Sparse Factorization Methods for Inference in Bayesian Networks

Ernesto Coutinho Colla*† and Julio Michael Stern*†

*Institute of Mathematics and Statistics, University of São Paulo.
†ecolla@ime.usp.br jstern@ime.usp.br

INTRODUCTION

Bayesian Networks (BNs) are probabilistic graphical models used to represent and encode uncertain expert knowledge. BNs stand out for dealing with uncertainty in decision making and statistical inference, and many algorithms were described for inference in BNs, see Dechter (1996), Heckerman (1995), Jensen (1996), Lauritzen (1988), Pearl (1988), and Zang (1996). The parallel algorithm described in this paper is based on the sequential variable elimination algorithm of Cozman(2000), using algebraic operations on potentials. These algebraic schemata for inference in BNs are not only relatively simple to understand and to implement, but also allow us to use the techniques, heuristics and abstract combinatorial structures from the sparse matrix factorizations literature, see George (1993) and Stern (1994, 2006, 2008).

The main goal of this paper is to show how variations of the variable elimination algorithm can be combined with sparse matrix factorization methods to implement a fast and efficient parallel algorithm for inference in BNs. This goal is achieved with the complete separation between a first symbolic phase, and a second numerical phase. In the symbolic phase the proposed algorithm explores the graphical structure of the model, without computing or even accessing probabilistic information. The second numerical phase can be fully vectorized and parallelized using static data structures previously defined in the first phase. This is done examining the decoupling or separation operators of sparse matrix factorization algorithms and BNs inference procedures from a unified combinatorial framework. This unified framework is the key for implementing efficiently this parallel algorithm.

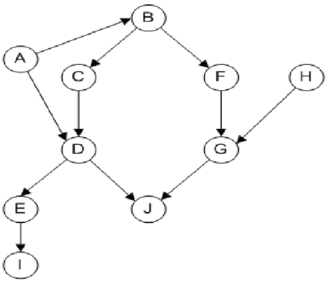
INFERENCE WITH BAYESIAN NETWORK

A BN, see Jensen (1996), is a graphical model that efficiently encodes the joint probability distribution for a set (or list) of random variables, $X = \{X_1, X_2, ..., X_n\}$, each of them having a finite number of possible states. A BN consists of two components: (i) A Directed Acyclic Graph (DAG) defining the network structure and encoding the conditional dependence relations between the variables in X; (ii) A set of local probability densities associated with each variable. Each node, i, of the DAG represents a random

CP1073, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering—28th International Workshop*
edited by M. de Souza Lauretto, C. A. de Bragança Pereira, and J. M. Stern

variable, X_i. In order to make the notation lighter, we may write a node index, i, instead of its random variable, X_i, and vice versa. We also use the vectorized notation, X_S, for the subset $\{X_i\}, i \in S$.

The DAG representing the BN structure has an arc from node i to node j, that is, i is a *parent* of j, $i \in \mathrm{pa}(j)$, if the probability distribution of variable X_j is directly dependent on variable X_i, and the strength of this influence is expressed by conditional probability distributions. In many specific statistical models an arc can be interpreted as a direct influence or causal effect of X_i on X_j, see Pearl (1988).



Probability Densities:
$P(A)$, $P(B \mid A)$, $P(C \mid B)$, $P(D \mid A,C)$, $P(E \mid D)$
$P(F \mid B)$, $P(G \mid F,H)$, $P(H)$, $P(I \mid E)$, $P(J \mid D,G)$

B and D are children of A
C is child of B
A and C are parents of D so A and C are spouses
D and G are parents of J so D and G are spouses
G and J are descendants of F
D, E, I and J are descendants of C
H is nondescendant of B

Figure 1. Bayesian network example.

The semantics of BNs implies a correspondence between the topology of a DAG and the network's probabilistic dependence relations, determined by the *Markov condition*: Every variable is independent of its nondescendants nonparents given its parents. Therefore, every X_i is associated with a local probability density, $P(X_i \mid X_{\mathrm{pa}(i)})$, as showed in Fig 1. Based on this condition, a BN encodes a unique probability distribution: $P(X) = \prod_i P(X_i \mid X_{\mathrm{pa}(i)})$.

Inference in BNs is based on queries, where the posterior marginal distribution for a set of *query variables*, X_Q, has to be computed given a set of observed variables, X_E. This set of observed variables is the *evidence* in the network and establishes the values of the variables in X_E. For example: $e = \{X_i = x_i, X_j = x_j\}$ establishes the values of X_i and X_j, so $E = \{i, j\}$.

The posterior probability of X_Q given e is:

$$P(X_Q \mid e) = \frac{P(X_Q, e)}{P(e)} = \frac{\sum_{X \setminus \{X_Q, X_E\}} P(X)}{\sum_{X \setminus X_E} P(X)} \,. \tag{1}$$

The expression $X \setminus Y$ indicates the set of all variables which belong to X but do not belong to Y, and the expression $\sum_X f(X,Y)$ indicates that all variables of X were *eliminated* or *marginalized out*, that is, were summed out from the function $f(X,Y)$.

Efficient computational algorithms rely on two important technical points:

(I) Given a BN over variables X, an evidence e and a query X_Q, not all variables of X may be required to compute $P(X_Q \mid e)$. If the local probability density $P(X_i \mid X_{\mathrm{pa}(i)})$ is required to compute $P(X_Q \mid e)$, then X_i is a *requisite variable*, $i \in R$. Fortunately there are simple polynomial algorithms able to identify the set R. We have used Bayes-Ball algorithm, see Shachter (1998). It is important to realize that the requisite variables, X_R, can be identified exploring only the DAG topology, without any numerical information concerning probability distributions. Hence, in order to reduce the problem dimension, the identification of R should be done at the very first stage of inference calculation.

(II) At intermediate computations, it is not necessary to compute the normalization constants, that is, the denominator $P(e)$ of (1). We only need the numerator in (1),

$$P(X_Q \mid e) \propto P(X_Q, e) = \sum\nolimits_{X_R \setminus \{X_Q, X_E\}} \left(\prod\nolimits_{X_i \in X_R} P(X_i \mid X_{\mathrm{pa}(i)}) \right) . \tag{2}$$

Hence, a basic rule for operation in BNs is: Compute the numerator $P(X_Q, e)$ and obtain normalization constant $P(e)$ only in the last stage. This rule means that we can perform the intermediate computations with un-normalized distributions, which are real-valued tables over a finite set of variables. These tables, ϕ, are called *potentials*, see Jensen (1996). A potential's domain, $\mathrm{dom}(\phi)$, is its correspondent set of variables. In the following, we give some important properties of the algebra of potentials:

(1) A variable X_i can be *marginalized out* of a potential ϕ resulting in a new potential $\phi'_{X_i} = \sum_{X_i} \phi$ over the domain $\mathrm{dom}(\phi'_{X_i}) = \mathrm{dom}(\phi) \setminus \{X_i\}$. Marginalization follows:
(1a) the commutative law: $\sum_{X_i} \sum_{X_j} \phi = \sum_{X_j} \sum_{X_i} \phi$; and
(1b) the distributive law: if $X_i \notin \mathrm{dom}(\phi_1)$, then $\sum_{X_i} \phi_1.\phi_2 = \phi_1.\sum_{X_i} \phi_2$.

(2) Two potentials can be *multiplied*, resulting in a new potential with
$\mathrm{dom}(\phi_1.\phi_2) = \mathrm{dom}(\phi_1) \cup \mathrm{dom}(\phi_2)$. Multiplication follows:
(2a) the commutative law: $\phi_1.\phi_2 = \phi_2.\phi_1$; and
(2b) the associative law: $(\phi_1.\phi_2).\phi_3 = \phi_1.(\phi_1.\phi_3)$.

As an example, consider the BN in Figure 1. The BN joint probability distribution can be rewritten as: $P(X) \propto \phi_A.\phi_B.\phi_C.\phi_D.\phi_E.\phi_F.\phi_G.\phi_H.\phi_I.\phi_J$, and the potentials specified for the network are: $P(A) \propto \phi_A(A), P(B \mid A) \propto \phi_B(B,A), P(C \mid B) \propto \phi_C(C,B), P(D \mid A,C) \propto \phi_D(D,A,C)$ and so on. Computing $P(I)$ can be accomplished by marginalizing out of $P(X)$ all the variables, except I.

$$P(I) = \sum\nolimits_{A,B,C,D,E,F,G,H,J} P(X) . \tag{3}$$

BNs are particularly useful for calculating new probabilities when we acquire new information. However, in the preceding calculations no evidence was entered into the network. Now, assume information e has been acquired, stating that "$A = a_t$", where A is a variable and a_t is the t-th state of A. Let A have s states with probability distribution $P(A) = (x_1, \ldots, x_t, \ldots, x_s)$. This observed evidence e means that all states except tth one are impossible. So the new (un-normalized) probability distribution is $P(A,e) = (0, \ldots., 0, x_t, 0, \ldots, 0)$ which is the result of multiplying $P(A)$ with $\underline{e}_A = (0, \ldots, 0, 1, 0, \ldots, 0)$ in which only tth value is 1. The s-dimensional 0-1 potential $\underline{e}_A$ is called *finding*.

In the current example, assume that we have the evidence $A = a$, $H = h$ and $J = j$. This evidence e would be represented using three findings $\underline{e}_A$, $\underline{e}_H$ and $\underline{e}_J$. The posterior marginal $P(I \mid e)$ can be obtained normalizing $P(I, E)$:

$$P(I, e) = \sum\nolimits_{A,B,C,D,E,F,G,H,J} P(X).\underline{e}_A.\underline{e}_H.\underline{e}_J . \tag{4}$$

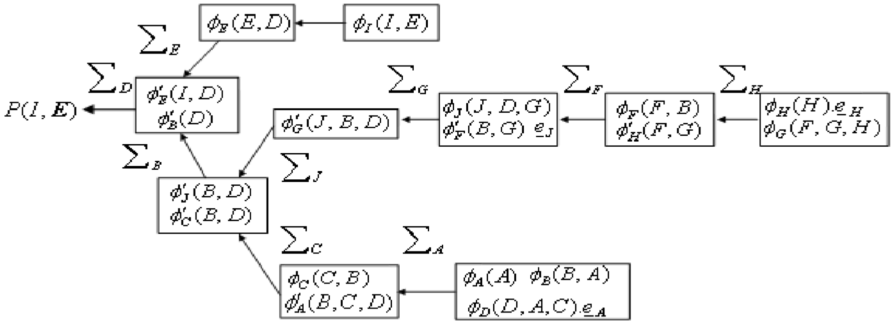
To avoid calculating the product of all potentials, we use the distributive law:

$$P(I\,|\,e) = \sum_D \sum_E \phi_I(I,E).\phi_E(E,D) \sum_B \sum_C \phi_C(C,B). \sum_A \phi_A(A).\phi_B(B,A).\phi_D(D,A,C).\underline{e}_A$$

$$\sum_J \sum_G \phi_J(J,D,G).\underline{e}_J \sum_F \phi_F(F,B). \sum_H \phi_G(G,F,H).\phi_H(H).\underline{e}_H\ .$$

First, calculate $\phi'_H = \sum_H \phi_G(G,F,H).\phi_H(H).\underline{e}_H$, then multiply $\phi'_H(F,G)$ on $\phi_F(F,B)$ and calculate $\phi'_F = \sum_F \phi_F(F,B).\phi'_H(F,G)$. The later result is multiply on $\phi_J(J,D,G).\underline{e}_J$, to calculate $\phi'_G = \sum_G \phi_J(J,D,G).\phi'_F(B,G).\underline{e}_J$, and so forth. All the operations involved are represented in Figure 2.

Because marginalization is commutative it can be done in any order. In the preceding calculation the marginalization, also called variable elimination, was done in a particular order, namely $q = [H,F,G,E,A,C,J,B,D]$.



$\phi_E(E,D)$
$\phi_I(I,E)$
$\sum_E$
$\sum_D$
$P(I,E)$
$\phi'_E(I,D)$ $\phi'_B(D)$
$\sum_G$
$\sum_F$
$\sum_H$
$\phi'_G(J,B,D)$
$\phi_J(J,D,G)$ $\phi'_F(B,G)$ $\underline{e}_J$
$\phi_F(F,B)$ $\phi'_H(F,G)$
$\phi_H(H).\underline{e}_H$ $\phi_G(F,G,H)$
$\sum_B$
$\sum_J$
$\phi'_J(B,D)$ $\phi'_C(B,D)$
$\sum_C$
$\sum_A$
$\phi_C(C,B)$ $\phi'_A(B,C,D)$
$\phi_A(A)$ $\phi_B(B,A)$ $\phi_D(D,A,C).\underline{e}_A$

Figure 2. The process of marginalizing down to *I*

The diagram in Figure 2 also portrays the dependencies among potential operations to calculate $P(I\,|\,e)$. Notice that some operations could be done simultaneously. For example, at very first stage, we could perform the required operations on A, H, and I, calculating $\phi'_A = \sum_A \phi_A(A).\phi_B(B,A).\phi_D(D,A,C).\underline{e}_A$, etc.

If a parallel computer is available, we can simultaneously execute all the marginalizations using already computed potentials. Hence, it is desirable to find:
(i) An efficient way to specify all dependencies among these marginalization operations.
(ii) A way to specify an elimination order entailing a “simple” dependence structure, so that many operations can be done simultaneously.

The dependence structure of these operations is exactly the same as the dependence structure for “pivoting” operations appearing in numerical linear algebra, namely, in the Cholesky factorization of sparse matrices, see George (1993), Pissanetzky (1984), and Stern (1994, 2006, 2008). We describe only the aspects pertinent to this paper.

An *Undirected Graph* (UG), $\mathscr{G} = (\mathscr{V},\mathscr{E})$, has undirected edges, $\{i,j\} \in \mathscr{E}$, standing for pairs of opposite directed arcs, (i,j) and (j,i). The *Moral Graph* of a DAG, $\mathscr{G}$, is the UG with the same nodes as $\mathscr{G}$, and edges joining nodes i and j if they are immediate relatives in $\mathscr{G}$. The immediate relatives of a node in $\mathscr{G}$ include its parents, children and spouses (but not brothers or sisters). i is a spouse of j is they have a child in common, that is, $i \in \mathrm{sp}(j) \Leftrightarrow \exists k \,|\, i,j \in \mathrm{pa}(k)$.

The *Markov Blanket* of X_i, $X_{\mathrm{mb}(i)}$ is defined as the minimal set of variables that makes a variable X_i independent from all other variables in the BN. This means that the Markov Blanket of a variable "decouples" this variable from the rest of network: $P(X_i \,|\, X_{\mathrm{mb}(i)}, X_j) = P(X_i \,|\, X_{\mathrm{mb}(i)})$. It can be shown that the set of immediate relatives of node i is the Markov Blanket of node i. Figure 3a shows the Moral Graph of the BN in Figure 1. It is important to realize that if X_i and X_j are both in the same domain, of a variable X_k of the BN, then the edge $\{i,j\}$ is in the Moral Graph.

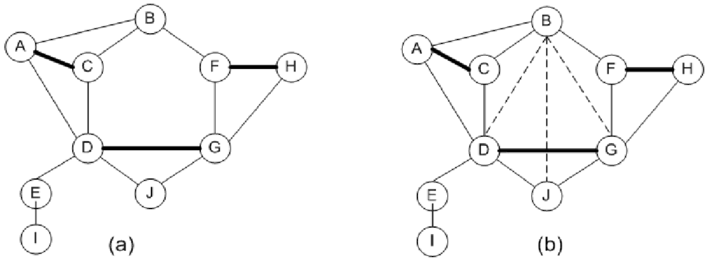


Figure 3. (a) Moral Graph (b) Filled Graph

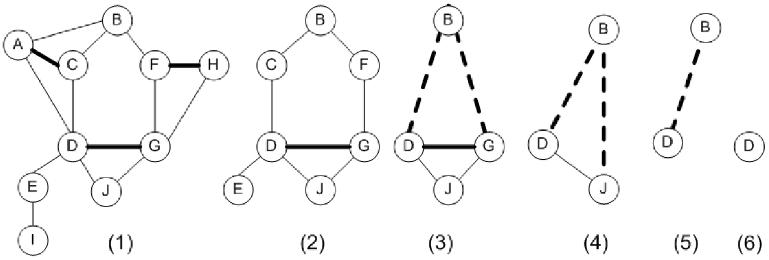


Figure 4. Elimination graphs sequence

Given an UG, $\mathcal{G} = (\mathcal{V},\mathcal{E})$, $\mathcal{V} = \{1,\dots n\}$, and an elimination order, $q = [q(1),\dots q(n)]$, we define the elimination process of its nodes as the sequence of *elimination graphs* $\mathcal{G}_k = (\mathcal{V}_k,\mathcal{E}_k)$, for $k = 1\dots n$, as follows: When eliminating node $q(k)$, we make its neighbors a *clique*, adding all missing edges between them.

$$\mathcal{V}_k = \{q(k), q(k+1), \dots q(n)\}, \quad \mathcal{E}_1 = \mathcal{E}, \quad \text{and, for } k > 1 \ ,$$

$$\{i,j\} \in \mathcal{E}_k \Leftrightarrow \left\{ \begin{array}{l} \{i,j\} \in \mathcal{E}_{k-1} \text{ , or} \\ \{q(k-1),i\} \in \mathcal{E}_{k-1} \text{ and } \{q(k-1),j\} \in \mathcal{E}_{k-1} \ . \end{array} \right.$$

The *Filled Graph* is the graph $(\mathcal{V},\mathcal{F})$, where $\mathcal{F} = \cup_{k=1}^{n} \mathcal{E}_k$. The *original* edges and the *filled* edges in $\mathcal{F}$ are, respectively, the edges in $\mathcal{E}$ and in $\mathcal{F} \backslash \mathcal{E}$. There is a computationally more efficient form of obtaining the Filled Graph, known as *simplified elimination*: In the simplified version of the elimination graphs, $\mathcal{G}_k^*$, when eliminating vertex $q(k)$, we add only the clique edges incident to its neighbor, $q(l)$, that is next in the elimination order.

The marginalization of variable X_i out of $P(X)$ corresponds to the elimination of the correspondent node in the elimination sequence. In order to marginalize on X_i, we have first to multiply all the potentials having X_i in its domain, and than sum out X_i. The domain of the resulting potential includes all the neighbors of X_i. In the Elimination Graphs, the corresponding elimination of X_i forms a clique with all of X_i's neighbors. Figure 3b and 4 show the Filled Graph and a synthetic version of the eliminations graphs for the order $q = [H,F,G,E,A,C,J,B,D]$.

The *Elimination Tree*, see George (1993), Pissanetzky (1984), and Stern (1994, 2006, 2008), portrays the dependencies among numeric operations on potentials, corresponding to dependencies in the node elimination process in the elimination graph. Hence, building the Elimination Tree for the corresponding Moral Graph makes it easy to see which variables can be eliminated simultaneously. Figure 5 shows the Elimination Tree for the order $q = [H,F,G,E,A,C,J,B,D]$.

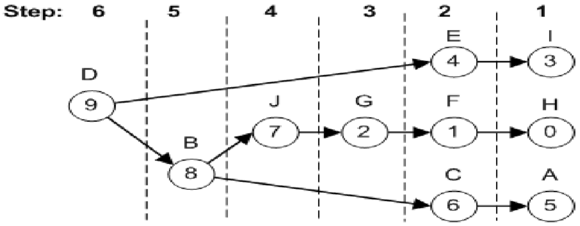
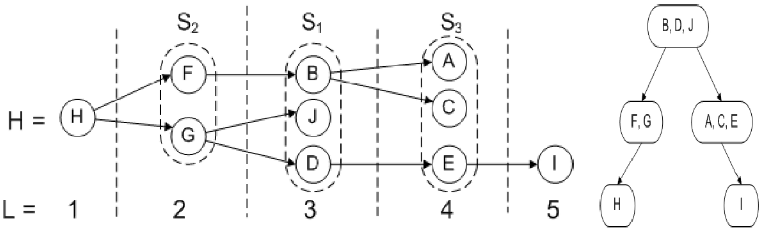


Figure 5. Elimination Tree for order $q = [H,F,G,E,A,C,J,B,D]$



Figures 6 and 7. Nested Dissection.

According to the Figure 5, six steps would be enough to eliminate all nodes: Variables *I*, *H* and *A* could be eliminated at first step and variables *E*, *F* and *C* at second one. Note that: (i) The Elimination Tree has the same structure of the tree of operations portrayed in Figure 2; (ii) A serial elimination would require 10 steps, one for each variable.

Clearly the Elimination Tree depends on the chosen elimination order, and the sparse matrix literature has many heuristics designed for finding good elimination orders. In this paper we adopted the an heuristic based on a nested dissections of the breadth-first tree rooted at a pseudo-peripheral vertex which, in turn, was found using the Gibbs heuristic, see Figures 6 and 7. These procedures are described in detail in George (1993), Pissanetzky (1984) and Stern (1994, 2006, 2008).

PARALLEL VARIABLE ELIMINATION ALGORITHM

The sequence of operations described in the previous section for inference in BNs can be summarized in the *parallel variable elimination algorithm*:

1. Symbolic phase:

1.1* Define the requisite variables X_R using, for example, the Bayes-Ball algorithm;

1.2 Build the Moral Graph (including only variables in X_R);

1.3 Choose a good elimination order using the Gibbs heuristics to find a pseudo-peripheral vertex used as a root for the Nested Dissection heuristic;

1.4 Symbolic Factorization: Execute the simplified elimination on the Moral Graph, and build the Elimination Tree;

1.5 Allocate the computation resources and prepare the data structures to execute the numeric operations.

2. Numeric phase: Using static data structures previously defined in the first phase:

2.1 While the root of the Elimination Tree was not executed: Based on the Elimination Tree hierarchy, trigger simultaneously threads to execute all variable eliminations ready to be done, including its numeric operations of multiplication and marginalization;

2.2 Normalize the remaining potential at the root.

RESULTS AND CONCLUSIONS

The proposed parallel algorithm was implemented and its performance was compared with a serial implementation. Both implementations were done in C and use the same functions to execute the basic operations for: Load the network; Multiply and marginalize potentials; and define the elimination order. The only difference between the two implementations is that the parallel version builds the Elimination Tree and, if possible, eliminate two or more variables simultaneously. Following this strategy we hope to isolate the effect of parallelization.

Table 1 displays some illustrative results. These experiments were done in a bi-processed machine running Linux and consists of 100 inferences for 7 distinct queries using the Hailfinder25 network (55 variables). The set of experiments suggests that the parallel implementation is much faster than the serial one for larger experiments. Queries requiring more variables or with a branched structure in the Elimination Tree allow the simultaneous elimination of several variables, for example experiments 1 to 6. Queries (or models) requiring less variables, or with a more linear structure in the elimination tree allow less parallelization of elimination operations. Consequently, in these examples, the serial implementation performed better due to the computational overheads imposed by the parallel version, namely, building of the Elimination Tree and the heavy context switch during execution. This was the case of experiment 7 in which the relations of dependence between the operations reduce the possibilities of parallelization.

Practitioners always want to solve larger models, most large models used in practice are sparse, and parallel or distributed computer are increasingly available. Hence, we see great potential for the parallel algorithm presented in this article.

Q.E.	N.R.	P.T.	S.T.	P.C.S.	S.C.S.
1	44	174	812	6498	901
2	44	95	125	6514	142
3	45	71	155	6553	183
4	46	74	155	6681	164
5	46	104	126	6817	138
6	48	106	125	7067	152
7	22	98	66	2948	78

Table 1:
Query example,
Numb. or requisite vars.,
Parallel time, Serial Time,
Parallel context switches.
Serial context switches,

REFERENCES

- E.C.Colla (2007). *Aplicação de Técnicas de Fatoração de Matrizes Esparsas para Inferência em Redes Bayesianas.* Ms.S. Thesis, Institute of Mathematics and Statistics, University of São Paulo.

- F.G.Cozman (2000). Generalizing variable elimination in Bayesian networks. *IBERAMIA/SBIA 2000 Workshop proceedings.* São Paulo, Tec Art, pp.27–32.

- A.Mandani, D.Heckerman, M.P.Wellman (1995). Real-world applications of Bayesian networks. *Comm. of the ACM* 38, 3, 24–26.

- R.Dechter (1996). Bucket elimination: An unifying framework for probabilistic inference. *12-th UAI proceedings*, 211–219. San Francisco: Morgan Kaufmann Publishers.

- A.George, J.R.Gilbert, J.W.H.Liu (ed.) (1993). *Graph Theory and Sparse Matrix Computation.* NY: Springer.

- F.V.Jensen (1996). *An introduction to Bayesian networks.* NY: Springer.

- S.L.Lauritzen, D.J. Spiegelhalter (1988). Local computations with probabilities on graphical structures and their application to expert systems. *J. Royal Statistical Society, B*, 50, 2, 157–224.

- J.Pearl (1988). *Probabilistic reasoning in intelligent systems: Networks of plausive inference*, Morgan Kaufmann, San Francisco.

- S.Pissanetzky (1984). *Sparse matrix technology.* Academic Press, New York, USA.

- R.Shachter (1998). Bayes-ball: The rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams). *14-th UAI proceedings,* 480–487. San Francisco, Morgan Kaufman.

- J.M. Stern (1994). *Esparsidade, Estrutura, Estabilidade e Escalonamento em Álgebra Linear Computacional.* IX Escola de Computação. UFPE, Recife.

- J.M. Stern (2006). *Decoupling, Sparsity, Randomization, and Objective Bayesian Inference.* Tech.Rep. MAC-IME-USP-2006-07.

- J.M.Stern (2008). *Cognitive Constructivism and the Epistemic Significance of Sharp Statistical Hypotheses.* Tutorial book for MaxEnt 2008, The 28th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, July 06-11, Boracéia, SP, Brazil.

- N.L.Zhang, Poole (1996). Exploiting casual independence in Bayesian network inference. *Journal of Artificial Intelligence Research,* 301–328.