1st International Symposium on Imprecise Probabilities and Their Applications, Ghent, Belgium, 29 June - 2 July 1999

Learning in Prevision Space

Stefan Arnborg
Kungliga Tekniska Högskolan, Stockholm, Sweden
stefan@nada.kth.se

Abstract

We investigate some problems related to implementation of uncertainty management, in particular the handling of computational and conceptual difficulties that easily appear in complex problems. The uncertainty polytope resulting from a set of inequality judgments on probabilities and means in a problem has very high dimension, but can be represented by a projection on a low-dimensional space if the judgments are structured into a graph with low tree-width. With this representation many judgments of independence become vacuous. The uncertainty polytope is high-dimensional and thus difficult to grasp or visualize. We propose a method to sample uniformly and efficiently from the polytope, as a means to obtain various summaries not obtainable by linear programming, such as volume, center of gravity, principal axes, etc.

Keywords. Learning, uncertainty, decomposition, uncertainty polytope

1 Introduction

The concept of uncertainty has rather deep philosophical connotations. How can I say something certain about uncertainty? What do I mean by saying that something (*e.g.*, some proposition A) is uncertain? One thing I mean is that I do not know whether or not A holds. It is a property of our language that this can mean different things. It can mean that A is permanently true or false, but I do not know which is the case. It can mean that A is dependent on an unstated context. It can mean that I do not know exactly what A means. The first of these cases can be analyzed using different schemes of Bayesian inference.

Bayesian reasoning can be regarded as an extension of Aristotelian deductive logic. Aristotle structured the reasoning process into assumption of premises, and a set of manipulative steps whereby new statements are found that are different but must be true if the assumptions are. We use terms invented by Aristotle when we talk about statements, *e.g.*, as premises and conclusions. So we have a vocabulary of statements, A, B, etc., and means to combine them and infer consequences. Deductive reasoning is the logic of certainty. It is a special case of any useful logic of uncertainty. Aristotle apparently realized the limitations of deduction as the only inference method in science. His attempts on the logic of uncertainty were quite influential but non-quantitative and apparently flawed[18], so now we must find inspiration from work done much later in order to finish his projects.

Bayesianism as a developing school of thought started in the 1930s with work by de Finetti and Jeffreys, among others. Instead of assuming certain statements to be true, we assume a number of statement plausibilities and derive plausibilities of other statements. When statement plausibilities change as a result of observations made, other statements will have their plausibilities changed. In 1946, R.T. Cox[4] published his findings on some properties required by any good calculus of plausibility of statements. He stated three requirements:

I: Plausibility is a real number, II: Consistency, III: Common sense.

A very lucid elaboration of Cox findings can be found in E. T. Jaynes posthumous manuscript[11], Ch 2. The conclusion was that every type of reasoning with the plausibility of statements satisfying I, II and III above, as they interpret them, is equivalent to computing with probabilities after a rescaling of the plausibility measure.

This is because he found a rescaling that must satisfy the two first rules of probability: $P(AB|C) = P(A|C)P(B|AC)$ and $P(\overline{A}|C) = 1 - P(A|C)$. Bayes rule $P(A|BC) = P(B|AC)P(A|C)/P(B|C)$ is an immediate consequence of the probability rule for con-

junction and commutativity of conjunction. A similar derivation (more related to de Finettis work) was published by Lindley[15]. The accompanying discussion is recommended reading as an illustration of how difficult this topic is. The analysis has been both praised and criticized regularly since 1946, see *e.g.,* [9]. In Cox argument it is easily seen that the conclusion hangs critically on (I), sometimes referred to as the Bayesian dogma of precision. Walley[21] made a similar derivation without assumption (I), resulting in a system where plausibility is measured by an interval of numbers. The accompanying methodology suggests that our uncertainty about the world is measured by a polytope in 2^n-space of probabilities for a world of n binary variables (atomic statements). This is in contrast with most applications of Bayesianism, where the situation would be measured either with a point in 2^n-space, for the purpose of decision making, or with a probability distribution over 2^n-space for the purpose of learning from observation or experience. If we have a pdf (a distribution over 2^n-space) to measure uncertainty, all its information required for Bayesian decision making is summarized in its mean.

The standard Bayesian view has no room for total ignorance - there is one prior, first- or second-order. But it is completely natural to model some types of ignorance with families of Bayesian assessments, among which we do not want to choose. Each such assessment can be thought of as coming from one expert or even a mode of an expert. Although standard Bayesianism advocates earliest possible fusion of such assessments, this may not always be possible or even desirable. An interesting question is then how learning should be realized: Do we never want to choose between the assessments, or is there a useful mechanism by which some assessments become downplayed by observations - as one would imagine happens when a decision maker tries to rationally use advice given by a collection of quarreling experts, using as input each experts past performance?

We will investigate problems related to combining previsions and observations, as a feasibility study concerning new tools for analysis of knowledge and observations related to medical and human brain informatics[8].

2 Definitions

The concepts underlying this discussion are considered in many application areas which have developed rather different terminologies. The usage in this paper is defined here:

We consider worlds that can be defined by the truth or falsity of each of n atomic statements in a vocabulary V, which we call *binary variables*. We consider binary variables here, since the generalization to variables taking more than two values is obvious. A *possible world* is specified by an assignment of true or false to each of these n variables. It is thus a corner of the n-dimensional hypercube, and there are 2^n possible worlds. A probability distribution over these worlds is specified with a point in the 2^n-dimensional hypercube, giving the relative probabilities of the 2^n possible worlds. We use the quantities y_s to denote such probabilities, where s is a binary string of length n. When referring to their computation by linear programming, the y_s will be called *LP variables*. A *subworld* (or subvocabulary) of the n-variable world is defined by a subset $W \subset V$ of its variables, and denoted in probabilities with a superscript which is a list of its variables. The possible subworld probabilities are obtained by summing the possible world probabilities over those indices corresponding to variables not included in the subworld. This process is also known as marginalization. The possible subworld probabilities are denoted x_s^W or z_s^W, where W is a list of binary variables and s is a list of corresponding binary indicators. The subworld probabilities will also sometimes be used as LP variables.

3 Assessment and Learning

In the first treatise on probability, Bernoulli gave the method of assigning probabilities to hypotheses, the principle of *insufficient reason* or *indifference*: If there are n exclusive and exhaustive hypotheses to choose from, and s of them imply success, and there is no information leading us to distinguish among them, then the probability of success is s/n. Even if we start out with some queer measure of plausibility satisfying requirements I, II and III, the basic rule of assigning plausibility refers to the probability, and not to the queer measure we started with. There is a problem with this recipe: it says that probability is conditioned by all information we possess that is relevant to the problem under consideration. Many questionable derivations in Bayesian analysis result from taking this condition too lightly. Bayes, being more humble, had the opposite problem: it has been reported that his ambivalence towards the uniform prior caused him not to publish his paper.

It is equally dubious to omit from the analysis information that we have, as it is to enter information which we do not have. In the notation $A|C$, C stands for this information. It is useful to call it the *context* of the analysis. It is important to note that two individuals with the same information state (context) will in principle assign the same probabilities, so they are not truly subjective. But this is a hypothetical

statement - it is not practically possible to measure the information state of an individual. The method of lower previsions seems to relate to the situation where we have a set of contexts that we do not want to fuse by weighting. The uncertainty polytope is the result of keeping an unweighted set of contexts and using their convex hull (all possible weightings) as a set of possible contexts.

All non-trivial applications of uncertainty reasoning must deal with the problem of learning from experience. In statistics oriented systems the learning principle seems to be application of Bayes rule to a second-order probability function, which is a generalization of the model choice principle using Bayes rule. This seems to be the most controversial part of Bayesianism, because conclusions are typically not robust with respect to choice of prior. In many application areas this is not felt as a problem, but in others it is. This topic is one of the most discussed in statistical methodology. It caused Bayes problems, Laplace was ridiculed for his choice of example. It looks as a manifestation of an eternal (Socratic) educational riddle: How can you learn if you do not already know? Recently proposals were made to use families of learning functions with some canonical properties like the imprecise Dirichlet prior[20]. This is a non-committing family of priors which however has a 'stiffness' parameter that controls speed of learning. It could be seen as a second-order manifestation of the Bayesian dogma of precision which is somewhat hard to escape. In the current discussion we can see that the view of the uncertainty polytope as a description of consensus among different experts can be augmented by introducing learning into the expert set. With this view we say that each expert should learn by experience, in which case the uncertainty polytope changes by observation, but remains the convex hull of the opinions of the different experts. The above is meant to suggest that it is a good idea to look at the uncertainty polytope and in particular to get handles on the computational and conceptual challenges it poses.

4 Joining Small Worlds

The method of lower previsions allows a user to impose judgments on the probability space of a problem, and each such judgment may become modified by observations made. Thus, a problem with n binary variables is described by a set of probability distributions over the 2^n possible worlds, each described by a 2^n-tuple of probabilities summing to one. A judgment in this system is a linear constraint on these probabilities. Thus, a state of uncertainty resulting after a set of judgments have been passed is described by a polytope in 2^n-space. Most inferences required are in the form of estimates of a linear function of the probability vector, and with the only information that the vector lies in a polytope, the answer to the inference problem is an interval of numbers that can be found in two linear programming optimizations, one maximization and one minimization. For large n it is not possible to attack the linear programming problem using standard methods. However, using the technique of decomposability, it is often possible to solve combination and optimization problems with many variables, provided the judgments are reasonably structured. The method has been used under many names in quite many application areas: [1, 17, 16, 14, 19]. We illustrate the method with an informal discussion with a simple example instead of introducing one of the rather complex notational systems invented to cover the general case.

The analysis starts with a graph where each of the n binary variables is a vertex. For each imposed judgment involving a set of k of these variables, and for each desired inference of a quantity referring to k variables, we draw edges in the graph that completely connect the corresponding k vertices with $\binom{k}{2}$ edges. Then we make a tree-decomposition of the graph, *i. e.,* we construct a number of subworlds we call small worlds and connect them in a tree structure in such a way that

1 For every judgment made, the variables mentioned in the judgment are all present in at least one small world.

2 For each variable, the set of small worlds in which it is present forms a contiguous part of the tree.

In general it is difficult to find a tree-decomposition with smallest possible size of its largest subworld. However, there are several methods proposed that work in linear time for a fixed largest subworld size [2, 3] and the method proposed in [2] was implemented in the Graphed system[10]. Now, we treat the different subworlds separately, and introduce a set of 2^w probabilities x_s^W for a world W of w binary variables, each representing the probability of one of the states of W. These new probabilities are related to the probabilities y_s of the original n-variable world by summation over those indices not appearing in W. Each judgment referring only to variables in subworld W can be translated to a linear constraint on these probabilities. In this way we get much less than 2^n variables for the whole problem, if the judgments are reasonably structured. But we must also connect the probabilities of different subworlds, since some variables exist in several of them. The following method does that.

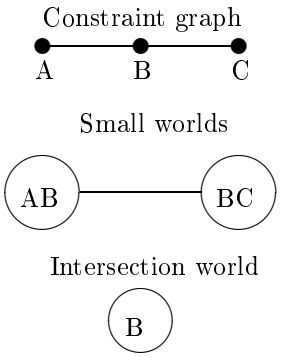

Figure 1: Decomposition of constraint graph

For every edge in the decomposition tree, we must connect the probabilities of the two subworlds on each side by marginalization, so that both subworlds give the same probabilities of the world defined by their intersection. Let the edge connect worlds W_1 and W_2, and let $W' = W_1 \cap W_2 = \{V_1, \ldots V_r\}$ be the variables common to W_1 and W_2. Then, for each of the 2^r states of these r variables we introduce the probability $z_s^{W'}$ (s is a binary indicator string). These quantities can be obtained both by summing over x^{W_1} and over x^{W_2} probabilities, and the two ways to obtain them must give equal results. The first condition **1** on the tree-decomposition will now guarantee that every probability appearing in a judgment can be expressed in the possible subworld probabilities x rather than in the many more possible world probabilities y. The second condition **2** is required for guaranteeing that the different subworlds are mutually consistent, so that a feasible point in the decomposed world problem corresponds to some global solution expressible with the y_s probabilities. The mechanism will be explained in the example.

Example: Assume we have variables A, B, and C. Then we have made judgments involving $P(A)$, $P(B)$, $P(AB)$ and $P(BC)$, and want to make an inference about $P(C)$. We decompose this problem into two worlds AB and BC, and their intersection world is B, see fig. 1.

The probabilities y_s are connected to the probabilities x_s^{AB} and x_s^{BC} by:

$$\begin{aligned} x_{10}^{AB} &= y_{101} + y_{100} \\ x_{11}^{AB} &= y_{111} + y_{110} \\ x_{01}^{AB} &= y_{011} + y_{010} \\ x_{00}^{AB} &= y_{001} + y_{000} \\ x_{10}^{BC} &= y_{110} + y_{010} \\ x_{11}^{BC} &= y_{011} + y_{111} \\ x_{01}^{BC} &= y_{101} + y_{001} \\ x_{00}^{BC} &= y_{100} + y_{000} \\ 1 &= \sum_{ijk} y_{ijk} \end{aligned} \tag{1}$$

The probabilities z_s^B of the intersection world are connected to the probabilities of W_1 and of W_2 by the following:

$$\begin{aligned} x_{00}^{AB} + x_{10}^{AB} &= z_0^B \\ x_{01}^{AB} + x_{11}^{AB} &= z_1^B \\ x_{00}^{BC} + x_{01}^{BC} &= z_0^B \\ x_{10}^{BC} + x_{11}^{BC} &= z_1^B \end{aligned}$$

When asking for the probability of C after assigning real numbers to p_A, P_B, P_{AB} and p_{BC}, we get the linear programming problems with non-negative LP variables of maximizing and minimizing $P_C = x_{01}^{BC} + x_{11}^{BC}$ under constraints (where the question marks indicate a relation $<$, $>$ or $=$, depending on the type of judgment made):

$$\begin{aligned} x_{10}^{AB} + x_{11}^{AB} &\;?\; p_A \\ x_{11}^{AB} &\;?\; p_{AB} \\ x_{00}^{AB} + x_{01}^{AB} + x_{10}^{AB} + x_{11}^{AB} &= 1 \\ x_{10}^{BC} + x_{11}^{BC} &\;?\; p_B \\ x_{11}^{BC} &\;?\; p_{BC} \\ x_{00}^{BC} + x_{01}^{BC} + x_{10}^{BC} + x_{11}^{BC} &= 1 \\ x_{00}^{AB} + x_{00}^{AB} - z_0^B &= 0 \\ x_{00}^{BC} + x_{01}^{BC} - z_0^B &= 0 \\ z_0^B z_1^B &= 1 \end{aligned} \tag{2}$$

The value of a conditional probability or conditional mean is not linear in the LP variables introduced, but its range can be found by checking the feasibility of the system after adding the linear constraint with a particular value of the conditional quantity, and iteration using duality[21]. In order to estimate $P(\overline{C}|B)$ we would thus consider the feasibility of the system obtained by adding the constraint $x_{10}^{BC} - r(x_{10}^{BC} + x_{11}^{BC}) = 0$, for different numerical values of r.

It should be clear that a feasible point of the entire system, with LP variables x, z, and y, can be projected to a feasible point in the system where equations involving the y, namely the set (1), have been removed. The reverse is also true. To see this, first note that all judgments have been expressed in the small world probabilities x_s^W. Given a feasible point in x-z-

space, we can construct a feasible set of y-values as follows: For one particular probability, say y_{ijk}, we fetch all x and z variables that represent its projections on small worlds (namely x_{ij}^{AB} and x_{jk}^{BC}) and intersection worlds (in this example z_j^B). If one of these is zero, y_{ijk} will also be zero. Otherwise, we get y_{ijk} by multiplying together all the x-values found and dividing by all the z-values found. This is a simple application of the general method of finding a pdf given as a set of potentials[5, 22]. In our example we would construct the probability of $A\overline{B}C$ as: $y_{101} = x_{10}^{AB} x_{01}^{BC} / z_0^B$, unless the denominator is zero, in which case both numerators are also zero and y_{101} too. We can now verify that equation set (1) is satisfied. For example, we get $y_{100} + y_{101} = x_{10}^{AB}(x_{01}^{BC} + x_{00}^{BC})/z_0^B$, which, by (2), equals x_{10}^{AB} as it should.

This argument shows that there is always a solution for the possible world probabilities y, but it seems not easy to characterize the full 2^n-space polytope without actually introducing exponentially many variables.

Another thing that this example shows is that a judgment that A and C are independent conditional on B does not change the problem, since there is always a solution where quantities in different worlds are independent conditional on the state of some intersection world between them. It seems as if this is a natural and no-cost judgment of independence. Other conditional independence judgments are possible to argue for, but they seem fairly difficult to interpret or relate to a causality argument. When they are needed, more complex solutions will still be required[21], but they would hopefully be used infrequently.

The number of variables that need to be introduced in the outlined method can be somewhat reduced using more optimal state reduction methods, like tree automata[1] or BDD technology[16]. This leads to some difficulties in interpretation and surprises in the number of states actually produced, however. The decomposition idea has been used before in uncertainty management, typically in computations of probabilities where non-edges mean independence, like for graphical decomposable probability models[13]. The applications closest to this one is the use in probabilistic logic[1] and in finding probability intervals for partially specified probability models[19]. The difference here is that we tried to analyze the problem in such a way that the question of independence can be decoupled from the analysis, and consequently compatible judgments of independence can be ignored (they will not influence intervals of probabilities that can be seen in the decomposed model), and that likewise conditional beliefs can be entered directly as judgments of conditional probabilities. Conditional probability intervals can be found using iteration, which gives a belief update function.

5 Summarizing the Uncertainty Polytope

The polytope describing the state of uncertainty will usually be high-dimensional and impossible to visualize in a comprehensible way, even in cases where the number of variables can be brought down by decomposition. For this reason it will be desirable to compute a summary. Such summaries can be measures of location and extent. The LP formulation makes it relatively easy to find ranges of the polytope in various directions. But some more handles seem to be called for in order to assess the adequacy of the modeling effort. It seems clear from looking at specific examples that the analysis of a problem benefits from other types of summaries like volumes, centers of gravity and principal axes. Such quantities depend on the geometry of the polytope and are typically obtained by integration over it. A high-dimensional polytope is too complex for standard numerical integration methods, and there are convincing results showing that no deterministic integration method will work[6]. The standard practical approaches to this problem have been ad hoc use of Monte Carlo simulations, but the convergence analysis of such methods has typically been missing, and there have been no guarantees that the simulation is statistically valid. Fortunately, recent results using the theory of rapidly mixing Markov chains has indicated that integration problems with required accuracy are possible with Monte Carlo methods. In [12], it is shown that the volume of a polytope of dimension n can be approximated within relative error ϵ using $O^*(n^5)$ polytope containment tests. The approximation is obtained by a random walk on a modified version of the polytope and is quite complex. We propose that the only feasible summarization of a high-dimensional polytope must be obtained via a uniform sample. A two-dimensional example is shown in fig. 2. One method is as follows: A set of independent variables is chosen such that their values determine, by equality constraints, the remaining variables. The projection of the polytope on these variables gives a full-dimensional body to work with. Initialize all LP variables to a point in the polytope. In each step, choose an independent variable and find a proposed new value, uniformly over its feasible values given the current values of the other independent variables. The mean probability over the polytope is estimated with the average of the coordinates occurring in the chain. As can be seen in fig. 3, the chain has high autocorrelation and it takes a long chain to

get a uniform sample.

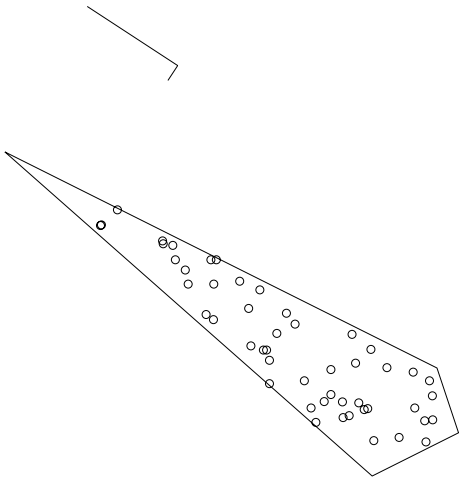

Figure 2: A uniform sample of a polytope, with principal components

The procedure has been tested on polytopes with known geometry, and converges rapidly for these, even for high dimensions. However, it is also clear that it is easy to construct cases (thin and tilted polytopes, as in the example) where the chain mixes intolerably slowly, and this indicates the need to look at the solution with guaranteed accuracy proposed in [12]. Two measures taken seem important for improving the performance considerably, even in an algorithm that works with heuristic assessment of statistical convergence: In a first step, the volume is transformed by a randomized algorithm so that the radius ratio of an enclosing and an enclosed ball is made small. This will make it possible to traverse the volume with fewer steps. This is done by taking a sample of points in the body, finding its principal components, and scaling to unit variance using the eigenvalue found for each principal axis. A second feature is a technical design of the chain (rounding) that is allowed to expand slowly from the center of the polytope, in order to improve provable mixing rates. These two techniques have an obvious intuitive appeal and would seem to speed up the computation in practice for difficult polytopes. In fig. 4 we can see how the chain is improved if a small sample is used to rescale the polytope by the principal components of the sample. This chain has the same length (40 steps) as that of fig. 3. It can easily be translated back to the original polytope by a linear transformation. In higher dimensions this effect becomes more pronounced.

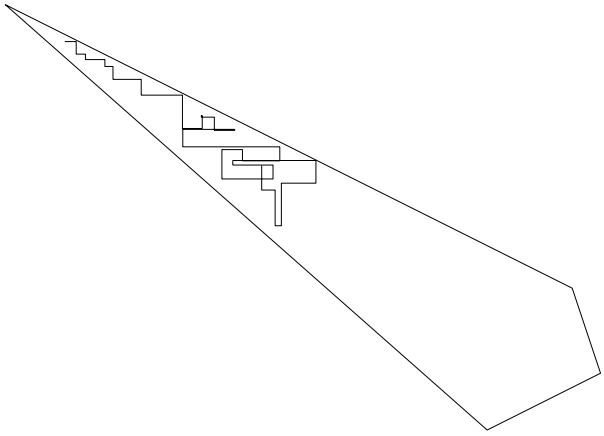

Figure 3: Walk in a thin and tilted polytope

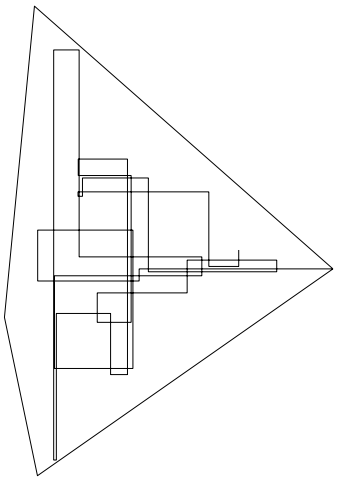

Figure 4: Walk in the rounded polytope

6 Conclusion

We investigated some problems of computational and conceptual complexity in imprecise probability methodology. Decomposing a problem by constraint structure makes it solvable with much fewer variables than originally needed, and facilitates many types of independence judgments.

The exploration of the uncertainty polytope for the purpose of presenting summaries was also described with a proposed solution, although preliminary.

Acknowledgments

The referees and Peter Walley have given me good advice on the structure of this presentation, and some additional references.

References

[1] S. Arnborg. Graph decompositions and tree automata in reasoning with uncertainty. *J. Expt. Theor. Artif. Intell.*, 5:335–357, 1993.

[2] Stefan Arnborg, Derek G. Corneil, and Andrzej Proskurowski. Complexity of finding embeddings in a k-tree. *SIAM Journal on Algebraic and Discrete Methods*, 8(2):277–284, 1987.

[3] Hans L. Bodlaender and Ton Kloks. Efficient and constructive algorithms for the pathwidth and treewidth of graphs. *Journal of Algorithms*, 21(2):358–402, September 1996.

[4] R.T. Cox. Probability, frequency, and reasonable expectation. *Am. Jour. Phys.*, 14:1–13, 1946.

[5] A. P. Dawid and S. L. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, 21:1272–1317, 1993.

[6] Martin Dyer, Alan Frieze, and Ravi Kannan. A random polynomial time algorithm for approximating the volume of convex bodies. In *Proceedings of the Twenty First Annual ACM Symposium on Theory of Computing*, pages 375–381, Seattle, Washington, 15–17 May 1989.

[7] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, 1996.

[8] Håkan Hall, Stig Larsson, and Göran Sedvall. Human brain informatics - HUBIN web site. 1999. `http://www.ki.se/cns/hubin/`.

[9] J. Halpern. A counterexample to theorems of Cox and Fine. *Journal of AI research*, 10:67–85, 1999.

[10] M. Himsolt. Graphed: a graphical platform for the implementation of graph algorithms. In R. Tamassia and I. G. Tollis, editors, *Graph Drawing*, volume 894 of *Lecture Notes in Computer Science*, pages 182–193. DIMACS, Springer-Verlag, October 1994. ISBN 3-540-58950-3.

[11] E. T. Jaynes. *Probability Theory: The Logic of Science*. Preprint: Washington University, 1996. `ftp://bayes.wustl.edu/Jaynes.book/`.

[12] Ravi Kannan, László Lovász, and Miklós Simonovits. Random walks and an $O(n^5)$ volume algorithm for convex bodies. *Random Structures & Algorithms*, 11:1–50, 1997.

[13] Steffen L. Lauritzen. *Graphical Models*. Clarendon Press, 1996.

[14] Thomas Lengauer and Egon Wanke. Efficient processing of hierarchical graphs for engineering design. *Bulletin of the European Association for Theoretical Computer Science*, 35:143–157, June 1988. Technical Contributions.

[15] D. V. Lindley. Scoring rules and the inevitability of probability (with discussion). *Internat. Stat. Rev*, 50:1–26, 1982.

[16] R. E. Bryant. Symbolic Boolean Manipulation with Ordered Binary-Decision Diagrams. *ACM Computing Surveys*, 24(3):293–318, September 1992.

[17] P. Shenoy. A valuation-based language for expert systems. *International Journal of Approximate Reasoning*, 3(5):383–411, September 1989.

[18] Robin Smith. Logic. In P. Barnes, editor, *Aristotle*, Cambridge, 1995. Cambridge University Press.

[19] L. van der Gaag. Computing probability intervals under independency constraints. In P. P. Bonissone, M. Henrion, L. N. Kanal, and J. F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 457–466. North-Holland, 1991.

[20] P. Walley. Inferences from multinomial data: Learning about a bag of marbles. *J. R. Statist. Soc. B*, 58, 1996.

[21] P. Walley. Measures of uncertainty in expert systems. *Artificial Intelligence*, 83, 1996.

[22] Dag Wedelin. Efficient estimation and model selection in large graphical model. *Statistics and Computing*, 6, 1996.