Technical Report TR-2010-06
University of British Columbia

Where do priors and causal models come from? An experimental design perspective

Hendrik Kueck and Nando de Freitas

April 7, 2010

Department of Computer Science, University of British Columbia
2366 Main Mall, Vancouver, BC Canada V6T 1Z4

Abstract

In this pedagogical note, we treat prior elicitation and experimental design within a common decision making framework. This is contrary to the standard practice of assuming that priors are already available when performing Bayesian experimental design. We argue instead that these processes are intertwined. We demonstrate our decision-theoretic stance in the setting of learning a causal Bayesian network prior sequentially. We choose among a set of actions (consulting an expert, observing more data and conducting interventions and experiments) to maximize the gain in information.

1 Introduction

It is unquestionable that priors play an important role in machine learning, artificial intelligence and statistical modeling (Bernardo and Smith, 1994; Bishop, 2006; Gelman et al., 1995; Mackay, 2006; Russell and Norvig, 2002). The question of prior formulation has also gained enormous momentum in cognitive science in recent years; for excellent surveys see Griffiths and Tenenbaum (2006); Tenenbaum et al. (2006); Shafto et al. (2008); Yuille and Kersten (2006). Priors allow us to incorporate preferences (*e.g.* sparsity) as well as subjective knowledge into the modeling process (Mackay, 2006). Priors may even arise naturally as a consequence of modeling assumptions, such as exchangeability (J. K. Ghosh, 2003). There are also theoretical reasons in the study of admissibility of estimators, *e.g.* Stein's paradox, that motivate the existence of priors (Robert, 2007).

Probably the two most immediate ways in which priors are used are inference and decision making. In inference, Bayes rule provides the mathematical tool for updating this prior belief in light of observed data:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)},$$

where $p(\theta|D)$ is the posterior distribution, $p(D|\theta)$ is the likelihood of the observed data and $p(D)$ the marginal likelihood of the data. In the rational approach to decision making, which is prevalent in game theory, experimental design, control and reinforcement learning, one defines a function $u(y, d, \theta)$ that measures the utility of an outcome y for decision d and model parameters θ (von Neumann and Morgenstern, 1947). An agent expecting to behave rationally must maximize its expected utility:

$$d^* = \arg\max_d \iint u(y, d, \theta)p(y|\theta, d)p(\theta)\, dy\, d\theta.$$

But, where do the priors come from?

Some have argued that priors (or sub-components of the priors) come from experts. In this line of work, the task of 'extracting' a person's belief and expressing it as a probability distribution is known as *prior elicitation* (O'Hagan and Oakley, 2004). Often it will be an expert on the particular system being modeled whose prior knowledge and beliefs need to be elicited. The problem is that people do not internally represent their beliefs and knowledge in terms of probability distributions. The representations and heuristics we employ to answer questions about our beliefs lead to various types of biases and inaccuracies when making judgements about

probabilities (Kahneman et al., 1982; Garthwaite et al., 2005). Despite these difficulties, there has been a significant amount of research on the problem of prior elicitation; see *e.g.* Jenkinson (2005); O'Hagan et al. (2006); Garthwaite et al. (2005).

The difficult and time-intensive prior elicitation process is most typically carried out in cases where the elicited expert beliefs are directly used as a basis for decisions. However even in Bayesian inference and experimental design, where the elicited distributions are used 'just' as a prior for analyzing subsequently gathered data, the effort can be justifiable. This is the case in settings where acquiring data is expensive and/or the model is not identifiable. Important applications include medical diagnosis, choice of treatment and clinical trials in medicine, safety assessment in the nuclear industry as well as applications in psychology (*e.g.* user studies (Myung and Pitt)), economics, engineering and many other fields. That is, if gathering data or conducting experiments is expensive, we want as much prior knowledge as possible to guide the decisions and reduce costs.

In Bayesian data analysis and experimental design the prior is usually assumed to have been acquired before any data is collected.. While traditionally the expert's belief would be elicited before any data is acquired, we argue that this is not necessarily optimal. For example eliciting the full belief of an expert about relationships of multiple variables in a complex system is practically intractable. However it might be possible to gather some data first and then use the knowledge gained from the data to ask the expert directed questions about important relationships between parameters which remain uncertain. The data can potentially narrow down the space of possible models to a size where elicitation becomes more tractable.

In this technical note, we argue that it is beneficial not to treat elicitation, inference and experimental design as separate stages, but to instead handle them jointly. Just as the prior can inform the design of experiments, available data can inform the choice of questions to ask an expert.

In recognizing this, we propose a decision theoretic approach to treat the problems of Bayesian inference, elicitation and experimental design jointly. The approach allows us to optimally decide whether to gather more data, perform an intervention (and which) or ask an expert (and what question) in the setup of learning a seemingly simple, but nonetheless very difficult, probabilistic causal model.

The expert is assumed to provide indirect knowledge about the world. That is, the expert provides answers about the state of the world and the statistician can use these answers to update his own beliefs. Fundamentally the expert's answers in this framework are treated very similarly to the out-

comes of experiments conducted in the system of interest. The qualitative differences are (i) the expert can potentially provide information about the system at a higher level of abstraction, (ii) because the expert's answers are (presumably) based on a significant amount of experience with the system, one can hope to gain more information than from a small number of observations or experiments, and (iii) the cost of asking the expert differs from the cost of conducting experiments. Depending on the setting it could be significantly higher or lower. The cost will typically also vary for different types of questions and experiments.

2 An application to causal discovery

We demonstrate the idea of interleaving prior elicitation and data acquisition via experiments using an illustrative causal discovery problem. Causal discovery or structure learning in general is a very hard class of problems. One of the reasons for this is that the space of all possible structures grows super-exponentially with the number of variables in the system. For a system with d variables there are $O(d!\, 2^{\binom{d}{2}})$ possible DAGs (directed acyclic graphs) (Robinson, 1973). As a consequence, eliciting an experts full prior belief over all possible structures is infeasible except for very small numbers (less than 4) of variables. In addition to the graph structure, the expert's opinion about the parameters of each network ideally would need to be elicited as well, adding another layer of infeasibility.

In contrast to Bayesian networks, causal networks encode causal relationships and not just independencies amongst the variables. This representation has many advantages (see Pearl (2000) for a comprehensive overview of causal networks). One of them is that people tend to understand the world in terms of causal relationships. Questions about causal links are therefore much more intuitive than questions about conditional independencies, making these kinds of questions easier to answer with high confidence. For example most people will be comfortable answering a question such as 'Do you think turning on the sprinkler will affect the wetness of the grass?'. On the other hand they will likely have a harder time with a question like 'Given that the grass is wet, do you think the state of the sprinkler and the weather are independent of each other?'

In the simple example presented in the following we are trying to learn the causal relationships between 5 variables. Figure 1 shows the true example network that our algorithm will be trying to reconstruct. This example network is taken from Friedman et al. (1998) (also used in Eaton and Mur-

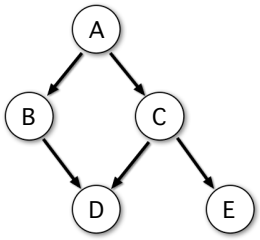

Figure 1: The example causal network that we are trying to reconstruct from observations, experiments with interventions and questions to an expert.

phy (2007b)). The 5 variables take on binary values. We chose the causal dependencies in our example network to be quite strong (the parameters for the conditional Bernoulli distributions for the pairwise causal links in the graph are 0.9 or 0.1). The 5 variables in the system are all assumed to be observable. Even for only 5 variables, eliciting an expert's belief on the probability of all possible structures is prohibitive (there are 29281 possible DAGs with 5 nodes).

Simply observing the values that the nodes take on jointly makes it possible to infer dependencies and independencies amongst the variables but is not sufficient to determine the direction of all causal dependencies. The 4 causal networks shown in Figure 2 are *likelihood equivalent* with the true network from Figure 1. This means that even given an infinite amount of observation data alone, it is not possible to distinguish between these 4 networks. In order to determine the direction of the causal links it is necessary to either use additional information such as prior belief elicited from an expert or to use experiments with controlled interventions.

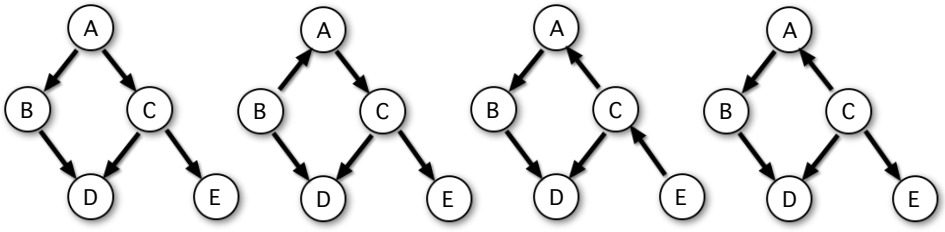

Figure 2: The likelihood equivalence class of the causal graph shown in Figure 1. Using observations alone it is not possible to distinguish between these 4 graphs.

In our example we assume that we have access to an expert with good knowledge of the system as well as the ability to carry out perfect interventions in which we control the value of one variable in the system. We chose to model the expert using the external approach, in which we are not trying to explicitly reconstruct the expert's internal probabilistic model. Instead we assume that we can ask the expert questions about the system and that she will provide us with answers which are with high probability correct. We treat this learning problem as a sequential optimal design problem where at each stage we can choose between 3 different types of actions:

- Ask the expert about the existence and direction of a specific causal link.
- Run purely observational experiments.
- Conduct experiments with perfect intervention, in which one of the variables in the system is controlled.

In total this yields 16 possible actions (10 possible edges to ask about and 5 possible nodes to intervene on). We take the myopic approach to optimal design, which means that we only try to find the optimal next action based on the information acquired so far but do not consider future effects. To decide between these 16 actions at each stage we first need to choose a utility function. In causal discovery the goal is often to understand the causal links between elements of a system. A reasonable goal would then be to minimize the uncertainty about the individual causal links, while also taking into account the costs of different types of experiments/questions. We formalize this next.

Let $G \in \mathcal{G}$ denote one particular causal network structure (out of the set of all possible DAGs $\mathcal{G}$) and let e_{ij} be the state of the edge between node i and j in the graph:

$$e_{ij}(G) = \begin{cases} 0 & \text{if there is no edge between nodes } i \text{ and } j \text{ in G,} \\ 1 & \text{if there is a directed edge from } i \text{ to } j \text{ in G,} \\ 2 & \text{if there is a directed edge from } j \text{ to } i \text{ in G.} \end{cases}$$

If we have some previous data D (consisting of observation trials, outcomes of experiments with interventions and expert answers to questions) then $p(G|D)$ is our posterior distribution over graphs. This induces a marginal distribution over edge states

$$p(e_{ij} = s|D) = \sum_{G \in \mathcal{G}} \delta_{s,e_{i,j}(G)}\, p(G|D). \tag{1}$$

The entropy of the state of one particular edge in the posterior is then

$$H(e_{ij}|D) = -\sum_{s=0}^{2} p(e_{ij} = s|D) \log p(e_{ij} = s|D). \tag{2}$$

The goal in this example is to minimize the sum of the entropy of all edge states $\sum_{i,j} H(e_{ij}|D)$. That is, we want to collect data such that we are maximally certain about the state of the individual causal links in the causal network.

Let ζ denote one of the 16 possible actions and let y be a possible outcome (answer to a question or outcomes of experimental trials). Each type of action has a different cost $c(\zeta)$ associated with it. This reflects the fact that it might for example be much more costly to perform an intervention in the system than to just passively observe. For the example presented in the following we used a cost of 1.0 for collecting a set of 20 observations, 3.0 for an experiment involving intervention on one node (and also 20 trials) and a cost of 2.0 for asking the expert about a specific edge.

The problem we are trying to solve is to choose the source of information ζ that we expect to give us the most information (reduction in entropy) per cost. The amount of information that an outcome y would provide about the state of one specific edge between node i and j is given by the difference in entropy between the belief distribution after and before observing y:

$$H(e_{ij}|D) - H(e_{ij}|D, \zeta, y).$$

The overall expected information gain per cost for a particular source of information ζ is then

$$EU(\zeta) = \sum_{y} p(y|\zeta, D) \frac{\sum_{i,j} H(e_{ij}|D) - H(e_{ij}|D, \zeta, y)}{c(\zeta)}, \tag{3}$$

and the goal is to choose $\zeta^* = \arg\max_{\zeta} EU(\zeta)$.

Computing the posterior distribution $p(G|D, \zeta, y)$ (incorporating the evidence from the different types of data) is fairly involved and we cannot describe it in full details here. We use a uniform prior $p(G)$ on structures and the BDeu prior (Heckerman et al., 1995) on the parameters, which allows for closed form computation of the marginal likelihood of experimental data (with or without interventions), and build on the computational approach presented in Eaton and Murphy (2007b).

Computing the expected utility of asking about an edge according to Equation (3) involves only 3 possible outcomes and is computationally feasible to do exactly. However in the case of experimental trials, we resort to

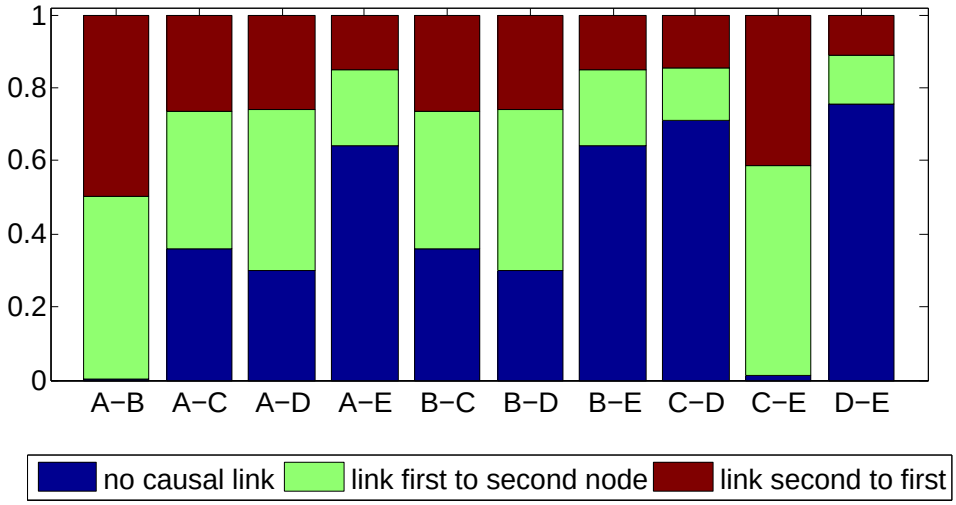

Figure 3: The marginal distributions $p(e_{ij}|D)$ over the existence and direction of causal links between any two nodes in the graph. The data D here consists of a first round of observations, consisting of 20 measurements of all variables. (Best viewed in color.)

sampling to simulate hypothetical data from $p(y|\zeta, D)$ by first sampling a structure G from $p(G|D)$ and then sampling the outcomes y of experimental trials (taking into account possible interventions). Equation (3) is then approximated using these samples. For the small example network used here the distribution $p(G|D)$ can actually be computed and stored exactly. For larger networks the technique proposed in Eaton and Murphy (2007a) could be used to efficiently sample from $p(G|D)$.

Given the completely uninformative prior and the costs of the experiments as given above it turns out that at the very beginning of the sequential learning process collecting observational data is the option expected to be most cost-effective. Figure 3 shows the edge marginals $p(e_{ij}|D)$ *after* a first round of collecting a set of 20 observations. Using these observations alone we see that there is basically no uncertainty left about the existence of an edge between A and B and C and E. However the direction of these edges is completely unclear. Based on the information gained from these first 20 observations we then evaluate the expected utility (information gain per cost) of all possible 16 actions. Figure 4 plots these expected utilities. In this case an intervention on node A promises to yield the highest amount of information per cost. Figure 5 shows the following rounds of the sequential

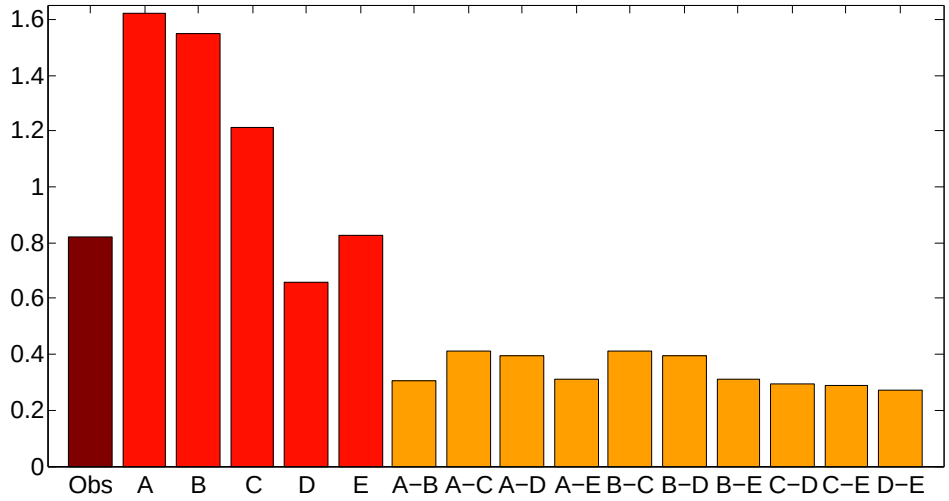

Figure 4: The expected utility (Equation (3)) of choosing one of the 16 possible next actions after an initial round of observations. The brown bar represents the expected utility of choosing additional observations. The red bars correspond to experiments with a perfect intervention on the specified node, while the orange bars show the expected utility of asking the expert about the state of one specific edge. In this case an experiment with intervention on variable A promises the highest payoff and is thus chosen as the next action.

learning procedure. We can see that the intervention on A removed close to all uncertainty about the edges connecting node A to B and C. This informs the action selection at the next step. For example the expected utility of asking the expert about edges $A-B$ or $A-C$ is now very low (as one would expect). For this particular example and choice of parameters and utility function it turned out to be optimal to only ask the expert towards the end, when little uncertainty remained.

We believe that this illustrative example already enjoys the benefits of using information from experimental data to inform the choice of questions posed to an expert. Treating belief elicitation and experimental design as a joint decision problem allows making more efficient use of the given resources (both the expert's and the experimenter's time and other associated costs).

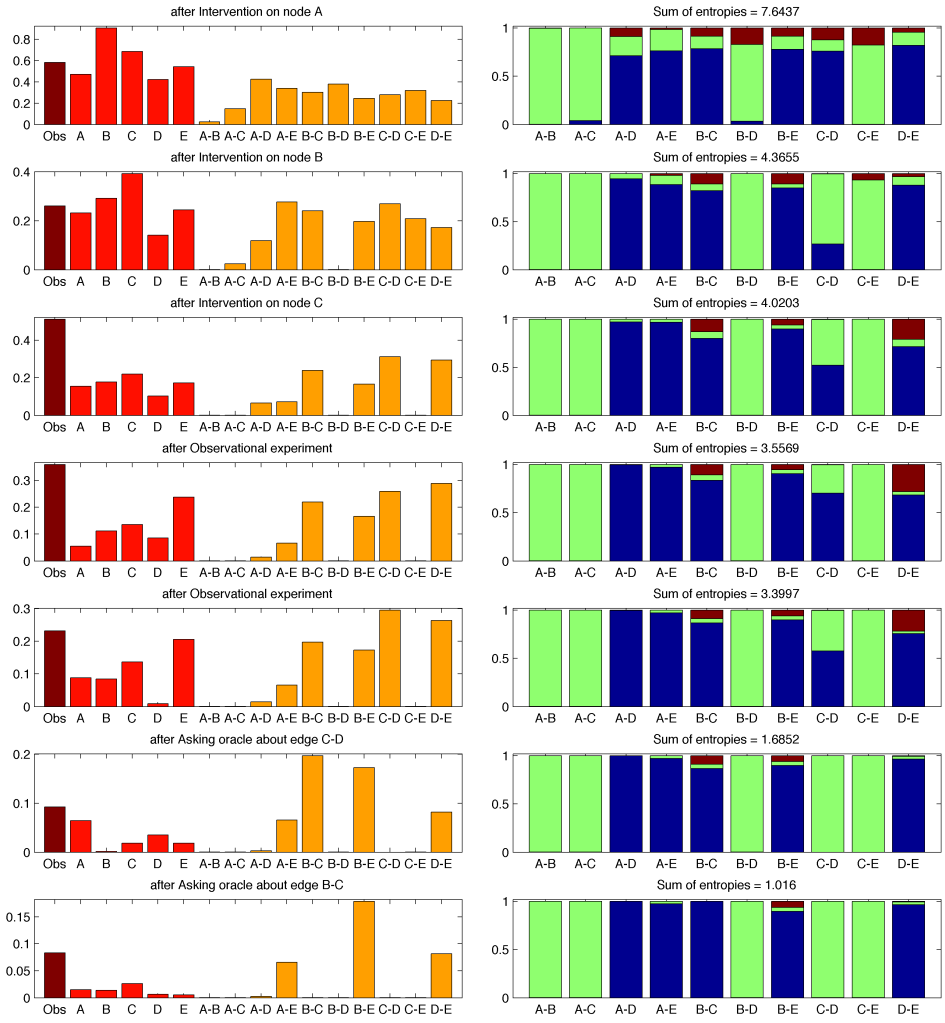

Figure 5: Marginal distributions of edge states (right column) and the expected utilities of the next experiments/question (left column) for rounds 2 to 8 of the sequential learning procedure. Please refer to the legend and captions in Figures 3 and 4 for explanations of the graphs. The titles of the plots on the left indicate the experiment or question chosen at the previous step.

3 Conclusions

To the best of our knowledge the combination of prior elicitation and experimental design into a joint sequential learning task has not been proposed

in the literature before. Even though in many applications the separation of prior elicitation and experimentation/data collection is likely inevitable due to practical constraints.

We plan to further explore the behavior of the approach for different utility functions. For example one might want to simply answer the question whether there exists any directed path going from node B to E. One would expect that in this case the optimal first action might be to ask the expert about the edge $B - E$ or to conduct an experiment with intervention on B.

Finally, the extension to continuous variables is also worth pursuing. The approaches discussed in Kueck et al. (2006); Myung and Pitt are likely to bear fruit in this context.

References

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley Series in Applied Prob. and Stats., 1994.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Daniel Eaton and Kevin Murphy. Bayesian structure learning using dynamic programming and MCMC. In *Uncertainty in Artificial Intelligence (UAI)*, 2007a.

Daniel Eaton and Kevin Murphy. Exact Bayesian structure learning from uncertain interventions. In *Artificial Intelligence and Statistics (AIStats)*, 2007b.

N. Friedman, K. Murphy, and S. Russell. Learning the structure of dynamic probabilistic networks. In *Proc. Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI'98)*, pages 139–147, 1998.

P. H Garthwaite, J. B. Kadane, and A O'Hagan. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100:680–701, 2005.

A Gelman, J B Carlin, H S Stern, and D B Rubin. *Bayesian Data Analysis*. Chapman and Hall, 1995.

T. L. Griffiths and J. B. Tenenbaum. Optimal predictions in everyday cognition. *Psychological Science*, 17:767–773, 2006.

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.

R. V. Ramamoorthi J. K. Ghosh. *Bayesian nonparametrics*. Springer, 2003.

D. Jenkinson. The elicitation of probabilities: A review of the statistical literature. BEEP working paper, Univ. Sheffield, 2005.

D. Kahneman, P. Slovic, and A. Tversky. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, 1982.

H. Kueck, N. de Freitas, and A. Doucet. SMC samplers for Bayesian optimal nonlinear design. In *Nonlinear Statistical Signal Processing Workshop (NSSPW)*, 2006.

D. J. C. Mackay. *Information Theory, Inference, and Learning Algorithms*. Cambridge U. Press, 2006.

J. I. Myung and M. A. Pitt. Optimal experimental design for model discrimination. *under review*.

A. O'Hagan and J.E. Oakley. Probability is perfect, but we can't elicit it perfectly. *Reliability Engineering and System Safety*, 85(1-3):239–248, 2004.

A. O'Hagan, Caitlin E Buck, and Alireza Daneshkhah. *Uncertain Judgements: Eliciting Experts' Probabilities*. Wiley, 2006.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

C.P. Robert. *The Bayesian Choice*. Springer, 2007.

R.W. Robinson. Counting labeled acyclic digraphs. *New Directions in the Theory of Graphs*, pages 239–273, 1973.

S.J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, 2002.

P. Shafto, C. Kemp, E. B. Bonawitz, J. D. Coley, and J. B. Tenenbaum. Inductive reasoning about causally transmitted properties. *Cognition*, 109:175–192, 2008.

Joshua B. Tenenbaum, Thomas L. Griffiths, and Charles Kemp. Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7):309 – 318, 2006.

J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behaviour*. Princeton U. P., 1947.

A. Yuille and D. Kersten. Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, 10(7):301 – 308, 2006. Special issue: Probabilistic models of cognition.