \sim \text{Beta}(2, 5)$, and $p(y = 1 \mid 1, 1, 1) \sim \text{Beta}(3, 4)$. The modes corresponding with these priors are given in table 5 on the left. Note that the prior modes are consistent with the order implied by the signs that have been specified by the expert. Combination with the observed data results in the posterior modes given in table 5 on the right. We find that these posterior modes no longer satisfy the order constraints, as they are increasing in the first row. Application of the minimum lower sets algorithm resolves the violation by averaging the posterior modes over the first row to obtain $p_0^*(y = 1 \mid 0, 0, 1) = p_0^*(y = 1 \mid 0, 1, 1) = 12/40 = 0.3$.

Table 5: Prior (left) and posterior (right) modes of the Beta distributions for the example

$X_3 = 1$	$X_2 = 0$	$X_2 = 1$	$X_3 = 1$	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	$4/10 = 0.4$	$1/5 = 0.2$	$X_1 = 0$	$8/30 = 0.27$	$4/10 = 0.4$
$X_1 = 1$	$8/10 = 0.8$	$2/5 = 0.4$	$X_1 = 1$	$17/20 = 0.85$	$22/45 = 0.49$

6 Complexity of the Minimum Lower Sets Algorithm

The dominant factor in the runtime complexity of the minimum lower sets algorithm is the number of lower sets involved. To address this number, we observe that for k binary parents, each element of $\Omega(\mathbf{X})$ is uniquely determined by the parents that have the value 1, that is, by a subset of $\{1, 2, \ldots, k\}$. The partial order that is imposed on $\Omega(\mathbf{X})$ by positive influences of all parents therefore, is isomorphic to the partial order generated by set inclusion on $\mathcal{P}(\{1, 2, \ldots, k\})$. Every lower set of $\Omega(\mathbf{X})$ now corresponds uniquely to an antichain in this partial order. The number of distinct lower sets of $\Omega(\mathbf{X})$ thus equals the number of distinct nonempty antichains of subsets of a k-set, which adheres to Sloane sequence A014466 [15]. Table 6 gives the number of distinct lower sets for different numbers of parents under the assumption that each parent exerts a signed influence on the variable for which we want to compute parameter estimates. From the table, we note that the minimum lower sets algorithm is feasible only for up to five or six parents with signed influences.

From our example network fragment, we have noted that unsigned influences serve to partition the set of parent configurations $\Omega(\mathbf{X})$ into disjoint subsets, such that no element of the one subset is order related to any element of the other subsets. We have argued that constrained estimates may be computed for these subsets separately, thereby effectively decomposing the parameter-learning problem into a number of independent smaller problems. This decomposition yields a considerable improvement of the overall runtime involved. Let k_1 denote the number of parents with a signed influence and let k_2 denote the number of parents with an unsigned influence. The number of configurations for the parents with an unsigned influence equals 2^{k_2}. The order graph therefore has 2^{k_2} separate components. If we would not exploit the problem's decomposition, the algorithm would have to establish the weighted average of

$$|\mathcal{L}(k_1 + k_2)| = (|\mathcal{L}(k_1)| + 1)^{2^{k_2}} - 1$$

lower sets, since a lower set of the entire set of parent configurations $\Omega(\mathbf{X})$ is constructed by arbitrarily combining lower sets of the 2^{k_2} subsets induced by the unsigned influences. By treating each of these subsets as a separate problem, the algorithm initially has to compute the weighted average of

$$|\mathcal{L}(k_1 + k_2)| = 2^{k_2} \cdot |\mathcal{L}(k_1)|$$

Table 6: The number of lower sets ($|\mathcal{L}|$) as a function of the number of parents with a signed influence (k)

k	$\lvert\Omega(\mathbf{X})\rvert$	$\lvert\mathcal{L}\rvert$
1	2	2
2	4	5
3	8	19
4	16	167
5	32	7580
6	64	7828353
7	128	2414682040997

lower sets. For $k_1 = 4$ and $k_2 = 3$, for example, the algorithm needs to compute the weighted average of $168^8 - 1 = 6.35 \cdot 10^{17}$ lower sets for the non-decomposed problem and $8 \cdot 167 = 1336$ lower sets for the decomposed problem.

7 Experimental Results

To study the behaviour of the isotonic estimator in a more involved setting, we compare it to the standard maximum-likelihood estimator on the *Brain Tumour* network introduced in section 3; the network and its associated qualitative influences are depicted in figure 1. For the network, we specified two sets of parameter probabilities; the resulting models will be referred to as Model A and Model B respectively. The parameter probabilities for Model A are as follows (those for Model B are given between parentheses):

$$\begin{array}{ll}
p_{MC}(1) = .2(.4) & \\
p_{CT|B}(1 \mid 1) = .95(.75) & p_{CT|B}(1 \mid 0) = .1(.25) \\
p_{B|MC}(1 \mid 1) = .2(.3) & p_{B|MC}(1 \mid 0) = .05(.15) \\
p_{SH|B}(1 \mid 1) = .8(.7) & p_{SH|B}(1 \mid 0) = .6(.6) \\
p_{ISC|MC}(1 \mid 1) = .8(.75) & p_{ISC|MC}(1 \mid 0) = .2(.25) \\
p_{C|B,ISC}(1 \mid 1,1) = .8(.7) & p_{C|B,ISC}(1 \mid 0,1) = .8(.7) \\
p_{C|B,ISC}(1 \mid 1,0) = .8(.7) & p_{C|B,ISC}(1 \mid 0,0) = .05(.3)
\end{array}$$

Note that for model B we have chosen less extreme probabilities, that is, closer to 0.5, than for Model A. We drew samples of different sizes, $n = 50, 150, 500, 1500$, from the two models, using logic sampling; for each sample size, 100 samples were drawn. From each sample, both the standard maximum-likelihood estimates and the constrained estimates of the various parameters were calculated; for this purpose, the minimum lower sets algorithm as well as an algorithm for efficiently generating the lower sets [16], were implemented in Splus. Given the computed parameter estimates, the joint distribution defined by the resulting network was established. This distribution then was compared to the true joint distribution defined by the original network. To compare the true and estimated distributions we used the well-known Kullback-Leibler divergence. The Kullback-Leibler divergence of Pr' from Pr is defined as

$$\mathrm{KL}(\mathrm{Pr}, \mathrm{Pr}') = \sum_{\mathbf{x}} \mathrm{Pr}(\mathbf{x}) \cdot \log \frac{\mathrm{Pr}(\mathbf{x})}{\mathrm{Pr}'(\mathbf{x})}$$

where a term in the sum is taken to be 0 if $\mathrm{Pr}(\mathbf{x}) = 0$, and infinity whenever $\mathrm{Pr}'(\mathbf{x}) = 0$ and $\mathrm{Pr}(\mathbf{x}) > 0$. The results are summarised in table 7. The columns labeled $\mathrm{KL} < \infty$ indicate for how many samples the KL divergence was smaller than infinity for both the maximum-likelihood and the isotonic estimator; the reported numbers are averages over these samples.

The results reveal that the isotonic estimator scores better than the standard maximum-likelihood estimator, although the differences are rather small. For the smaller samples the differences are more marked than for the larger samples. This finding conforms to our expectations, since for smaller samples the maximum-likelihood estimator has a higher probability of yielding estimates that violate the constraints. For larger samples, the standard estimator and the isotonic estimator are expected to often result in the same estimates.

To illustrate the benefits of our isotonic estimator in terms of acceptance of the resulting network, we consider again the counterintuitive example from section 3. The data in that example included two patients who both had a brain tumour and an increased level of serum calcium. Since one of these patients fell into a coma, the maximum-likelihood estimator set the probability that a patient with this combination of symptoms falls into a coma to 0.5, that is, $\hat{p}_{C|B,ISC}(1 \mid 1,1) = \frac{1}{2} = 0.5$. With $\hat{p}_{C|B,ISC}(1 \mid 0,1) = \frac{13}{14} = 0.929$, this estimate violates the constraint that originates from the positive influence of B on C. The isotonic estimator therefore pools the two estimates to obtain $p^*_{C|B,ISC}(1 \mid 1,1) = p^*_{C|B,ISC}(1 \mid 0,1) = \frac{14}{16} = 0.875$. The counterintuitive behaviour resulting from the basic estimates has thus been eliminated.

Table 7: Results of the experiments with the *Brain Tumour* network

	Model A			Model B		
n	$\mathrm{KL}(\mathrm{Pr}, \widehat{\mathrm{Pr}})$	$\mathrm{KL}(\mathrm{Pr}, \mathrm{Pr}^*)$	$\mathrm{KL} < \infty$	$\mathrm{KL}(\mathrm{Pr}, \widehat{\mathrm{Pr}})$	$\mathrm{KL}(\mathrm{Pr}, \mathrm{Pr}^*)$	$\mathrm{KL} < \infty$
50	0.13	0.11	2	0.13	0.12	61
150	0.033	0.030	17	0.048	0.044	97
500	0.013	0.011	83	0.013	0.012	100
1500	0.0045	0.0043	100	0.0043	0.0041	100

We observe that the order-constrained estimates yielded by the isotonic estimator imply that the probability of falling into a coma for a patient with an increased serum calcium level equals 87.5% *regardless* of whether or not he has a brain tumour. Although these estimates are an improvement over the basic estimates, they may still not be entirely satisfactory. The estimates nevertheless satisfy the specified qualitative knowledge. If we want to enforce a strict increase of the probability of falling into a coma for a patient with a brain tumour compared to that for a patient without a brain tumour, then the expert will have to specify some minimum numerical difference between the two probabilities. As an illustration, we suppose that the expert specifies, in addition to the two positive influences, the following minimum difference:

$$p_{C|B,ISC}(1 \mid 1,1) - p_{C|B,ISC}(1 \mid 0,1) \geq 0.1$$

The order-constrained estimates given all available knowledge now are

$$\begin{array}{ll} p^*_{C|B,ISC}(1 \mid 1,1) = 0.914 & p^*_{C|B,ISC}(1 \mid 0,1) = 0.814 \\ p^*_{C|B,ISC}(1 \mid 1,0) = 0.5 & p^*_{C|B,ISC}(1 \mid 0,0) = 0.031 \end{array}$$

Note that the difference between the estimates for the violating parameters now equals the specified minimum difference. Since the minimum lower sets algorithm is no longer applicable when minimum differences are to be enforced, the above estimates have been computed by numerical optimization.

A problem with the suggested approach to enforcing differences is that an expert is required to specify numerical information in addition to qualitative signs. This problem can be alleviated to some extent by eliciting such information only if necessary, that is, if averaging leads to unwanted equalities. An alternative approach would be to enforce some predefined minimum difference between $p^*(y = 1 \mid \mathbf{x})$ and $p^*(y = 1 \mid \mathbf{x}')$ whenever there is an arrow from $\mathbf{x}$ to $\mathbf{x}'$ in the order graph. It is questionable however whether such an approach would yield results that are acceptable to experts in the domain of application.

8 Conclusions and Further Research

We showed that, upon estimating the parameters of a Bayesian network from data, prior knowledge about the signs of influences can be taken into account by computing order-constrained estimates. Since these isotonic estimates are consistent with the knowledge specified by experts, the resulting network is more likely to be accepted in its domain of application than a network with basic maximum-likelihood estimates. Our experimental results moreover revealed that the isotonic estimator results in a slightly improved fit of the true distribution. For smaller samples the improvement will generally be more marked than for larger samples, since smaller samples are more likely to give rise to maximum-likelihood estimates that violate the order constraints.

We see various challenging directions for further research. It would be interesting, for example, to investigate whether other types of constraint, such as additive synergies and product synergies, can be exploited to further improve parameter learning. Another interesting extension of our method would be to allow for non-binary variables with linearly-ordered discrete values. An influence on such a variable is defined in terms of stochastic dominance of the distributions involved, which in essence also imposes a constraint on the estimates. The use of qualitative influences to improve parameter learning from incomplete data in our opinion also merits further investigation.

References

[1] J. Pearl, Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann, 1988.

[2] M. Druzdzel, Probabilistic reasoning in decision support systems: From computation to common sense, Ph.D. thesis, Department of Engineering and Public Policy, Carnegie Mellon University (1993).

[3] M. Druzdzel, L. van der Gaag, Building probabilistic networks: "Where do the numbers come from ?" Guest editors introduction, IEEE Transactions on Knowledge and Data Engineering 12 (2000) 481–486.

[4] E. Helsper, L. van der Gaag, F. Groenendaal, Designing a procedure for the acquisition of probability constraints for Bayesian networks, in: E. Motta, N. Shadbolt, A. Stutt, N. Gibbins (Eds.), Engineering Knowledge in the Age of the Semantic Web: 14th International Conference, Springer, 2004, pp. 280–292.

[5] M. Wellman, Fundamental concepts of qualitative probabilistic networks, Artificial Intelligence 44 (1990) 257–303.

[6] G. Cooper, NESTOR: a computer-based medical diagnostic aid that integrates causal and probabilistic knowledge, Tech. Rep. HPP-84-48, Stanford University (1984).

[7] T. Robertson, F. Wright, R. Dykstra, Order Restricted Statistical Inference, Wiley, 1988.

[8] H. Brunk, Conditional expectation given a σ-lattice and applications, Annals of Mathematical Statistics 36 (1965) 1339–1350.

[9] M. Ayer, H. Brunk, G. Ewing, W. Reid, E. Silverman, An empirical distribution function for sampling with incomplete information, Annals of Mathematical Statistics 26 (1955) 641–647.

[10] R. Ahuja, J. Orlin, A fast scaling algorithm for minimizing separable convex functions subject to chain constraints, Operations Research 49 (5) (2001) 784–789.

[11] H. Brunk, Maximum likelihood estimates of monotone parameters, Annals of Mathematical Statistics 26 (1955) 607–616.

[12] C. van Eeden, Maximum likelihood estimation of ordered probabilities, in: Proceedings Koninklijke Nederlandse Akademie van Wetenschappen A, 1956, pp. 444–455.

[13] D. Heckerman, D. Geiger, D. Chickering, Learning Bayesian networks: The combination of knowledge and statistical data, Machine Learning 20 (1995) 197–243.

[14] R. Barlow, D. Bartholomew, J. Bremner, H. Brunk, Statistical Inference under Order Restrictions, Wiley, 1972.

[15] N. Sloane, The on-line encyclopedia of integer sequences, http://www.research.att.com/~njas/sequences/ .

[16] G. Steiner, An algorithm to generate the ideals of a partial order, Operations Research Letters 5 (6) (1986) 317–320.