SAMPLING WITHOUT REPLACEMENT IN JUNCTION TREES

by

R COWELL

Statistical Research Paper No. 15

Department of Actuarial Science and Statistics
City University
London
June 1997

ISBN 1 901615 14 6

Northampton Square
London EC1V 0HB
Telephone 0171-477 8000
“Any opinions expressed in this paper are my/our own and not necessarily those of my/our employer or anyone else I/we have discussed them with. You must not copy this paper or quote it without my/our permission.”
Sampling without replacement in junction trees

Robert Cowell

School of Mathematics, Actuarial Science and Statistics
City University
Northampton Square
London EC1E OHT
June 24, 1997

Abstract

It is shown that a highly efficient algorithm recently presented by Nilsson (1997) for finding the M-most probable configurations in a probabilistic expert system can be adapted for generating samples without replacement from a junction-tree representation of the joint distribution.

Keywords: Probabilistic expert system, Bayesian network, junction tree, max-propagation, sampling without replacement.
1 Introduction

For a probabilistic expert system, (also called a Bayesian network), of discrete variables Dawid (1992) showed how one could find the most-probable configuration of the variables by a single max-propagation on an associated junction tree. Recently Nilsson (1997) exploited this calculational method and combined it with a clever partitioning scheme to formulate an efficient method for finding the M most likely configurations of the joint distribution for any integer M. In his paper Dawid also showed how one can sample directly and efficiently from the posterior distribution represented by the marginal-potential representation on the junction tree. It is the purpose of this short paper to show that, using the partitioning scheme of Nilsson, one can sample without replacement directly and efficiently from the posterior distribution. The next section presents a summary of the key points of Nilsson's method. Sampling without replacement is developed by analogy in the subsequent section. We shall assume the reader is familiar with the papers of (Dawid 1992) and (Nilsson 1997), and we will borrow notations, definitions and results from these papers.

2 Finding the M-most likely configurations

Suppose we are given a connected\textsuperscript{3} junction tree $\mathcal{T}$ having a potential representation of a probability function $f$ in the discrete variables $X_u, u \in U$ given by

$$f = \frac{\prod_{C \in \mathcal{C}} a_C}{\prod_{S \in \mathcal{S}} b_S},$$

where the clique and separator potentials $a_C$ and $b_S$ are known non-negative functions. Then by a process called max-propagation on the junction tree, consisting of a set of local messages passed between the cliques, the joint distribution $f$ can be

\textsuperscript{3}This is not a problematic restriction, the disconnected case is readily treated.
expressed in the form

$$f = \frac{\prod_{C \in \mathcal{C}} f_C}{\prod_{S \in \mathcal{S}} f_S}$$

(2)

where $f_C$ is the maximum of $f$ over the subset of variables $x_C \subset U$. (See (Dawid 1992) for proof.) From this max-marginal representation one can readily find the configuration which has the highest probability, by the following search procedure (called max-distribute). First one designates a clique as the root, and orders the cliques starting with the root so that they obey the running intersection property. This is equivalent to making the junction tree directed, with all edges pointing away from the root, and then finding a topological ordering of the cliques beginning with the root. Such an ordering is always possible. For convenience label the ordered cliques sequentially thus $(C_0, C_1, \ldots, C_k)$, with separators $(S_1, \ldots, S_k)$ such that the separator $S_j$ joins the clique $C_j$ to the adjacent clique which is closer to the root. (see Figure 1 for an example). Denote the corresponding ordered residual sets by $(R_0, R_1, \ldots, R_k)$ where $R_j \equiv C_j \setminus S_j$ for every $j > 0$ and we define $R_0 \equiv C_0$.

To find the most likely configuration, one then proceeds outwards from the root as follows: First search the root clique $C_0$ for the configuration of its variables, $\hat{x}_{C_0}$ say, which maximizes the clique potential value (breaking ties arbitrarily). Distribute this as extra “evidence”, fixing successively the remaining variables in the cliques further from the root in their running intersection order by finding a maximal configuration consistent with the neighbouring cliques which have also been fixed, and including the states of the newly fixed variables as extra evidence. Thus for clique $C_j$ one finds the configuration of the residual variables $R_j$ which maximizes $f_C$, subject to the separator variables $S_j$ taking the configuration fixed by the earlier steps. The union of the “evidence” yields the most likely configuration. If there is “real” evidence then this will have been incorporated by the original max-propagation which put the
The algorithm of (Nilsson 1997) for finding the $M$-most likely configurations is an iterative divide-and-conquer process built upon this algorithm for finding the most likely configuration. Following Nilsson, let $H$ denote the complete set of configurations, and for $L = 1, 2$, etc let $x^L$ denote the $L$th most probable configuration. The procedure for finding the $L+1$th configuration, which is essentially a search over the space $H \setminus \{x^1, \ldots, x^L\}$, consists of three steps:

1. **Partition phase**: Partition $H \setminus \{x^1, \ldots, x^L\}$ into disjoint representative subsets, which together are consistent with all remaining configurations.

2. **Candidate phase**: Search these partition-generated subsets for the set consistent with the $L+1$th most probably configuration.

3. **Identification phase**: Extract the configuration from the selected candidate.

First we discuss a simplified (and far less efficient) version to help explain Nilsson’s algorithm. Denote the set of initial evidence (if any) by $\mathcal{E}$, and order the variables in some way, say $(X_1, X_2, \ldots, X_n)$ where the $j$-th node has $T_j$ states $(x_{j}^{1}, \ldots, x_{j}^{T_j})$. 
From the preceding discussion, one can find the most likely configuration, $x^1 = (x_1^1, \ldots, x_n^1)$. Now the second most likely configuration, $x^2$ must differ from $x^1$ in the state of at least one variable. So now perform a further $n$ max-propagations, (each requires a re-initializing the junction tree), each with a distinct "pseudo-evidence" $\mathcal{E}_{j,8}^i$ defined by:

$$
\begin{align*}
\mathcal{E}_{1,8}^1 &= \mathcal{E} \text{ and } X_1 \neq x_1^1, \\
\mathcal{E}_{2,8}^1 &= \mathcal{E} \text{ and } X_1 = x_1^1 \text{ and } X_2 \neq x_2^1, \\
\mathcal{E}_{3,8}^1 &= \mathcal{E} \text{ and } X_1 = x_1^1 \text{ and } X_2 = x_2^1 \text{ and } X_3 \neq x_3^1, \\
\vdots \\
\mathcal{E}_{n,8}^1 &= \mathcal{E} \text{ and } X_1 = x_1^1 \text{ and } \ldots \text{ and } X_{n-1} = x_{n-1}^1 \text{ and } X_n \neq x_n^1.
\end{align*}
$$

(3)

By this procedure we partition the search space for the set of all configurations, excluding the most likely one $x^1$ we have already found. Hence one of these must be consistent with the second most likely configuration $x^2$, and only one if it is unique; which one can be found by calculating the max-normalizations of each and picking the one with the largest normalization.

If the second most likely configuration is found to arise from the $j$th set, then by a max-distribute operation it can found using initial evidence $\mathcal{E}_{j,8}^i$. Denote it by $x^2 = (x_1^2, \ldots, x_n^2)$, (where by the partition constraint $t_i^1 = t_i^2$ for all $i < j$).

Now the third most likely configuration $x^3$ is either consistent with one of the set alternative evidences of (3) excluding the $j$th, or it has a configuration which disagrees with both the $x^1$ and $x^2$ configurations in at least one place for some variable having index $\geq j$ while agreeing for all indices less that $j$. To find out which perform a further partition of the search space to obtain an extra set of $n - j + 1$
disjoint evidences of the form:

\[ E_{j,j}^2 = E_{j,j}^1 \text{ and } X_j \neq x_j^q, \]
\[ E_{j+1,j}^2 = E_{j,j}^1 \text{ and } X_j = x_j^q \text{ and } X_{j+1} \neq x_{j+1}^q, \]
\[ \vdots \]
\[ E_{n-1,n}^2 = E_{j,j}^1 \text{ and } X_{n-1} = x_{n-1}^q \text{ and } X_n \neq x_n^q. \]

(4)

Finding the max-normalizations of these, and comparing with those of (3) (excluding the jth) we pick out that which attains the highest max-normalization and the use max-distribute to find the actual configuration. The representative evidence is then further partitioned, etc, until the desired number \( M \) of configurations have been found.

The partitioning the search space is one of the key steps in Nilsson’s method, and will work with any ordering of the variables. However it does require a large number of junction tree initializations and propagations. The other key step is to select a specific ordering of the variables so that only one max-propagation is required, thus making the process much more efficient. Specifically, one orders the variables according to the ordered residual sets \( R \), and in the partitioning phase excludes joint states of the residual variables (rather than states of individual variables). Thus, in terms of the residual sets, suppose the most likely configuration has the form \( x^1 = (r_1^q, \ldots, r_k^q) \), (where now the ts range over the configurations of the residuals).

Then as dummy evidence for finding the second most likely configuration, one could perform \( k + 1 \) propagations of the form:
\[ \mathcal{E}_{i,0} = \mathcal{E} \text{ and } R_0 \neq x_{i}^{0}, \]
\[ \mathcal{E}_{1,0} = \mathcal{E} \text{ and } R_0 = x_{i}^{0} \text{ and } R_1 \neq x_{i}^{1}, \]
\[ \mathcal{E}_{2,0} = \mathcal{E} \text{ and } R_0 = x_{i}^{0} \text{ and } R_1 = x_{i}^{1} \text{ and } R_2 \neq x_{i}^{2}, \]
\[ \vdots \]
\[ \mathcal{E}_{k,0} = \mathcal{E} \text{ and } R_0 = x_{i}^{0} \text{ and } \ldots \text{ and } R_{k-1} = x_{i}^{k-1} \text{ and } R_k \neq x_{i}^{k}. \]

The point about using this ordering is that, for example, if one were to do a max-collect using \( \mathcal{E}_{j,0} \) then, because no new dummy-evidence has been entered beyond the \( j \)th residual set or clique, all messages received by the subtree induced by \( (C_0, \ldots, C_j) \) are identical to those which were sent in the original max-propagation using the evidence \( \mathcal{E} \). This means that sending such messages is unnecessary, and one need only consider the max-collect operation on the subtree \( (C_0, \ldots, C_j) \) to find the max-normalization. But such an operation is trivial in that, because of the evidence \( \mathcal{E}_{j,0} \) entered, only one non-zero value from each clique and separator (denoted by \( x_{C_i}^{*} \) and \( x_{S_i}^{*} \)) survive for all indices \( i < j \), thus the normalization can be found very efficiently by the simple formula:

\[ \frac{\prod_{i=0}^{j-1} f_C(x_{C_i}^{*})}{\prod_{i=0}^{k-1} f_S(x_{S_i}^{*})} \max f_C(x_{C_j}, \mathcal{E}_{j,0}) \]

which requires a simple search in the \( j \)th clique for the last term. This is Theorem 3 of (Nilsson 1997). \( \mathcal{E}_{j,0} \) is then partitioned similarly to (4) into:

\[ \mathcal{E}_{2,j} = \mathcal{E}_{j,0} \text{ and } R_j \neq x_{j}^{0}, \]
\[ \mathcal{E}_{2,j+1} = \mathcal{E}_{j,0} \text{ and } R_j = x_{j}^{0} \text{ and } R_{j+1} \neq x_{j+1}^{0}, \]
\[ \vdots \]
\[ \mathcal{E}_{n,j} = \mathcal{E}_{j,0} \text{ and } R_{n-1} = x_{n-1}^{0} \text{ and } R_k \neq x_{k}^{0}. \]
3 Sampling without replacement algorithm

Besides showing how one can find the most likely configuration, in his paper Dawid (1992) also showed how one can sample with replacement directly from the posterior distribution using the junction tree. The first stage is to put the junction tree potentials into their marginal-potential form by a sum-propagation of evidence, thus:

$$f(U,E) = \frac{\prod_C f(C,E)}{\prod_S f(S,E)}$$  \hspace{1cm} (8)

Then to get a sample, one uses again the ordering of the cliques as for the previous maximization story. One begins by sampling a configuration in the root clique, drawing from $f(R_0, E)$ to obtain $r_0^{(1)}$ say. Next one samples from $f(R_1 \mid R_0 = r_0^{(1)}, E)$ found from clique $C_1$, etc, processing the cliques in their running intersection ordering, so that when sampling for $R_i$ from clique $C_i$ the variables $X_S$ will already have been fixed by earlier sampling. When all cliques have been so processed a configuration will have been generated with its correct probability. This process can be repeated to draw as many independent samples as may be required. We shall call such a search process a sample-distribute operation.

However, instead of sampling independent samples we can employ Nilson’s partitioning scheme to draw samples without replacement correctly from the posterior distribution. The differences are (1) sum-propagation rather than max-propagation is used to initialize the tree, and (2) sampling rather than maximization is performed in the candidate partition search stages.

Specifically, suppose that we have put the junction tree into sum-marginal representation by sum-propagation of evidence. Then one samples one configuration by the method described above - denote this first sample by $x^1$ writing it in terms of the residual sets as: $x^1 = (r_0^{(1)}, \ldots, r_k^{(1)})$. Then to sample a second configuration, in analogy to finding the second most likely configuration one generates $k + 1$ candi-
dates using evidences given by (5), with the evidence defined by $E_{j,b}^1$ being assigned a weight given by

$$\frac{\prod_{i=0}^{l-1} f_{C_i}(x_{C_i})}{\prod_{j} f_{S_j}(x_{S_j})} \sum f_{C_j}(x_{C_j}, E_{j,b}^1)$$

(cf (6)). This weight could alternatively be found much more inefficiently using a sum-collect operation with evidence $E_{j,b}^1$, but again in direct analogy to the max-collect operation, by sum-consistence the cliques and separators having indices greater than $j$ have their contributions cancel.

Next from this set of weighted candidates one samples a candidate according to its weight, (max-distribute-search picks the highest weight configuration), and then proceed to complete the configuration by a sample-distribute operation, which ensures that the final configuration obtained is drawn from the distribution correctly.

Thus assume that in this process the $j$th term was sampled. Then in analogy to drawing a single sample, the configuration is completed by a sample-distribute-search using $E_{j,b}^1$, (which can begin from the $j$ clique as all previous cliques have completely specified residuals). Its probability can then be found easily. After completion, if a third configuration is required then this must also be partitioned to form new extra candidates. These candidates will have respective evidences as in (7), with weights given in analogy to (9).

4 Conclusions

We have shown informally how samples may be drawn without replacement using a variation of the top $M$-configuration search algorithm of Nilsson (1997), in which max-propagation is replaced with sum-propagation, and max-distribute-searching with sample-distribute-searching etc. With these simple replacements, plus sampling
DEPARTMENT OF ACTUARIAL SCIENCE AND STATISTICS

Actuarial Research Papers

1. Haberman S. Long Term Prognosis after Cerebral Infarction. 78 pages, 1984 (out of print).
   ISBN 1 874 770 01 8

2. Haberman S. Long Term Prognosis after Subarachnoid Haemorrhage. 43 pages, 1984 (out of
   print).
   ISBN 1 874 770 02 6

3. Haberman S. Age Variation in Incidence of Cerebrovascular Disease. 36 pages, 1985 (out of
   print).
   ISBN 1 874 770 03 4

   ISBN 1 874 770 04 2

   [Also: Haberman S. and Bloomfield D.S.F. Social Class Differences in Mortality Around 1981.

   ISBN 1 874 770 05 0

6. Kaye G.D. Life Assurance Taxation - Comments in Response to the Board of the Inland
   ISBN 1 874 770 06 9

   ISBN 1 874 770 07 7

   circulation).
   ISBN 1 874 770 08 5

   time Lost to Sickness, Unemployment and Stoppages. 100 pages. 1989.
   ISBN 1 874 770 09 3

    ISBN 1 874 770 10 7


30. Puzey, A.S. An Investigation into the Effect of Averaging Age Nearest Birthday at the Beginning of a Rate Year to Exact Age. October 1991. 33 pages. ISBN 1 874 770 30 1


ISBN 1 874 770 41 7

ISBN 1 874 770 42 5

ISBN 1 874 770 43 3

ISBN 1 874 770 44 1

ISBN 1 874 770 45 X

ISBN 1 874 770 46 8

ISBN 1 874 770 47 6

ISBN 1 874 770 48 4

ISBN 1 874 770 49 2

ISBN 1 874 770 50 6

ISBN 1 874 770 51 4

ISBN 1 874 770 52 2

ISBN 1 874 770 53 0

ISBN 1 874 770 54 9


86. Booth P.M.  Long-Term Care for the Elderly: A Review of Policy Options. April 1996. 45 Pages. ISBN 1 874 770 89 1


95. Chadburn R.G.  The Use of Capital, Bonus Policy and Investment Policy in the Control of Solvency for With-Profits Life Insurance Companies in the UK. April 1997. 29 Pages. ISBN 1 901615 05 7

ISBN 1 901615 07 3

ISBN 1 901615 08 1

ISBN 1 901615 09 X

ISBN 1 901615 12 X

Statistical Research Papers

ISBN 1 874 770 83 2

ISBN 1 874 770 86 7

ISBN 1 874 770 87 5

ISBN 1 874 770 92 1

ISBN 1 874 770 95 6

ISBN 1 874 770 96 4

ISBN 1 874 770 97 2

ISBN 1 874 770 98 0

ISBN 1 901615 00 6


ISBN 1 901615 14 6