An upper bound for BDeu local scores

James Cussens[1]

Abstract. An upper bound on BDeu log local scores is derived using an existing upper bound on the beta function with r variables. Two bounds on this bound are derived, one of which is suitable for pruning the search for optimal parent sets of a variable in Bayesian network learning. Empirical results concerning the tightness of bounds are given.

1 BDeu log local scores

If Dirichlet parameters are used for the parameters of a BN and the data D is complete then the log marginal likelihood for BN structure G with variables $i = 1, \ldots, p$ is:

$$\log P(G|D) = \sum_{i=1}^{p} z_i(G)$$

where

$$z_i(G) = \sum_{j=1}^{q_i(G)} \left(\log \frac{\Gamma(\alpha_{ij})}{\Gamma(n_{ij} + \alpha_{ij})} + \sum_{k=1}^{r_i} \log \frac{\Gamma(n_{ijk} + \alpha_{ijk})}{\Gamma(\alpha_{ijk})} \right) \tag{1}$$

Defining notation in (1): $q_i(G)$ is the number of joint instantiations of the parents of i in G; r_i is the number of values variable i can take; n_{ijk} is the count of how often in the data variable i takes its kth value when its parents in G take their jth joint instantiation; and α_{ijk} is the Dirichlet parameter corresponding to n_{ijk}. We also have $n_{ij} = \sum_{k=1}^{r_i} n_{ijk}$ and $\alpha_{ij} = \sum_{k=1}^{r_i} \alpha_{ijk}$. This equation (in non-log form) is given in [7] and was first derived by Cooper and Herskovits [2]. In this paper the parameters α_{ijk} will be set to $\alpha/(r_i q_i(G))$ where α (known as the *equivalent sample size*) is set by the user. This variant is known as the log BDeu score (Bayesian Dirichlet equivalent uniform score) where the "uniform" reflects that α_{ijk} is the same for all values k of i. With this restriction, and abbreviating $q_i(G)$ to q_i, (1) becomes:

$$z_i(G) = \sum_{j=1}^{q_i} \left(\log \frac{\Gamma(\frac{\alpha}{q_i})}{\Gamma(n_{ij} + \frac{\alpha}{q_i})} + \sum_{k=1}^{r_i} \log \frac{\Gamma(n_{ijk} + \frac{\alpha}{q_i r_i})}{\Gamma(\frac{\alpha}{q_i r_i})} \right) \tag{2}$$

$z_i(G)$ is the *BDeu log local score* for variable i and is determined by the parents i has in G (as well as by the data). Note that if $n_{ij} = 0$ for some j then also $n_{ijk} = 0$ for all k and the summand for j in (2) is zero. Let $q^{(0)}$ be the number of values of j where $n_{ij} = 0$ and $q^{(+)}$ be the number of values of j where $n_{ij} \geq 1$ (so that $q_i = q^{(0)} + q^{(+)}$). Suppose, without loss of generality, that $j = 1, \ldots q^{(+)}$ are the values of j where $n_{ij} \geq 1$, then we have that:

$$z_i(G) = \sum_{j=1}^{q^{(+)}} \left(\log \frac{\Gamma(\frac{\alpha}{q_i})}{\Gamma(n_{ij} + \frac{\alpha}{q_i})} + \sum_{k=1}^{r_i} \log \frac{\Gamma(n_{ijk} + \frac{\alpha}{q_i r_i})}{\Gamma(\frac{\alpha}{q_i r_i})} \right) \tag{3}$$

Since the log BDeu score exclusively is used here *BDeu log local score* will be abbreviated to *local score*. A more compact formulation of (3) is possible using the beta function of r variables. This function $B(x_1, \ldots, x_r)$ is defined in (4).

$$B(x_1, \ldots, x_r) = \frac{\Gamma(x_1) \ldots \Gamma(x_r)}{\Gamma(x_1 + \ldots + x_r)} \tag{4}$$

so that

$$\log B(x_1, \ldots, x_r) = \left(\sum_{k=1}^{r} \log \Gamma(x_k) \right) - \log \Gamma \left(\sum_{k=1}^{r} x_k \right)$$

Noting that

$$\sum_{k=1}^{r_i} \left(n_{ijk} + \frac{\alpha}{q_i r_i} \right) = n_{ij} + \frac{\alpha}{q_i}$$

we have

$$\begin{aligned} z_i(G) &= \sum_{j=1}^{q^{(+)}} \left(\log \frac{\Gamma(\frac{\alpha}{q_i})}{\Gamma(n_{ij} + \frac{\alpha}{q_i})} + \sum_{k=1}^{r_i} \log \frac{\Gamma(n_{ijk} + \frac{\alpha}{r_i q_i})}{\Gamma(\frac{\alpha}{q_i r_i})} \right) \\ &= \sum_{j=1}^{q^{(+)}} \left[\log \Gamma \left(\frac{\alpha}{q_i} \right) - \left(\sum_{k=1}^{r_i} \log \Gamma \left(\frac{\alpha}{q_i r_i} \right) \right) \right. \\ &\quad \left. + \left(\sum_{k=1}^{r_i} \log \Gamma \left(n_{ijk} + \frac{\alpha}{q_i r_i} \right) \right) - \log \Gamma \left(n_{ij} + \frac{\alpha}{q_i} \right) \right] \\ &= q^{(+)} \log \Gamma \left(\frac{\alpha}{q_i} \right) - q^{(+)} r_i \log \Gamma \left(\frac{\alpha}{q_i r_i} \right) \\ &\quad + \sum_{j=1}^{q^{(+)}} \log B \left(n_{ij1} + \frac{\alpha}{q_i r_i}, \ldots n_{ijr_i} + \frac{\alpha}{q_i r_i} \right) \end{aligned} \tag{5}$$

2 Computational benefits of bounding BDeu scores

An important problem is to identify a BN structure with maximal BDeu score for a given data set. Upper bounds on local scores can help in this task. To see this first note that the local score $z_i(G)$ depends only on the parents that i has in the graph G (that is what makes it 'local'), so it is clearer to write the local score as $z_i(W)$ where $W \subseteq \{1, \ldots p\} \setminus \{i\}$ are the parents of i. We have the following useful result:

[1] Dept. of Computer Science & York Centre for Complex Systems Analysis, University of York, York, UK email:james.cussens@york.ac.uk

Theorem 1 *If $W \subset W'$ and $z_i(W) > z_i(W')$ then W' cannot be a parent set for i in an optimal BN.*

This is because any BN where W' were the parents of i would have a worse score than one where the edges from $W' \setminus W$ to i were removed.

This simple result has been used by many working on BN learning [8, 3, 5, 4], but it has been exploited most fruitfully by de Campos and Ji [6]. Let $q_i(W')$ be the number of joint instantiations of variables in the set W', and let $q^{(+)}(W')$ be the number of associated non-zero counts. Translating the results of de Campos and Ji to the notation of the current paper they show that if (i) $W \subset W'$ and (ii) $\alpha/q_i(W') \leq 0.8349$ and (iii) $z_i(W) > -q^{(+)}(W') \log r_i$ then neither W' *nor and any superset of* W' can be an optimal parent set for i. (Their result is actually more general since they consider any BDe score, not just as here, the BDeu score.)

This is a tremendously useful result. Before computing the local score of a candidate parent set W' for i, de Campos and Ji inspect the scores of any previously computed $z_i(W)$ where $W \subset W'$ to see if conditions (ii)-(iii) are satisfied. If so, there is no need to compute $z_i(W')$. More importantly, all supersets of W' can be ruled out as optimal parent sets for i. Note that the required quantities for this check are either readily available or appropriately bounded: α is fixed by the user, $\log r_i$ requires a simple look-up, $q_i(W') \geq q_i(W)$ and $q^{(+)}(W') \geq q^{(+)}(W)$.

de Campos and Ji establish their result by considering upper bounds on BDe scores. The goals of the current paper are the same (restricted to BDeu). For an upper bound to be useful it important that (i) it allows all supersets of some W' to ruled out as parents set for i and (ii) it is defined in terms of cheaply computable quantities. The next section derives a promising upper bound for this.

3 Exploiting Alzer's bound

The key contribution of this paper is that we can obtain a useful upper bound on (BDeu log) local scores using an upper bound on the beta function discovered by Alzer [1]. Here is Alzer's result in its general form.

Theorem 2 *(From [1]). Let $c > 0$ be a real number and let $r \geq 2$ be an integer. Then we have for all real numbers $x_k \geq c\,(k = 1, \ldots, r)$:*

$$B(x_1, \ldots, x_r) \leq \beta_r(c) \frac{\prod_{k=1}^r x_k^{-1/2+x_k}}{\left(\sum_{k=1}^r x_k\right)^{-1/2+\sum_{k=1}^r x_k}}$$

with the best possible constant $\beta_r(c) = r^{rc-1/2} c^{(r-1)/2} \frac{(\Gamma(c))^r}{\Gamma(rc)}$.

It will be convenient to work with Alzer's bound in its log form:

$$\begin{aligned} &\log B(x_1, \ldots, x_r) \\ &\quad \leq \quad \log \beta_r(c) + \left(\sum_{k=1}^r \left(x_k - \frac{1}{2}\right) \log x_k\right) \\ &\quad - \quad \left(-\frac{1}{2} + \sum_{k=1}^r x_k\right) \log \left(\sum_{k=1}^r x_k\right) \end{aligned}$$

with the best possible constant $\log \beta_r(c) = (rc - 1/2) \log r + ((r-1)/2) \log c + r \log \Gamma(c) - \log \Gamma(rc)$.

By choosing $c = \frac{\alpha}{q_i r_i}$ (which is always positive since α must be) this theorem provides an upper bound for $\log B\left(n_{ij1} + \frac{\alpha}{q_i r_i}, \ldots n_{ijr_i} + \frac{\alpha}{q_i r_i}\right)$. With this choice of c we have:

$$\begin{aligned} \log \beta_{r_i}(c) \quad &= \quad (\frac{\alpha}{q_i} - 1/2) \log r_i + ((r_i - 1)/2) \log(\frac{\alpha}{q_i r_i}) \\ &\quad + r_i \log \Gamma(\frac{\alpha}{q_i r_i}) - \log \Gamma(\frac{\alpha}{q_i}) \end{aligned}$$

and so

$$\begin{aligned} &\log B\left(n_{ij1} + \frac{\alpha}{q_i r_i}, \ldots n_{ijr_i} + \frac{\alpha}{q_i r_i}\right) \\ &\quad \leq \quad (\frac{\alpha}{q_i} - 1/2) \log r_i + ((r_i - 1)/2) \log(\frac{\alpha}{q_i r_i}) \\ &\qquad + r_i \log \Gamma(\frac{\alpha}{q_i r_i}) - \log \Gamma(\frac{\alpha}{q_i}) \\ &\qquad + \left(\sum_{k=1}^{r_i} \left(n_{ijk} + \frac{\alpha}{q_i r_i} - \frac{1}{2}\right) \log(n_{ijk} + \frac{\alpha}{q_i r_i})\right) \\ &\qquad - \left(-\frac{1}{2} + \sum_{k=1}^{r_i} n_{ijk} + \frac{\alpha}{q_i r_i}\right) \log \left(\sum_{k=1}^{r_i} n_{ijk} + \frac{\alpha}{q_i r_i}\right) \\ &\quad = \quad (\frac{\alpha}{q_i} - 1/2) \log r_i + ((r_i - 1)/2) \log(\frac{\alpha}{q_i r_i}) + \\ &\qquad r_i \log \Gamma(\frac{\alpha}{q_i r_i}) - \log \Gamma(\frac{\alpha}{q_i}) \\ &\qquad + \left(\sum_{k=1}^{r_i} \left(n_{ijk} + \frac{\alpha}{q_i r_i} - \frac{1}{2}\right) \log(n_{ijk} + \frac{\alpha}{q_i r_i})\right) \\ &\qquad - \left(-\frac{1}{2} + n_{ij} + \frac{\alpha}{q_i}\right) \log \left(n_{ij} + \frac{\alpha}{q_i}\right) \end{aligned} \tag{6}$$

Plugging (6) into (5) and replacing G with W we get:

$$\begin{aligned} &z_i(W) \\ &\quad \leq \quad q^{(+)} \log \Gamma\left(\frac{\alpha}{q_i}\right) - r_i q^{(+)} \log \Gamma\left(\frac{\alpha}{q_i r_i}\right) \\ &\qquad + \sum_{j=1}^{q^{(+)}} \left[(\frac{\alpha}{q_i} - 1/2) \log r_i + ((r_i - 1)/2) \log(\frac{\alpha}{q_i r_i})\right. \\ &\qquad + r_i \log \Gamma(\frac{\alpha}{q_i r_i}) - \log \Gamma(\frac{\alpha}{q_i}) \\ &\qquad + \left(\sum_{k=1}^{r_i} \left(n_{ijk} + \frac{\alpha}{q_i r_i} - \frac{1}{2}\right) \log(n_{ijk} + \frac{\alpha}{q_i r_i})\right) \\ &\qquad \left. - \left(-\frac{1}{2} + n_{ij} + \frac{\alpha}{q_i}\right) \log \left(n_{ij} + \frac{\alpha}{q_i}\right)\right] \\ &\quad = \quad q^{(+)}(\alpha/q_i - 1/2) \log r_i + q^{(+)}(r_i/2 - 1/2) \log(\frac{\alpha}{q_i r_i}) \\ &\qquad + \sum_{j=1}^{q^{(+)}} \left[\left(\sum_{k=1}^{r_i} \left(n_{ijk} + \frac{\alpha}{q_i r_i} - \frac{1}{2}\right) \log(n_{ijk} + \frac{\alpha}{q_i r_i})\right)\right. \\ &\qquad \left. - \left(-\frac{1}{2} + n_{ij} + \frac{\alpha}{q_i}\right) \log \left(n_{ij} + \frac{\alpha}{q_i}\right)\right] \\ &\quad = \quad q^{(+)}(\alpha/q_i - 1/2) \log r_i + q^{(+)}(r_i/2 - 1/2) \log(\frac{\alpha}{q_i r_i}) \\ &\qquad + \sum_{j=1}^{q^{(+)}} \left[\left(\sum_{k=1}^{r_i} \left(n_{ijk} + \frac{\alpha}{q_i r_i}\right) \log(n_{ijk} + \frac{\alpha}{q_i r_i})\right)\right. \end{aligned} \tag{7}$$

$$
\begin{aligned}
&\quad -\left(n_{ij}+\frac{\alpha}{q_i}\right)\log\left(n_{ij}+\frac{\alpha}{q_i}\right)\Bigg] \\
&\quad -\frac{1}{2}\left[\sum_{j=1}^{q^{(+)}}\left(\sum_{k=1}^{r_i}\log(n_{ijk}+\frac{\alpha}{q_i r_i})\right)-\log\left(n_{ij}+\frac{\alpha}{q_i}\right)\right] \\
= \;& q^{(+)}(\alpha/q_i-1/2)\log r_i + q^{(+)}(r_i/2-1/2)\log(\frac{\alpha}{q_i r_i}) \\
&+\sum_{j=1}^{q^{(+)}}\sum_{k=1}^{r_i}\Bigg[\left(n_{ijk}+\frac{\alpha}{q_i r_i}\right)\log(n_{ijk}+\frac{\alpha}{q_i r_i}) \\
&-\left(n_{ijk}+\frac{\alpha}{q_i r_i}\right)\log\left(n_{ij}+\frac{\alpha}{q_i}\right)\Bigg] \\
&-\frac{1}{2}\left[\sum_{j=1}^{q^{(+)}}\left(\sum_{k=1}^{r_i}\log(n_{ijk}+\frac{\alpha}{q_i r_i})\right)-\log\left(n_{ij}+\frac{\alpha}{q_i}\right)\right] \\
= \;& q^{(+)}(\alpha/q_i-1/2)\log r_i + q^{(+)}(r_i/2-1/2)\log(\frac{\alpha}{q_i r_i}) \\
&+\sum_{j=1}^{q^{(+)}}\sum_{k=1}^{r_i}\left(n_{ijk}+\frac{\alpha}{q_i r_i}\right)\log\frac{n_{ijk}+\frac{\alpha}{q_i r_i}}{n_{ij}+\frac{\alpha}{q_i}} \\
&+\frac{1}{2}\left[\sum_{j=1}^{q^{(+)}}\log\left(n_{ij}+\frac{\alpha}{q_i}\right)-\sum_{k=1}^{r_i}\log\left(n_{ijk}+\frac{\alpha}{q_i r_i}\right)\right]
\end{aligned}
\tag{8}
$$

The right-hand side of (8) breaks down naturally into what is almost a prior component [$q^{(+)}(\alpha/q_i-1/2)\log r_i + q^{(+)}(r_i/2-1/2)\log(\frac{\alpha}{q_i r_i})$] and a strongly data-dependent component (the rest). Note that this 'almost-prior' component is not entirely prior to the data since it depends on the value of $q^{(+)}$, the number of positive values of n_{ij}.

To understand the strongly data-dependent component consider the following Bayesian approach to parameter estimation in the saturated model (where no conditional independence relations are assumed). Let $r=\prod_i r_i$ be the number of full joint instantiations of the variables, and associate a Dirichlet parameter α/r with each full joint instantiation. Imagine now updating this Dirichlet prior with the observed data. It is not difficult to see that the posterior mean distribution gives probability $\tilde{p}_\iota = (n_\iota+\alpha/r)/(N+\alpha)$ to each full joint instantiation ι where n_ι is the observed frequency of ι in the data, and N is the size of the data. Marginal probabilities are easy to compute. Let $\tilde{p}_{ijk}$ be the joint probability of variable i taking its kth value and its parents (in some fixed graph) taking their jth instantiation, then clearly $\tilde{p}_{ijk}=(n_{ijk}+\frac{\alpha}{q_i r_i})/(N+\alpha)$. Similarly, $\tilde{p}_{ij}=(n_{ij}+\frac{\alpha}{q_i})/(N+\alpha)$, where $\tilde{p}_{ij}$ is the probability that the parents take their jth instantiation.

From this it is not difficult to see that:

$$
\sum_{j=1}^{q_i}\sum_{k=1}^{r_i}\left(n_{ijk}+\frac{\alpha}{q_i r_i}\right)\log\frac{n_{ijk}+\frac{\alpha}{q_i r_i}}{n_{ij}+\frac{\alpha}{q_i}} = -(N+\alpha)H_{\tilde{p}}(i|W)
$$

where $H_{\tilde{p}}(i|W)$ is the conditional entropy of variable i given its parents W. (Note that there already exists valuable work linking BDeu scores to conditional entropy [9].) Since:

$$
\begin{aligned}
&\sum_{j=1}^{q^{(+)}}\sum_{k=1}^{r_i}\left(n_{ijk}+\frac{\alpha}{q_i r_i}\right)\log\frac{n_{ijk}+\frac{\alpha}{q_i r_i}}{n_{ij}+\frac{\alpha}{q_i}} \\
= \;& \sum_{j=1}^{q_i}\sum_{k=1}^{r_i}\left(n_{ijk}+\frac{\alpha}{q_i r_i}\right)\log\frac{n_{ijk}+\frac{\alpha}{q_i r_i}}{n_{ij}+\frac{\alpha}{q_i}} \\
&-\sum_{j=1}^{q^{(0)}}\sum_{k=1}^{r_i}\left(n_{ijk}+\frac{\alpha}{q_i r_i}\right)\log\frac{n_{ijk}+\frac{\alpha}{q_i r_i}}{n_{ij}+\frac{\alpha}{q_i}} \\
= \;& -(N+\alpha)H_{\tilde{p}}(i|W) \\
&-q^{(0)}\frac{\alpha}{q_i}\log(1/r_i)
\end{aligned}
\tag{9}
$$

and $-q^{(0)}\frac{\alpha}{q_i}\log(1/r_i)=\alpha\log r_i-\alpha(q^{(+)}/q_i)\log r_i$, plugging (9) into (8) and rearranging we have:

$$
\begin{aligned}
&z_i(W) \\
&\leq \;\alpha\log r_i + q^{(+)}[(r_i-1)\log(\alpha/q_i r_i)-\log(r_i)]/2 \\
&\quad -(N+\alpha)H_{\tilde{p}}(i|W) \\
&\quad +\frac{1}{2}\left[\sum_{j=1}^{q^{(+)}}\log\left(n_{ij}+\frac{\alpha}{q_i}\right)-\sum_{k=1}^{r_i}\log\left(n_{ijk}+\frac{\alpha}{q_i r_i}\right)\right]
\end{aligned}
\tag{10}
$$

It remains to obtain a easily computable upper bound on the term in square brackets in (10). We have:

$$
\sum_{j=1}^{q^{(+)}}\log\left(n_{ij}+\frac{\alpha}{q_i}\right) = q^{(+)}\log G(n_{ij}+\frac{\alpha}{q_i})_j
$$

where $G(n_{ij}+\frac{\alpha}{q_i})_j$ is the geometric mean of the set $\{n_{ij}+\frac{\alpha}{q_i} : j=1,..q^{(+)}\}$. Since the geometric mean is never greater than the arithmetic mean we have:

$$
\sum_{j=1}^{q^{(+)}}\log\left(n_{ij}+\frac{\alpha}{q_i}\right) \leq q^{(+)}\log(N/q^{(+)}+\alpha/q_i)
$$

Finally a lower bound on $\sum_{j=1}^{q^{(+)}}\sum_{k=1}^{r_i}\log\left(n_{ijk}+\frac{\alpha}{q_i r_i}\right)$ is needed. Suppose $s^{(+)}$ of the $q^{(+)}r_i$ terms are positive, then

$$
\begin{aligned}
&\sum_{j=1}^{q^{(+)}}\sum_{k=1}^{r_i}\log\left(n_{ijk}+\frac{\alpha}{q_i r_i}\right) \\
&\quad = \;(q^{(+)}r_i-s^{(+)})\log\frac{\alpha}{q_i r_i}+\sum_{jk:n_{ijk}\geq 1}\log\left(n_{ijk}+\frac{\alpha}{q_i r_i}\right)
\end{aligned}
$$

where $|\{jk : n_{ijk}\geq 1\}| = s^{(+)}$. The quantity $\sum_{jk:n_{ijk}\geq 1}\log\left(n_{ijk}+\frac{\alpha}{q_i r_i}\right)$ is minimised when the distribution of the n_{ijk} is as 'uneven' as possible, with one of these values having the value $N-s^{(+)}$ and all others with the value 1. So we have:

$$
\begin{aligned}
&\sum_{jk:n_{ijk}\geq 1}\log\left(n_{ijk}+\frac{\alpha}{q_i r_i}\right) \\
&\quad \geq \;(s^{(+)}-1)\log\left(1+\frac{\alpha}{q_i r_i}\right)+\log\left(N-s^{(+)}+\frac{\alpha}{q_i r_i}\right)
\end{aligned}
$$

and therefore:

$$
-\sum_{j=1}^{q^{(+)}}\sum_{k=1}^{r_i}\log\left(n_{ijk}+\frac{\alpha}{q_i r_i}\right)
$$

$$
\begin{aligned}
&\leq -(q^{(+)}r_i - s^{(+)}) \log \frac{\alpha}{q_i r_i} \\
&\quad -(s^{(+)} - 1) \log \left(1 + \frac{\alpha}{q_i r_i}\right) \\
&\quad - \log \left(N - s^{(+)} + \frac{\alpha}{q_i r_i}\right)
\end{aligned} \tag{11}
$$

Applying (11) gives the following upper bound on $z_i(W)$:

$$
\begin{aligned}
&z_i(W) \\
&\leq \alpha \log r_i + q^{(+)}[(r_i - 1)\log(\alpha/q_i r_i) - \log r_i]/2 \\
&\quad -(N + \alpha) H_{\tilde{p}}(i|W) \\
&\quad +\frac{1}{2} q^{(+)} \log(N/q^{(+)} + \alpha/q_i) \\
&\quad -\frac{1}{2}(q^{(+)}r_i - s^{(+)}) \log \frac{\alpha}{q_i r_i} \\
&\quad -\frac{1}{2}(s^{(+)} - 1) \log \left(1 + \frac{\alpha}{q_i r_i}\right) \\
&\quad -\frac{1}{2} \log \left(N - s^{(+)} + \frac{\alpha}{q_i r_i}\right)
\end{aligned}
$$

which can be rearranged to give:

$$
\begin{aligned}
&2z_i(W) \\
&\leq 2[\alpha \log r_i - (N + \alpha) H_{\tilde{p}}(i|W)] \\
&\quad + \left(s^{(+)} - q^{(+)}\right) \log \left(\frac{\alpha}{q_i r_i}\right) \\
&\quad +q^{(+)} \log \left(\frac{N}{q^{(+)} r_i} + \frac{\alpha}{q_i r_i}\right) \\
&\quad - \left(s^{(+)} - 1\right) \log \left(1 + \frac{\alpha}{q_i r_i}\right) \\
&\quad - \log \left(N - s^{(+)} + \frac{\alpha}{q_i r_i}\right)
\end{aligned} \tag{12}
$$

From (12) it is clear that $\frac{\alpha}{q_i r_i}$ is a key quantity. Since q_i grows exponentially with the number of parents, $\frac{\alpha}{q_i r_i}$ will be very small for large parent sets and in such cases $\log\left(\frac{\alpha}{q_i r_i}\right)$ will be a highly negative number. Consider now the term $\left(s^{(+)} - q^{(+)}\right) \log\left(\frac{\alpha}{q_i r_i}\right)$. We have that $s^{(+)} - q^{(+)} \geq 0$ since for each positive n_{ij} there must be at least one positive n_{ijk}. We have $s^{(+)} - q^{(+)} = 0$ only in the extreme case where, for each parent instantiation for which we have observed data ($n_{ij} > 0$), all the associated datapoints have the same value for the child i. In other words, in the observed data, the joint value of the parents *determines* that of the child. In this unusual case even a very negative value of $\log\left(\frac{\alpha}{q_i r_i}\right)$ does not drive the upper bound down. When the data rule out a deterministic relation between parents and child we have $s^{(+)} - q^{(+)} > 0$ and $\log\left(\frac{\alpha}{q_i r_i}\right)$ will push down the upper bound. In summary, the term $\left(s^{(+)} - q^{(+)}\right) \log\left(\frac{\alpha}{q_i r_i}\right)$ penalises parent sets according to how much the values of the parents fail to determine that of the child.

Further insight into this issue can be gained by rewriting (7) as:

$$
\begin{aligned}
&z_i(W) \\
&\leq q^{(+)}(\alpha/q_i - 1/2) \log r_i + q^{(+)}(r_i/2 - 1/2) \log \left(\frac{\alpha}{q_i r_i}\right) \\
&\quad + \sum_{j=1}^{q^{(+)}} \sum_{k=1}^{r_i} \left[\left(n_{ijk} + \frac{1}{r_i}\left(\frac{\alpha}{q_i} - \frac{1}{2}\right)\right) \log \left(\frac{n_{ijk} + \frac{\alpha}{q_i r_i}}{n_{ij} + \frac{\alpha}{q_i}}\right) \right. \\
&\quad \left. -\frac{1}{2}\left(1 - \frac{1}{r_i}\right) \log \left(n_{ijk} + \frac{\alpha}{q_i r_i}\right) \right]
\end{aligned} \tag{13}
$$

For each n_{ijk}, if $n_{ijk} > 0$ then $n_{ijk} \geq 1 > ((1/2) - (\alpha/q_i))/r_i$ and both terms of the summand for n_{ijk} are negative, driving down the upper bound. (So if all n_{ijk} are positive a simple upper bound on $z_i(W)$ is obtained by deleting the double summation in (13)). However for each n_{ijk} where $n_{ijk} = 0, n_{ij} > 0$ the second term is positive if $\alpha/q_i < r_i$ and the first term also if $\alpha/q_i < 1/2$. Both of these conditions will hold for typical choices of α ($\alpha = 1$, for example) and the upper bound will be pushed up. Each n_{ijk} for which $n_{ijk} = 0, n_{ij} > 0$ is an example of 'determinism in the data': the kth child value never occurs when the parents are in their jth configuration (and this jth configuration occurs at least once in the data). The parents are 'rewarded' each time this occurs by a positive boost to the upper bound on their score.

4 How tight are the bounds?

In this section local scores are compared against their upper bounds as computed by the direct application of Azler's bound (8) and also using the less data-dependent bound given in (12). A selection of such comparisons is presented here with the aim of exemplifying the main points.

In one experiment 100 datapoints were sampled from the well-known 'Asia' network and local scores for all possible parent sets were computed for the variable *Dyspnea* with α set to 1. Fig 1 shows the 128 local scores ordered by their value together with the upper bound computed by (8) and by (12), labelled `true`, `alzer` and `easy` respectively. Figs 2–8 shows results in the same form, for different numbers of datapoints, values of α and Bayesian network. See figure captions for details. The most strking finding is that both bounds become much tighter as the amount of data increases. This is as expected, since from (12) we can see that the entropy dominates as N increases. Such asymptotic behaviour has been analysed in some detail by Ueno [9].

5 Potential for pruning the search for local scores

A key motivation for obtaining bounds on BDeu local scores is to be able to (cheaply) prune the search for optimal parent sets. de Campos and Ji [6] have already achieved impressive pruning results.

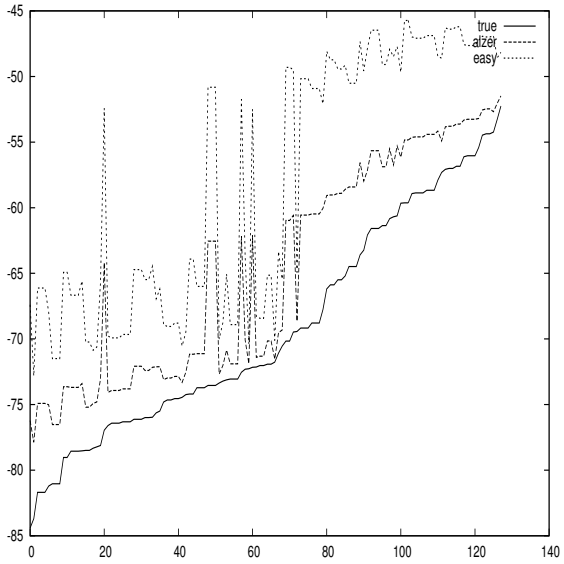

Figure 1. Local scores and upper bounds for parent sets of the variable *Dyspnea* in the 'Asia' network. Using 100 datapoints sampled from the 'Asia' network and $\alpha = 1$

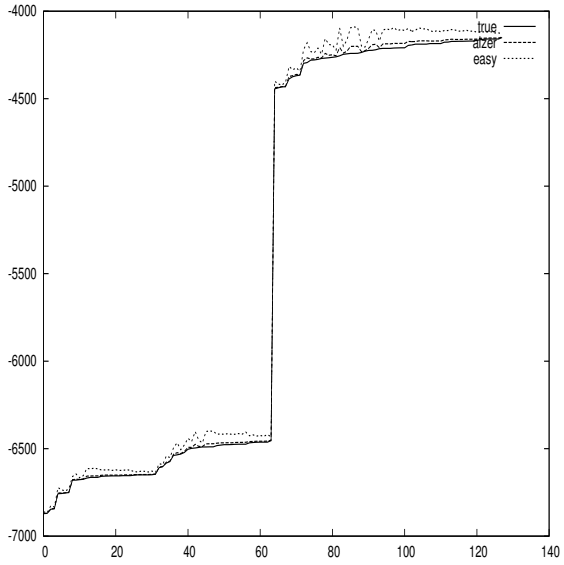

Figure 2. Local scores and upper bounds for parent sets of the variable *Dyspnea* in the 'Asia' network. Using 10000 datapoints sampled from the 'Asia' network and $\alpha = 1$

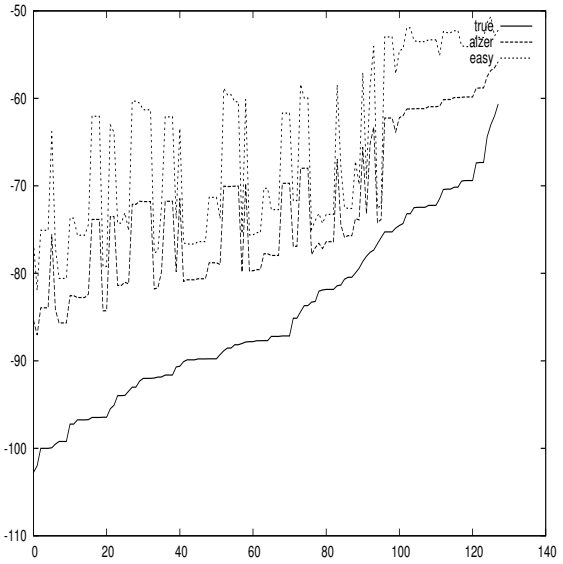

Figure 3. Local scores and upper bounds for parent sets of the variable *Dyspnea* in the 'Asia' network. Using 100 datapoints sampled from the 'Asia' network and $\alpha = 0.01$

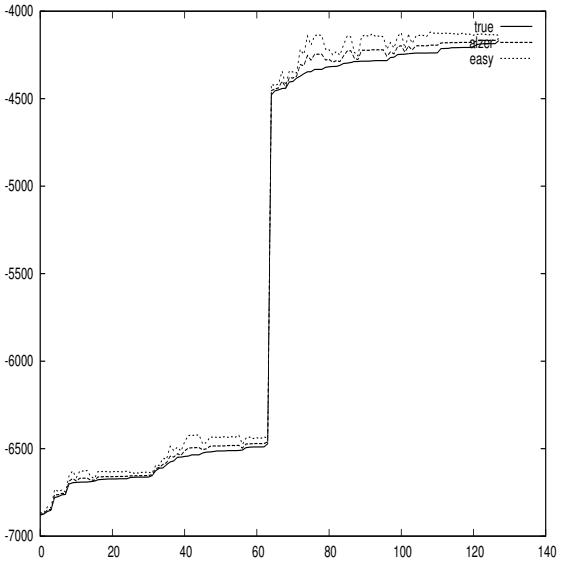

Figure 4. Local scores and upper bounds for parent sets of the variable *Dyspnea* in the 'Asia' network. Using 10000 datapoints sampled from the 'Asia' network and $\alpha = 0.01$

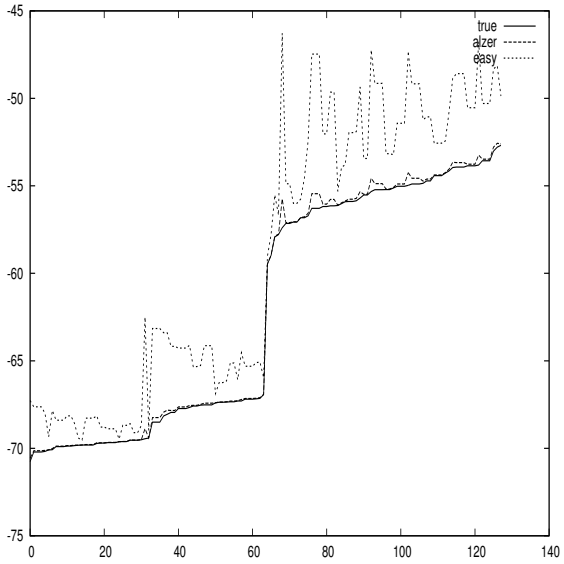

Figure 5. Local scores and upper bounds for parent sets of the variable *Dyspnea* in the 'Asia' network. Using 100 datapoints sampled from the 'Asia' network and $\alpha = 100$

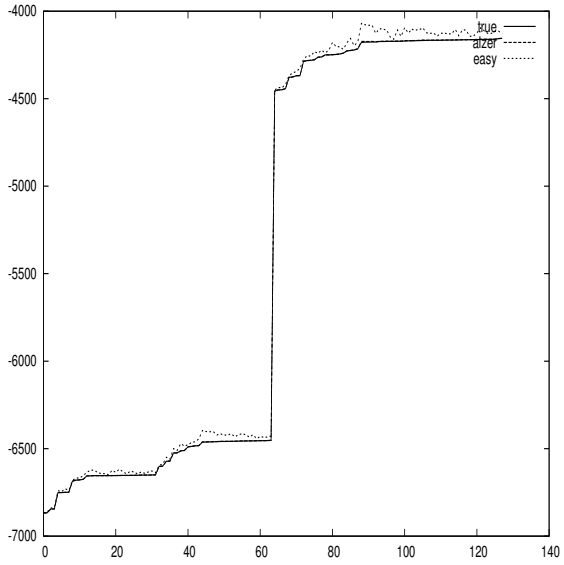

Figure 6. Local scores and upper bounds for parent sets of the variable *Dyspnea* in the 'Asia' network. Using 10000 datapoints sampled from the 'Asia' network and $\alpha = 100$

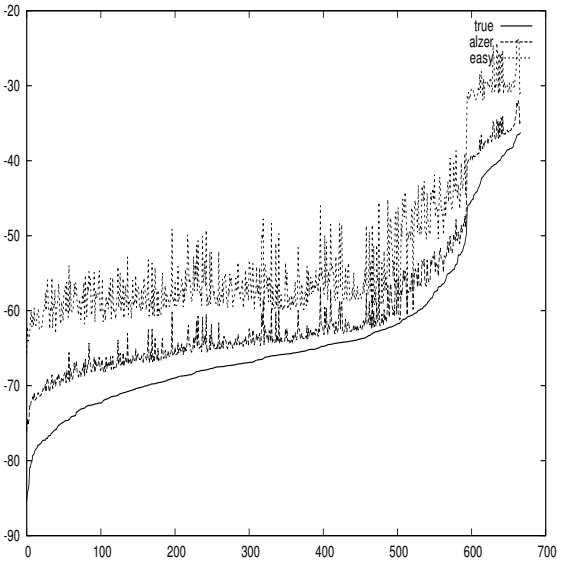

Figure 7. Local scores and upper bounds for small parent sets of the variable *HYPOVOLEMIA* in the 'alarm' network. Using 100 datapoints sampled from the 'alarm' network and $\alpha = 1$

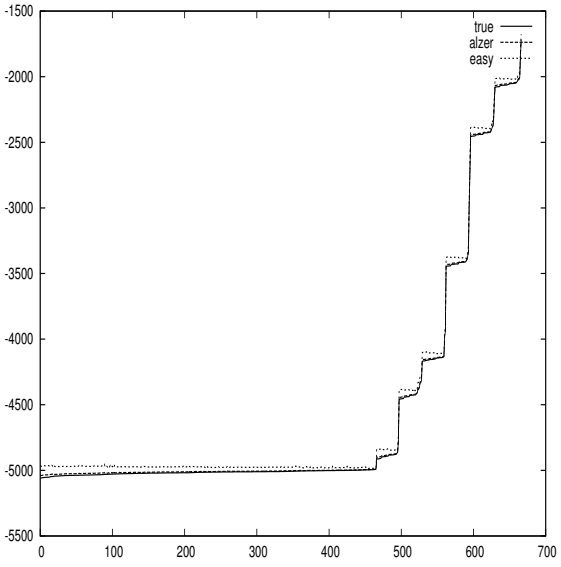

Figure 8. Local scores and upper bounds for small parent sets of the variable *HYPOVOLEMIA* in the 'alarm' network. Using 10000 datapoints sampled from the 'alarm' network and $\alpha = 1$

The bound (8), equivalent to (10), is tight but not appropriate for pruning. The looser bound (12) depends only on: $q^{(+)}(W), s^{(+)}(W), q_i$ and $H_{\tilde{p}}(i|W)$. These quantities can be bounded by the corresponding quantities for subsets and supersets. Let L and U be such that $L \subset W \subset U$. Abbreviate $(\min_{i' \in W \setminus L} r_{i'}) q_i(L)$ to $q'_i(L)$ and $(\min_{i' \in U \setminus W} r_{i'})^{-1} q_i(U)$ to $q'_i(U)$. We have

- $q'_i(L) \leq q_i(W) \leq q'_i(U)$
- $q^{(+)}(L) \leq q^{(+)}(W) \leq q^{(+)}(U)$
- $s^{(+)}(L) \leq s^{(+)}(W) \leq s^{(+)}(U)$
- $s^{(+)}(L) - q^{(+)}(L) \leq s^{(+)}(W) - q^{(+)}(W) \leq s^{(+)}(U) - q^{(+)}(U)$
- $-H_{\tilde{p}}(i|L) \leq -H_{\tilde{p}}(i|W) \leq -H_{\tilde{p}}(i|U)$

Note that the last double inequality is the well-known result that conditional entropy is always non-increasing as variables are added to the conditioning set. Conditional entropy only remains constant if the relevant conditional independence relation obtains.

If $\frac{\alpha}{q_i r_i} \leq 1$ there is the following bound:

$$
\begin{aligned}
& 2z_i(W) \\
& \quad \leq \quad 2[\alpha \log r_i - (N+\alpha) H_{\tilde{p}}(i|U)] \\
& \qquad + \left(s^{(+)}(L) - q^{(+)}(L)\right) \log \left(\frac{\alpha}{q'_i(L) r_i} \right) \\
& \qquad + q^{(+)}(U) \log \left(\frac{N}{q^{(+)}(L) r_i} + \frac{\alpha}{q'_i(L) r_i} \right) \\
& \qquad - \left(s^{(+)}(L) - 1\right) \log \left(1 + \frac{\alpha}{q'_i(U) r_i} \right) \\
& \qquad - \log \left(N - s^{(+)}(U) + \frac{\alpha}{q'_i(U) r_i} \right)
\end{aligned}
\tag{14}
$$

If the bound for $z_i(W)$ given by (14) is less than $z_i(L)$ it follows that W cannot be an optimal parent set for variable i. An obvious choice for U is $\{1, \ldots, p\} \setminus \{i\}$. The next step in this work is to see to what extent (14) provides effective pruning and compare to the existing method of de Campos and Ji [6].

ACKNOWLEDGEMENTS

I would like to thank the referees for their comments, which helped improve this paper considerably. This work has been supported by the UK Medical Research Council (Project Grant G1002312).

REFERENCES

[1] Horst Alzer. Inequalities for the Beta function of n variables. *The ANZIAM Journal*, 44:609–623, 2003.

[2] Gregory F. Cooper and Edward Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[3] James Cussens. Bayesian network learning by compiling to weighted MAX-SAT. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI 2008)*, pages 105–112, Helsinki, 2008. AUAI Press.

[4] James Cussens. Bayesian network learning with cutting planes. In Fabio G. Cozman and Avi Pfeffer, editors, *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 153–160, Barcelona, 2011. AUAI Press.

[5] C de Campos, Z Zeng, and Q Ji. Structure learning of Bayesian networks using constraints. In *Proc. of the 26th International Conference on Machine Learning*, 2009.

[6] Cassio P. de Campos and Qiang Ji. Properties of Bayesian Dirichlet scores to learn Bayesian network structures. In *Proc. AAAI-10*, pages 431–436, 2010.

[7] David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

[8] M. Teyssier and D. Koller. Ordering-based search: A simple and effective algorithm for learning Bayesian networks. In *Proc. UAI 2005*, pages 584–590, 2005.

[9] Maomi Ueno. Learning networks determined by the ratio of prior and data. In *Proc. UAI-2010*, pages 598–605, 2010.