Ratio-Scale Elicitation of Degrees of Support

Shimon Schocken

Leonard N. Stern School of Business
New York University

44 West 4th Street, 9-80
New York, NY 10012-1126
(212) 998-0841 (tel)
(212) 995-4228 (fax)
schocken@stern.nyu.edu

Working Paper Series
IS-93-29

Ratio-Scale Elicitation of Subjective Degrees of Support

Abstract: During the last decade, the computational paradigms known as *influence diagrams* and *belief networks* have become to dominate the diagnostic expert systems field. Using elaborate collections of nodes and arcs, these representations describe how propositions of interest interact with each other through a variety of causal and predictive links. The links are parameterized with inexact degrees of support, typically expressed as subjective conditional probabilities or likelihood ratios. To date, most of the research in this area has focused on developing efficient belief-revision calculi to support decision making under uncertainty. Taking a different perspective, this paper focuses on the *inputs* of these calculi, i.e. on the human-supplied degrees of support which provide the currency of the belief revision process. Traditional methods for eliciting subjective probability functions are of little use in rule-based settings, where propositions of interest represent causally related and mostly discrete random variables. We describe ratio-scale and graphical methods for (i) eliciting degrees of support from human experts in a credible manner, and (ii) transforming them into the conditional probabilities and likelihood-ratios required by standard belief revision algorithms. As a secondary contribution, the paper offers a new graphical justification to eigenvector techniques for smoothing subjective answers to pair-wise elicitation questions.

1 Introduction

Diagnostic expert systems are designed to support inference tasks that are inherently inexact: medical diagnosis, mechanical fault-detection, classification of fuzzy sonar images, and other problems that involve a mixture of normative decision models and informal human judgement. In these systems, inference is carried out by traversing chains of rules that link a set of prospective hypotheses H (e.g. copier machine malfunctions) to a set of pieces of evidence, or observations, E (e.g. error messages) through several layers of intermediate propositions (e.g. paper jams and other disorders). Normally, the propositions are arranged as a *belief network* or an *influence diagram*, as we treat later in the paper. When a body of evidence $E' \subseteq E$ is known to obtain, the system's "inference engine" attempts to discern a subset of hypotheses $H' \subseteq H$ which seems to provide the best explanation to E'. Typically, though, the rules that link H and E are non-categorical, describing causal, diagnostic, or simply correlated relationships. For example, consider the following reasoning chain, taken from a medical diagnosis example: the habit of smoking (a disposition) increases the likelihood of a coronary heart disease (an hypothesis), which, in turn, is sometimes manifested through a swollen ankles symptom (an observation). Note that even though this line of reasoning is plausible, it is not categorical; many smokers will not develop heart problems, and swollen ankles is not a unique manifestation of a heart disease. Hence, although causal information is generally useful, any inference drawn from it must be qualified by the impreciseness of the underlying rules and the uncertainty associated with the available observations.

The rule-bases of early diagnostic expert systems, most notably Mycin and Prospector (Duda & Shortliffe, 1981), were constructed as *forward reasoning* architectures. Specifically, the rules followed the pattern of "*IF a particular observation* e_i *is present, THEN conclude hypothesis* h_j *with degree of support* $d(h_j|e_i)$," where the latter parameter was typically implemented as a certainty factor (Buchanan and Shortliffe, 1984). The present generation of diagnostic systems is quite different in two important respects. First, following results from cognitive psychology and descriptive decision theory, AI researchers seem to agree that *backward reasoning* – from hypotheses to evidence – is a far more credible elicitation technique than forward reasoning – from evidence to hypotheses (Shachter & Heckerman, 1987). Today, instead of trying to replicate in machine form the cognitive biases that characterize the reasoning of human diagnosticians, knowledge engineers attempt to uncover and then simulate the "physical" process through which prospective hypotheses

cause the manifestation of various observations. The basic argument is that whereas human experts are quite good at suggesting causal relationships in their respective domains of expertise, they are not nearly as good in performing diagnosis – a cognitive process which is mired by such biases as illusionary correlation, confirmation, representativeness, and availability (Einhorn & Hogarth, 1987, Kahneman, Slovic, and Tversky, 1982).

The second important characteristic of contemporary diagnostic systems is that they no longer rely on quasi-probabilistic and ad-hoc belief revision calculi. Instead, most of today's systems are inherently probabilistic – using subjective probabilities and Bayesian algorithms to represent and combine, respectively, the degrees of support the parameterize the system's rules. The renewed interest in Bayesian methods for uncertainty management in AI systems has led to the development of several computational architectures that are consistent with probability theory. Today, the two leading architectures in this field are *influence diagrams* (Howard & Matheson, 1981) and *Bayes networks*, also known as *belief networks* (Pearl, 1986).

Formally, a belief network is an acyclic, directed graph, consisting of propositional nodes and dependency arcs. The network has a dual logical/probabilistic interpretation, as follows. From a logical perspective, the arc $x \xrightarrow{d} y$ is normally interpreted as x *causes* y or x *explains* y. From a probabilistic standpoint, the nodes x and y are viewed as discrete random variables, and the arc $x \xrightarrow{d} y$ codes that y is conditionally dependant on x. The 'strength' of this association is modeled through the conditional probability $P(y|x)$, which is bound to the arc's label d. Since its inception about ten years ago, the belief network paradigm was implemented in many areas, ranging from sleep disorder analysis to gas turbine diagnosis to oil price forecasting. Perhaps the largest belief network implementation to date has been the QMR (Quick Medical Reference) system (Shwe et al, 1991), an excerpt of which is depicted in figure 1. Developed by researchers at the University of Pittsburgh, Carnegie-Mellon University, and Stanford University, QMR encodes textbook and human-supplied knowledge about 600 diseases, 4,000 observations (dispositions, symptoms, lab results, and patient data), and 40,000 links between them.

Automatic reasoning in a belief network begins by clamping a subset of nodes to observed values, and then letting a belief revision algorithm propagate their impact on other propositions. In the process, the system computes the posterior probabilities of certain propositions that are interpreted as hypotheses. The values of these revised "beliefs" are then used to direct the system's inference engine to pursue additional information (through consulta-

tion with the user) about promising hypotheses. Although the general task of computing posterior probabilities in an arbitrary belief network is *NP*-hard (Cooper, 1987), some network topologies lend themselves to efficient belief revision algorithms. For example, in the case of singly-connected networks (polytrees), there exists belief revision algorithms whose run-time is polynomial with the network's size (Pearl, 1986).

Indeed, most of the research on belief networks to date has focused on developing normative and heuristic belief revision calculi for a variety of different network topologies. The consistency and validity of the network's *inputs* – the human-supplied degrees of support that parameterize the network's edges – has received little attention in the AI literature. The research reported in this paper is an attempt to fill in this void by drawing and integrating relevant results from decision theory and cognitive psychology. In particular, we present a new elicitation procedure that enables human experts to uncover subjective degrees of support using graphical and comparative terms. The plan of the paper is as follows. Section 2 gives a formal description of the elicitation problem. Section 3 presents three independent ideas that can be used to promote the validity and consistency of human-supplied degrees of support. This material sets the stage for sections 4 and 5, which describe a general purpose elicitation procedure that can support the construction of influence diagrams and belief networks. Section 6 comments on applicability and future research issues.

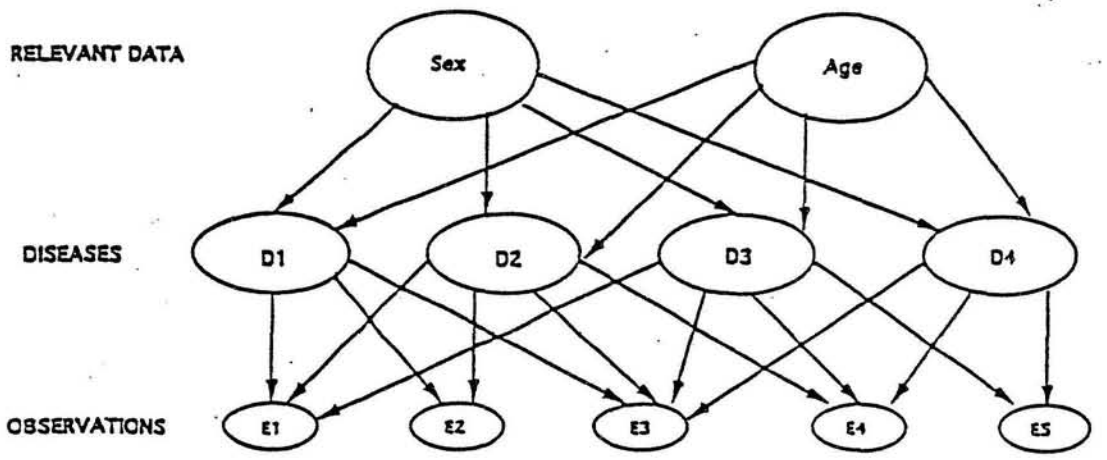

Figure 1: An excerpt from QMR-BN – one of the largest belief network implementations to date. A belief network is a directed acyclic graph in which nodes represent (mostly random) variables and links represent *causes* or *explains* relationships. In order to fully specify a belief network, each arc of the form $x \rightarrow y$ must be parameterized with the conditional probability $P(y|x)$. These probabilities are elicited through an exchange between a knowledge engineer and a domain expert.

2 The Problem

Consider the left side of figure 2, which describes a many-to-many relationship between two sets of propositions labeled $H = \{h_1, \ldots, h_m\}$ and $E = \{e_1, \ldots, e_n\}$. For convenience, we refer to elements of H and E as *hypotheses* and *observations*, respectively. Let us suppose that the goal of the elicitation procedure is to obtain the conditional probabilities that characterize *all* the possible rules that relate every single proposition in E to every single proposition in H, and vice versa. That is, for each observation e_i and hypothesis h_j, we wish to estimate both $P(e_i|h_j)$ and $P(h_j|e_i)$. For the sake of brevity, we denote the collection of these probabilities P, and the triplet $< E, H, P >$ a *model*. Our objective is to describe an elicitation procedure that, given E and H, helps a knowledge engineer elicit P from a human expert in a credible way.

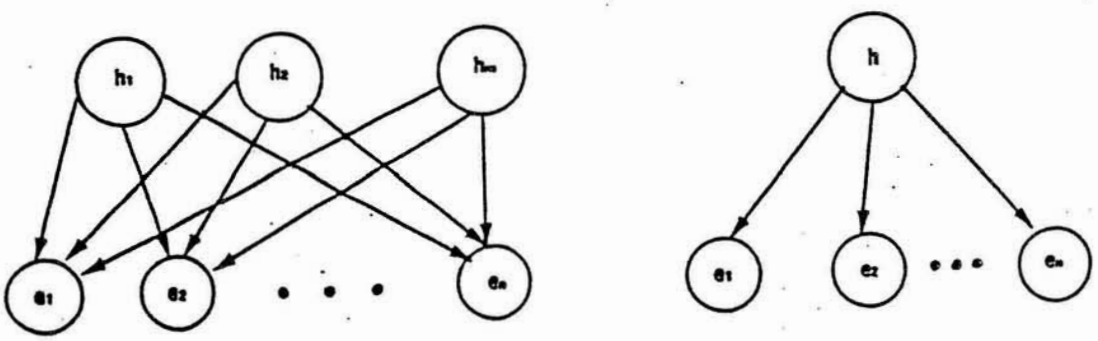

Figure 2: A canonical many-to-many model $< E, H, P >$ (left side) and a canonical one-to-many model $< E, h, P >$ (right side). In both cases, the h's are interpreted as prospective hypotheses and the e's as observations that are likely to be "caused" by the hypotheses, according to the expert whose knowledge the network represents. A key idea in modern diagnostic systems is to elicit and construct networks in a forward fashion (from hypotheses to observations) and then use Bayesian techniques to carry out backward, or "abductive" reasoning (from available observations to prospective hypotheses).

Our line of attack consists of two stages. First, we describe an elicitation procedure for models of the form $< E, h, P >$, i.e. models that describe a one-to-many relationship between a single hypothesis and a set of relevant observations (see right side of figure 2). Next, we extend the method to deal with many-to-many models. It's important to note in passing that any belief network can be seen as a modular collection of interacting $< E, H, P >$ models. For example, the network depicted in figure 1 consists of two layers, each being an instance of our notion of a many-to-many *model*. Thus, if the elicitation problem will receive a satisfactory solution at the single model's level, the same solution can be applied et cetera for all the models that make up a given network.

For simplicity, we assume for now that all the propositions in question are two-valued, and we denote the assertions q *is true* and q *is false* by q and $\overline{q}$, respectively. Focusing on the right side of figure 2, we interpret h as a possible cause of the e_i's, and we take the model to represent a set of inexact rules of the form "$h \rightarrow e_i$ *with degree of support* $d(e_i|h)$," $i = 1, \ldots, m$. Further, we encode $d(e_i|h)$ as the conditional likelihood-ratio $d(e_i|h) = P(e_i|h)/P(e_i|\overline{h})$. Importantly, our choice of degrees of support is free of any semantic interpretation. That is, it simply states that the odds of observing e_i when h obtains are $P(e_i|h)/P(e_i|\overline{h})$, irrespective of whether the relationship between the two propositions is causal, diagnostic, or simply correlational.[1]

If the conditional likelihood-ratios were credibly available for all the observations in question, a belief revision procedure could be used to compute their combined impact on the h hypothesis. To illustrate, suppose that a subset of observations were known to obtain. Without loss of generality, and in order to avoid reindexing, we denote this subset $E' = \{e_1, \ldots e_n\}$. If the prior odds favoring h on $\overline{h}$ were known to be $P(h)/P(\overline{h})$, the posterior odds in light of the body of evidence E' could be computed according to the ratio-form version of Bayes rule, as follows:

$$\frac{P(h|e_1, \ldots, e_n)}{P(\overline{h}|e_1, \ldots, e_n)} = \frac{P(e_1, \ldots, e_n|h)}{P(e_1, \ldots, e_n|\overline{h})} \cdot \frac{P(h)}{P(\overline{h})} \qquad (1)$$

Further, if the n observations are assumed to be ratio-independent with respect to the proposition h – an assumption that we treat later in the paper – the computation could be

[1] To avoid clutter, we use q to refer both to the uninstantiated proposition q as well as to the assertion q *is true*. The distinction between the two references will be clear from the context of the sentence.

factored into:

$$\frac{P(h|e_1,\ldots,e_n)}{P(\overline{h}|e_1,\ldots,e_n)} = \frac{P(e_1|h)}{P(e_1|\overline{h})} \cdot \ldots \cdot \frac{P(e_n|h)}{P(e_n|\overline{h})} \cdot \frac{P(h)}{P(\overline{h})} \tag{2}$$

In the decision theory literature, formula (2) is sometimes referred to as the *Bayesian belief revision* process – a "rational" prescription for revising one's belief in a pair of propositions when new and relevant evidence is brought to bear. In addition for being a normative belief revision prescription that can be easily derived from the axioms of subjective probability, formula (2) has certain extra-probabilistic and qualitative properties that go beyond mathematical details. For example, human diagnosticians are prone to many evidence presentation and clustering biases, such as primacy, recency, salience, and bundling effects (Fischhoff & Beyth-Marom, 1983). In contrast, formula (2) is commutative and associative, and therefore it is insensitive to the order and packaging in which the evidence unfolds. Therefore, if we let E' stand for the currently available body of evidence, formula (2) can be used to compute the current odds $P(h|E')/P(\overline{h}|E')$ recursively, as follows. First, in the absence of any relevant evidence, which we denote by $E' = \emptyset$, the current odds are initialized to the prior odds $P(h)/P(\overline{h})$. When the truth value of a certain observation e_i becomes available, the current odds are revised, or updated, through the step-wise formula:

$$\frac{P(h|E' \cap e_k)}{P(\overline{h}|E' \cap e_i)} = \frac{P(e_i|h)}{P(e_i|\overline{h})} \cdot \frac{P(h|E')}{P(\overline{h}|E')} \tag{3}$$

And E' becomes $E' \cap e_k$. As more observations become available, the current odds are updated in a similar fashion. Alan Turing, who had a side interest in belief revision models, called the ratio $P(e_i|h)/P(e_i|\overline{h})$ the "weight of evidence carried by e_i to the assertion h is more likely than $\overline{h}$"(Good, 1950). Note that the neutral element in this multiplicative calculus is the observation e_j for which $P(e_j|h) = P(e_j|\overline{h})$. This observation provides no "added value" in terms of discriminating between h or $\overline{h}$.

Suppose now that some version of formula (3) were embodied in a diagnostic system designed to carry out rule-based reasoning under uncertainty. In order to construct such a system, a knowledge engineer would have to elicit, for each rule of the form $h \rightarrow e_i$, a degree of support of the form $\frac{P(e_i|h)}{P(e_i|\overline{h})}$. How can such numbers be elicited from human experts in

a credible way? The next section presents three independent techniques that address this challenge.

3 Elicitation Support Techniques

To reiterate, we consider a two-valued hypothesis h and n rules of the form $h \rightarrow e_i$, $i = 1, \ldots, n$. For now, we take the goal of the elicitation procedure to acquire, through interactions with a human expert, a set of likelihood-ratios of the form $\frac{P(e_i|h)}{P(e_i|\overline{h})}$ $i = 1, \ldots, n$. We propose three steps that can be used to promote the validity and efficiency of this task:

- To minimize the use of numeric guestimates, we describe a graphical elicitation interface that combines the logical and probabilistic backdrops of the inference model in one representation.
- To help experts overcome certain estimation biases, we propose to use an isotropic elicitation procedure in which questions can be turned around, allowing reasoning from hypotheses to evidence and vice versa.
- To minimize cognitive strain and to promote consistency, we describe an elicitation modality in which the expert is asked to *compare*, rather than *specify*, the evidential impacts of various propositions.

Although the above three points are independent of each other, implementing them within the same elicitation procedure can lead to synergistic results. The remainder of this section discusses the three points in detail, as they unfold in the context of an illustrative diagnostic problem.

A graphical user-interface: Consider a population of "cases" partitioned into cases which are characterized by the condition h and cases which are characterized by the complementary condition $\overline{h}$. Each case can be subjected to a series of individual "tests" $e_i, i = 1, \ldots, n$ that come up either positive or negative. The prevalences of each positive test result in the population are given by the parameters $P(e_i|h)$ and $P(e_i|\overline{h})$. Given the above nomenclature, what should be the evidential impact of learning that a certain test

comes up positive? An inspection of the belief revision process (3) reveals that the admittance of a new piece of evidence e_i can either increase, decrease, or leave the same, the current belief in h. More specifically, we have five prototypical cases, as follows:

$$\text{impact of } e_i \text{ on } h \;=\; \begin{cases} e_i \text{ confirms h} & \text{if} \quad P(e_i|h)/P(e_i|\overline{h}) \;\to \infty \\ e_i \text{ supports h} & \text{if} \quad P(e_i|h)/P(e_i|\overline{h}) \;> 1 \\ e_i \text{ is irrelevant to h} & \text{if} \quad P(e_i|h)/P(e_i|\overline{h}) \;= 1 \\ e_i \text{ supports } \overline{h} & \text{if} \quad P(e_i|h)/P(e_i|\overline{h}) \;< 1 \\ e_i \text{ confirms } \overline{h} & \text{if} \quad P(e_i|h)/P(e_i|\overline{h}) \;\to -\infty \end{cases} \tag{4}$$

Said otherwise, the conditional likelihood ratios $P(e_i|h)/P(e_i|\overline{h})$ form the currency of the belief revision process. One way to "determine" the values of these ratios is to ask a domain expert to estimate them directly. Unfortunately, there is no evidence that human beings – whether laymen or experts – are capable of translating implicit degrees of support into a numeric $[-\infty, \infty]$ scale whose neutral point is 1. Therefore, it seems prudent to seek alternative means to express primitive beliefs. We propose a computer graphics technique that can be used to elicit degrees of support from humans *indirectly*, via a certain representation that we call d-graphs (d for "diagnostic").

The notion of d-graphs is illustrated in figure 3, which depicts the evidential impacts of three independent observations on the same hypothesis. In each graph, the exterior left (right) bar represents the subset of the population that has (does not have) the condition h. The interior left (right) bar represents the h cases (cases who have no h) for whom e_i is also known to be true, $i = 1, 2, 3$. For example, let h be a *prostate cancer* condition. Figure 3 (left) describes the diagnostic impact of a supportive observation, e.g. e_1 = *urinary disorders*. Figure 3 (middle) describes the impact of a confirmatory test, e.g. e = *positive biopsy*. Figure 3 (right) describes the problematic nature of a non-categorical test, e.g. $4 \leq \text{PSA} \leq 20$. The clinical characteristics of the Prostate Specific Antigen blood test are such that rates above 20 and below 4 are indicative of h and $\overline{h}$, respectively. "Grey" PSA values in the interval $[4, 20]$ are inconclusive. The test is controversial because (i) many middle-aged males who take it score in the grey area, and (ii) even though grey scores are prevalent in healthy as well as in predisposed patients, they tend to cause a great deal of concern to the tested individual, to the extent that some urologists recommend not to take the test unless other evidence suggests that the patient is predisposed. We speculate that

this anxiety and confusion could be alleviated if the clinical characteristics of the PSA test were presented to the individual using a d-graph like representation.

In general, d-graphs can play a role in (i) eliciting degrees of support from human experts, and in (ii) explaining the results of a diagnostic process to consulting users. Because of its dual format, the d-graph forces the expert (or the user) to give equal and simultaneous consideration to both h and $\overline{h}$; in doing so, it serves to mitigate potential confirmation biases. According to the well-known Popperian principle, when one suspects that h is true, the most rational course of action is to try to falsify h, i.e. to seek evidence that supports $\overline{h}$ (Popper, 1950). Yet in reality, humans are known to behave in an opposite way: a suspected h typically leads to a judicious search for confirmatory clues (Hamilton, 1979), resulting with an undue suppression of the possibility $\overline{h}$. For example, in the process of building "forward reasoning" rule based systems, experts are routinely asked to specify numbers that describe the extent to which various clues support various conclusions like h *is true*. In the context of a d-graph, this amounts to eliciting only half of the picture. If the system will not be fed with the degree to which the same observations support $\overline{h}$, the posterior belief in h will be overestimated during consultations. Since the posterior beliefs are normally used to guide the inference engine to promising directions, the system will attempt to pursue reasoning chains that collect additional evidence about h. Ironically, the reasoning process might exhibit the very same confirmation bias that characterizes the human expert that the system is attempting to simulate.

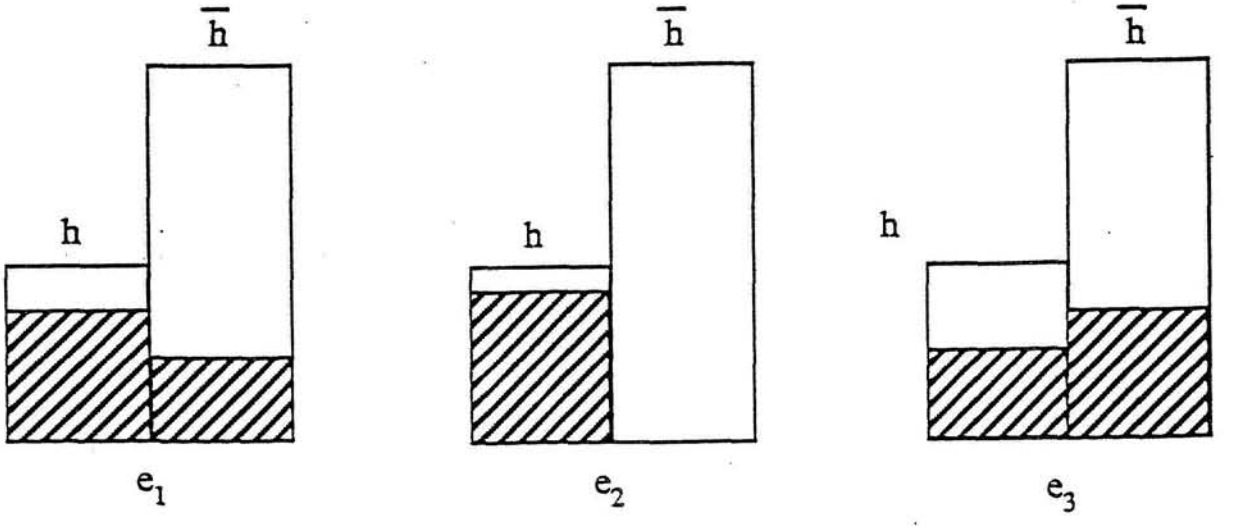

Figure 3: A series of *d-graphs*, describing the clinical characteristics of three different observations related to the same hypothesis: supportive evidence (left), confirmatory evidence (middle), and non-categorical evidence (right). Note that the transformation between the ith d-graph and the conditional likelihood ratio $P(e_i|h)/P(e_i|\overline{h})$ is straightforward.

In addition to their balanced format, d-graphs are explicit about the prior likelihood of the background hypothesis, which is the ratio between the areas of the two exterior bars. A well documented bias – representativeness – is known to systematically cause human experts to ignore or discount base rate probabilities and to overestimate the clinical impact of inexact evidence, especially if the evidence is more salient, recent, or interesting, than the background information (Tversky & Khaneman, 1974). The spatial arrangement of a d-graph can effectively debias this tendency, since the background information (exterior bars) is as vivid and explicit as the clinical information (interior bars). The d-graph representation packs all this information into one representation, along with other characteristics such as hit-rates, false alarm rates, and type II errors (the part of the left exterior bar which is not covered by the left interior bar).

In sum, we believe that d-graphs can play an important role both in building and in explaining diagnostic expert systems. First, we foresee a graphical elicitation interface that enables experts to *express* implicit degrees of supports by performing spatial manipulations on d-graphs. Second, we foresee a user interface that employs d-graphs to *explain* the clinical characteristics of various tests (along with their costs) and the process through which newly admitted evidence causes the system to update its beliefs in competing hypotheses.

Bi-directional elicitation procedures: When eliciting inexact inferential relationships, some questions can be easier to answer if you turn them around. To illustrate, suppose that e and h stand for the propositions *the patient smokes* and *the patient is predisposed to a certain heart condition*, respectively. Which subjective probability is more credibly available from a human expert, $P(h|e)$ or $P(e|h)$? Several authors addressed this question from a cognitive perspective, suggesting that "backward reasoning" and "thinking forward in reverse" (from hypotheses to evidence) is generally a more effective elicitation modality (Shachter & Heckerman, 1987, Einhorn & Hogarth, 1987). Although we agree with this recommendation when everything else is held equal, we postulate that the direction of the elicitation should also depend on the clinical experience of the expert, and, in particular, on his ability to retrieve relevant examples from the e and h populations.

If the expert is a general physician who knows relatively more smokers (e) than patients with heart diseases (h), it is probably safer to use the smokers population as a reference group and go on to assess $P(h|e)$. If a rigid elicitation language will force this expert to reason "forward," he will have to resort (in his mind) to the small conditioning sample h, thus yielding an unreliable estimate of $P(e|h)$. The situation would be quite different if the

expert were a cardiologist. in that case, the expert would probably find it easier to assess $P(e|h)$, due to the large sample of people with heart diseases that he can retrieve from his clinical work experience.

We see that the validity of elicited degrees of support is influenced by two factors: (i) availability, and (ii) the "law of small numbers" (Tversky and Kahneman, 1974). The availability heuristic leads the expert to focus on the sample of cases which is more salient or vivid in his mind. This heuristic would be beneficial only if it coincides with a large sample of cases. If the selected background sample is small, the literature indicates that many experts will still be willing to use it as a characteristic image of a much larger population – a manifestation of the "law of small numbers" bias. To correct this fallacy, the expert should be encouraged to retrieve as many examples as possible from both populations, using his clinical experience as well as organizational memories and relevant case histories. The larger sample should then be selected as the conditioning assumption.

Hence, we propose to employ an isotropic elicitation modality that enables knowledge engineers and domain experts to reason about the same inference problem using either diagnostic (from evidence to hypotheses) or clinical (from hypotheses to evidence) modes of reasoning. It is here where the reliance on Bayesian inference methods is particularly useful: unlike uni-directional belief calculi, such as the certainty factors model (Buchanan and Shortliffe, 1984) and the theory of evidence (Shafer, 1987), Bayesian methods allow the knowledge engineer to invert and validate elicited probabilities using a normative framework. We'll return to this point later in the paper, when we discuss the notion of a two-way elicitation procedure.

Relative versus absolute elicitation questions: A substantial body of psychometric evidence indicates that when expressing physical stimuli like brightness, weight, and distance, humans find it easier to use ordinal, rather than cardinal, scales of measurement (Stevens, 1959, Stevens & Galanter, 1964, Krantz, 1972). In a typical experiment, most subjects displayed considerable errors when asked to specify the aerial distances between different cities. At the same time, the subjects were quite good at providing answers such as "Cairo is about twice as far away from New York as London is." In general, the evidence suggests that people find it easier to gauge physical quantities (as well as priorities) using relative and pair-wise judgements. This observation is the hallmark of the *analytic hierarchy process* (AHP) technique for eliciting preferences over multi-attribute alternatives (Saaty, 1980).

There are two reasons why Saaty's technique could lend itself to eliciting conditional probabilities and likelihood ratios as well. From an analytic standpoint, the elicitation of probabilities is very similar to the elicitation of normalized preferences over uncertain outcomes (Savage, 1954). Since the advantages and limitations of Saaty's method for the latter are well understood, it seems reasonable, in light of Savage's insight, to try to apply them for the former as well. Further, it is unlikely that people are capable of encoding primitive beliefs in various propositions using a closed-form, cardinal scale. It seems more prudent to attempt to elicit implicit beliefs indirectly, in a "revealed" fashion, using pair-wise and perhaps verbal comparisons.

To illustrate the notion of ratio-scale pairwise comparisons, consider an hypothesis h and three relevant observations e_1, e_2, and e_3. First, we ask the expert to assume that h is true. Next, we ask him to compare the likelihoods of the three potential observations in light of that information. Suppose that the expert responds that in light of h, e_1 is about twice as likely as e_2 and about three times as likely as e_3. Further, the expert says that e_2 is about twice as likely as e_3. Before we go on to describe how this information should be processed, we point out that the expert's responses are inconsistent: if e_1 is twice as likely as e_2 and three times as likely as e_3, e_2 must be 1.5 as likely as e_3.

The example illustrates the impreciseness that is likely to mire any subjective, or human-supplied, set of inputs. In this particular case, the inconsistency could be resolved in a number of different ways, e.g. by setting the third human-supplied ratio to 1.5. However, such a "correction" would amount to tinkering with the genuine data provided by the expert. If the expert is a qualified specialist, his inputs must be treated with great care; Further, there is absolutely no way to pinpoint the culprit of the inconsistency, which may well be any one of the expert's answers (or combinations thereof).

There is a "positive" way to think about this inconsistency, though. First, the inconsistency is inevitable. Second, because of the algebraic properties of the elicited inputs, the inconsistency produces a *variance* of subjective opinions about each one of the desired ratios, and these opinions can perhaps be "pooled" into some sort of a "mean opinion." For example, the value of $P(e_3|h)/P(e_1|h)$ can be derived by (1) asking the human to specify it directly, (2) inverting the human-supplied value of $P(e_1|h)/P(e_3|h)$, and (3) multiplying the human-supplied values $P(e_3|h)/P(e_2|h)$ and $P(e_2|h)/P(e_1|h)$. If the expert were a perfect estimator, the three "opinions" would be identical. What should we do when they are not?

This is a general problem that goes beyond the specialized context of this paper. The problem arises in any elicitation procedure that asks the same question several times, either explicitly or implicitly. In such cases, the expert's answers can be viewed as noisy observations that are dispersed around a certain subjective mean. Indeed, several authors proposed to use regressive methods to smooth such inputs, e.g. logarithmic least squares (De Graan, 1980), and ridge regression (Harker & Vargas, 1987). However, when the expert's answers form a square and reciprocal matrix, as we will see shortly, the most robust smoothing method is the normalized eigenvector technique proposed by Saaty. The next section justifies the technique and adapts it to our probabilistic elicitation context.

4 One-Way Elicitation

The one-way elicitation problem is defined as follows. Given a single hypothesis h and a series of inexact rules of the form $h \rightarrow e_i$, $i = 1, \ldots, n$, the goal is to obtain a credible estimate of the degrees of support that parameterize these rules, which we denote by the vector $P = (P(e_1|h), \ldots, P(e_n|h))$ (the task of obtaining the vector $\overline{P} = (P(e_1|\overline{h}), \ldots, P(e_n|\overline{h}))$ follows exactly the same procedure which we now turn to describe). The choice of a vector notation reflects our *relative* approach to the elicitation problem: instead of asking the expert to specify $P(e_i|h)$ and $P(e_j|h)$ directly, as in commonly done in rule-based settings, we ask him to estimate the extent to which e_i is *more likely* to be observed than e_j when h obtains, $1 \leq i < j \leq n$. To illustrate, consider the following excerpt from such an elicitation scenario:

> Suppose that a patient has a certain heart condition, denoted h, and consider the following potential observations:
>
> e_1: the patient has a chest ache radiating to the left arm
>
> e_2: The patient has swollen ankles
>
> In your opinion, which observation is more likely in light of h?
>
> (*let us assume that the expert answered* e_1)

> To what extent is e_1 more likely than e_2? If you think that they are equally likely, enter the number 1. If you think that e_1 is twice as likely as e_2, enter the number 2. Feel free to enter any number greater than or equal to 1, using your clinical judgement and work experience.

With n observations, the expert is required to answer no more than $\frac{1}{2} \cdot n \cdot (n-1)$ such questions (the rationale for this upper bound will be explained shortly). Note that before the conditional likelihood ratio $P(e_i|h)/P(e_j|h)$ is elicited, the expert is asked which of the two observations is more likely in light of h. With n observations, the elicitation procedure will begin by asking the expert to rank-order the observations in terms of their (perceived) decreasing likelihoods in light of h. Without loss of generality, we can use the expert's ranking to reindex the observations, so that e_1 and e_n represent the most and the least likely observation, respectively, in light of h. As a result, $(e_1, \ldots, e_n)$ becomes an ordered set in which $P(e_i|h) \geq P(e_j|h)$ if $i < j$, according to the expert.

In what follows, we use the symbol P to stand for a subjective estimate of an unknown probability, which we denote P_0. We record the expert-supplied estimates in an $n \times n$ likelihood matrix, denoted A, in which $a_{i,j}$ represents $P(e_i|h)/P(e_j|h)$ – the expert's estimate of $P_0(e_i|h)/P_0(e_j|h)$. For example, suppose that the three rules $h \rightarrow e_i, \quad i = 1, 2, 3$ were parameterized by the following (objective) conditional probabilities[2]:

$$P_0 = (P_0(e_1|h), P_0(e_2|h), P_0(e_3|h)) = (0.8, 0.2, 0.1) \tag{5}$$

Had we had access to an "ideal expert" whose subjective judgement were perfectly calibrated with reality, we would obtain the A_0 matrix given in the left side of figure 4.

[2]All the vectors that are mentioned in this paper are column vectors. We use row notation as in (5) in order to conserve space.

$$\begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \end{array} \begin{array}{c} \begin{array}{ccc} e_1 & e_2 & e_3 \end{array} \\ \begin{pmatrix} 1 & 4 & 8 \\ \frac{1}{4} & 1 & 2 \\ \frac{1}{8} & \frac{1}{2} & 1 \end{pmatrix} \end{array} = A_0 \qquad \begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \end{array} \begin{array}{c} \begin{array}{ccc} e_1 & e_2 & e_3 \end{array} \\ \begin{pmatrix} 1 & 5 & 6 \\ \frac{1}{5} & 1 & 3 \\ \frac{1}{6} & \frac{1}{3} & 1 \end{pmatrix} \end{array} = A$$

Figure 4: A consistent (left) and inconsistent (right) likelihood matrices. Every likelihood matrix is square and reciprocal, with $a_{i,j} = 1/a_{j,i}$. A likelihood matrix is said to be *consistent* if and only if its column vectors are all proportional to each other. In this research, each likelihood matrix is constructed under a fixed hypothesis h.

In a likelihood matrix, each entry $a_{i,j}$ stands for a subjective (expert-supplied) ratio $P(e_i|h)/P(e_j|h)$. That is, the expert estimates that in light of h, e_i is $a_{i,j}$ more likely to be observed than e_j. For example, $a_{1,3} = \frac{P(e_1|h)}{P(e_3|h)} = \frac{0.8}{0.1} = 8$.

Likelihood matrices have three desirable properties that can be used to structure the elicitation process. First, since $a_{j,i} = 1/a_{i,j}$ throughout the matrix, only the entries above the diagonal can be elicited, giving a total of $\frac{1}{2} \cdot n \cdot (n-1)$ pair-wise comparison questions. Second, all the elicited entries must be greater than or equal to 1. Third, the relation $a_{i,j} < a_{i,k}$ must persist for all rows i and columns $j < k$. The latter two properties are a consequence of our preprocessing stage, in which the observations were reindexed according to their decreasing (perceived) likelihoods in light of h. Since the three properties constrain the inputs that go into the matrix, they can be used to test the expert's raw answers at the point of elicitation for first-order inconsistency violations.

Second order inconsistencies occur when the expert's answers induce an inconsistent matrix. Formally, we have the following definitions:

Definition 1: Two vectors U and V are said to be *proportional* if $U = \alpha \cdot V$ for some scalar α, which we call the proportionality constant.

Definition 2: Two vectors $U = (u_1, \ldots, u_n)$ and $V = (v_1, \ldots, v_n)$ are said to be *ratio-equivalent* if $u_i/u_j = v_i/v_j$ for all $1 \leq i, j, \leq n$.

Definition 3: A likelihood matrix is said to be *consistent* if all its columns are proportional to each other.

It is not difficult to show that two vectors are proportional if and only of they are ratio-equivalent. Hence, either definition 1 or 2 is theoretically unnecessary. Yet from a cognitive standpoint, proportionality is sometimes more salient than ratio-equivalence, and vice versa. Therefore, we will use both definitions interchangeably, according to the discussion's context.

Using definition 3, we see by inspection that A_0 is a consistent matrix. This is not surprising, because A_0 represents the opinions of an "ideal expert" whose estimates of the ratio-properties of P_0 are perfectly calibrated with reality. Since all the column vectors in A_0 are proportional to each other, any one of them can serve as a credible ratio-estimate of P_0. Yet in reality, P_0 is unknown ex-ante, and the very goal of the elicitation procedure is to estimate it from a set of *imperfect* human inputs. Since such inputs are bound to be biased, they induce an inconsistent likelihood matrix – a matrix which contains at least two disproportional columns. The source of the inconsistency can be traced to the fact that some of the answers to the $\frac{1}{2} \cdot n \cdot (n-1)$ questions presented to the expert exhibit what may be termed "multiplicative intransitivity." For example, suppose that three entries in a certain likelihood matrix A were such that:

$$a_{i,j} \cdot a_{j,k} \neq a_{i,k} \tag{6}$$

This, along with the semantics of the $a_{i,j}$'s, would imply the nonsensical relationship:

$$\frac{P(e_i|h)}{P(e_j|h)} \cdot \frac{P(e_j|h)}{P(e_k|h)} \neq \frac{P(e_i|h)}{P(e_k|h)}, \tag{7}$$

Clearly, such a result would indicate that one or more of the human-supplied $a_{i,j}$'s is off-target. There seem to be two ways to deal with the problem: the "revision method,"

and the "smoothing method." The revision method is based on going back to the expert and asking him to revise one or more of his original estimates in order to "correct" the inconsistency. This approach is deceptively simple, because it breaks down rather quickly when $n > 3$. Further, there reaches a point at which the expert will no longer be able or willing to tinker with his estimates, at which stage we would still have to deal with an inconsistent matrix. It is at that point that the second method – *smoothing* – enters the picture.

The smoothing method is based on the pragmatic premise that regardless of the efforts of the knowledge engineer and the domain expert, the likelihood matrix is bound to be inconsistent. Yet for the smoothing method, the inconsistency is not altogether bad. To explain this subtlety, we reiterate that the inconsistency stems from the fact that the elicitation procedure involves "too many" questions. Theoretically, if the goal is to elicit all the ratios of the form $P(e_i|h)/P(e_j|h)$, $1 \leq i, j \leq n$, it is sufficient to elicit the $n-1$ ratios $P(e_i|h)/P(e_{i+1}|h)$, $i = 1, \ldots, n-1$. Given this series, any desired ratio of the form $P(e_i|h)/P(e_j|h)$ can be computed through a product of some of the ratios in the series. Yet, going back to our example, in addition to asking the expert to specify $P(e_1|h)/P(e_2|h)$ and $P(e_2|h)/P(e_3|h)$, we have also asked him to specify $P(e_1|h)/P(e_3|h)$. Mathematically, one of these questions is redundant, because every one of the three ratios could be computed from the other two. By asking the expert to estimate *all* ratios directly, we open the door to inconsistency. In the general case, instead of asking the expert to specify a vector of n ratios directly, our series of pair-wise comparisons yields an $n \times n$ likelihood matrix. Taken together, each column of the matrix can be interpreted as a different ratio-scale estimate of the unknown vector P_0. The columns may be disproportional, but every one of them represents genuine information that must be taken into consideration, especially if the expert is highly qualified.

Said otherwise, the smoothing approach views the (inconsistent) likelihood matrix as a collection of inexact but nonetheless genuine opinions (column vectors) from which a composite ratio-scale estimate can be *synthesized* using a certain algebraic procedure. To illustrate, let us return to figure 4 and assume that the expert's above-diagonal judgement $a_{1,2}$, $a_{1,3}$, and $a_{2,3}$ were +25%, -25% and +50% off the mark compared to their A_0 counterparts. Such an expert would yield the inconsistent matrix A given in the right hand side of figure 4. Following standard methods from linear algebra, one way to "summarize" the column vectors of a square and reciprocal matrix is to compute the normalized eigenvector associated with the matrix's maximal eigen value. In the case of A, this eigenvector is:

$$W = (0.717, 0.195, 0.088) \tag{8}$$

At the end of the paper, we offer an appendix that shows that W is as good a ratio-estimate of P_0 as we can get from A. Now, if this assumption were valid, then had we had some ex-ante knowledge of one of the elementary probabilities in the objective P_0, we could compute the proportionality constant between W and P_0 and then use it to unfold the entire estimate of the latter. For example, suppose we knew that $P_0(e_2|h) = 0.2$. Using this information, we compute $\alpha = \frac{P(e_2|h)}{w_2} = \frac{0.2}{0.195}$ and proceed as follows:

$$\begin{aligned} P &= \tfrac{0.2}{0.195} \cdot W \\ &= (\tfrac{0.2}{0.195} \cdot 0.717,\ \tfrac{0.2}{0.195} \cdot 0.195,\ \tfrac{0.2}{0.195} \cdot 0.088) \\ &= (0.73,\ 0.2,\ 0.09) \end{aligned} \tag{9}$$

The difference between this result and the target vector $P_0 = (0.8, 0.2, 0.1)$ stems from the imperfect knowledge that A represents. However, considering the biasdeness of the expert's original answers, this result is surprisingly good, illustrating the robustness of the normalized eigenvector to data perturbations.

Of course, the technique's ability to construct P_0 hinges on a-priori knowledge of one of the desired probabilities. Anticipating that requirement, we can augment the initial set of observations $(e_1, \ldots, e_n)$ with an additional clue, say e^*, whose conditional probability $P_0(e^*|h)$ is known to the expert. For example, in the heart disease scenario, e^* can stand for the observation *the patient is a male*, the assumption being that the probability P_0(male|heart disease) is credibly available to the expert.

There exist situations, however, in which no prior knowledge will be available about any one of the constituent probabilities in P_0. In such a case, instead of estimating the vector $(P_0(e_1|h), \ldots, P_0(e_n|h))$, an extension of the technique described in this section can be used to estimate the likelihood-ratio vector $(P_0(e_1|h)/P_0(e_1|\overline{h}), \ldots, P_0(e_n|h)/P_0(e_n|\overline{h}))$. Methods for estimating such likelihood vectors for dichotomous and multi-valued propositions are described in the remainder of the paper. We conclude the present section with a comment on verbal and graphical elicitation techniques.

Non-numeric elicitation inputs: Up to this point, our approach to the elicitation problem was based on the assumption that humans are capable of describing likelihood-ratios using numbers. This controversial assumption has been challenged by many, not the least of them is H.R. Haldeman, President Nixon's Chief of Staff. Describing Kissinger's persistent concern about a Russian attack on China, Haldeman recalls how "I used to tease him about his use of percentages. He would say there was a 60% chance of a Soviet strike on China, for example, and I would say: why 60, Henry? Couldn't it be 65% or 58%? (Kotz and Stroup, 1983, p. 18). Clearly, Haldeman's point is well taken, and the credibility of any numeric measure of intuitive judgement should always be suspect for error. With that in mind, it is important to note that the eigenvector method yields a *proportional estimate* that is insensitive to the absolute magnitudes of the expert's inputs, so long as the inputs retain certain ratio-scale properties. This observation led Saaty to propose a 9-point scale of verbal assessments in which the value 1 corresponds (in our context) to the proposition "*given h, e_i is as likely as e_j*," and the value 9 corresponds to "*given h, e_i is absolutely more likely than e_j.*"

The choice of the 9-points scale of reference is not arbitrary, and can be justified on analytical and experimental grounds (Saaty, 1980). Further, Lichtenstein and Newman (1967) have shown empirically that verbal descriptions of uncertainty can be mapped quite effectively on ranges of probabilities. That said, it is important to note that nothing in the eigenvector method *requires* a 9-point scale. As Harker and Vargas (1987) pointed out, "One scale may be appropriate for one application and may not be appropriate for another ... a different scale could and should be chosen for each application." For example, in situations where little is known about a particular set of propositions, a 1 to 5 or even a 1 to 3 scale could be more credible than a 1 to 9 scale. Clearly, the freedom to modify the input scale or use more than one scale in the context of the same problem makes the elicitation task more flexible.

In addition to the numerical and verbal methods, the language of d-graphs (figure 3) can also serve as a vehicle for expressing primitive beliefs. Specifically, we foresee a system that presents the expert with of a series of d-graphs, one for each observation e_i, arranged on the same frame of reference (e.g. a single computer screen). Next, the expert is asked to simultaneously adjust the relative heights of the interior bars in all the d-graphs. The heights of the exterior bars are kept constant, since they reflect the prior odds on h, which is fixed across all the e_i's. When the expert signals that the data entry has been completed, the ratios among the bar heights that he has specified can be fed into a standard likelihood

matrix. Since the d-graphs are structured around the $h/\overline{h}$ dichotomy, the expert will end up providing all the inputs that are necessary to compute $P(e_1|h), \ldots, P(e_n|h)$ as well as $P(e_1|\overline{h}), \ldots, P(e_n|\overline{h})$ simultaneously. This will entail the manipulation of two independent likelihood matrices, but the details of the computations will be exactly the same in each case. Note in passing that the computations and the matrix notation are completely hidden from the expert, although the outputs of the computations can be displayed graphically using the d-graphs interface.

5 Two-way Elicitation

All the examples that we have discussed thus far were structured as $< E, h, P >$ models (right side of figure 2), where subsets of a series of n observations were used to update one's belief in a single dichotomous hypothesis. Needless to say, most interesting diagnostic problems involve multiple hypotheses (left side of figure 2). For example, in a lymph node pathology case, a human expert attempts to map a subset of n microscopic observations from a section of a lymph tissue (obtained through biopsy) onto a set of m classes of malignant lymphoma. As it turns out, this classification is one of the most difficult tasks of surgical pathology (Henrion, Breese, & Horvitz, 1991). Technically speaking, the problem is that every one of the hypotheses can manifest itself through overlapping combinations of observations, leading to an alarming number of misdiagnoses, especially by inexperienced pathologists (Velez-Garcia et al, 1983). Thus, the ability to augment the clinical findings of a practicing pathologist by computing their posterior beliefs (in light of different observations and based on the experience of expert pathologists) is an important decision support objective.

Formally, we consider a many-to-many model $< E, H, P >$ which describes a set of relationships between m hypotheses, labeled $H = \{h_1, \ldots, h_m\}$, and n related observations, labeled $E = \{e_1, \ldots, e_n\}$. For the sake of brevity, we introduce the following notation:

$$E_{i,j} = P(e_i|h_j) \tag{10}$$

$$H_{i,j} = P(h_i|e_j) \tag{11}$$

$$E_{i,j,k,l} = \frac{P(e_i|h_j)}{P(e_k|h_l)} \tag{12}$$

$$H_{i,j,k,l} = \frac{P(h_i|e_j)}{P(h_k|e_l)} \tag{13}$$

Let us suppose that, given a particular $n \cdot m$ model $< E, H, P >$ (where P is unknown), we are required to estimates *all* the conditional likelihood-ratios of types (12) and (13). The brute force approach, which involves asking a domain expert to specify all these ratios directly, requires a total of $nm \cdot (nm - 1)$ questions, one question per ratio. For example, a 3×10 model would entail 870 questions – obviously an unrealistic number, given that the model can be merely a small subset in a complex belief network or influence diagram. The elicitation procedure that we now turn to describe requires an upper-bound of $n \cdot (m + 1)$ questions, which, in the 3×10 example, entails 40 questions.

The two-way elicitation procedure makes an extensive use of a construct that we call *likelihood graphs*, or *L-graphs* for brevity. Formally, we have the following definition:

Definition 4: Each model $< E, H, P >$ can be associated with a *causal* L-graph and a *diagnostic* L-graph, as follows. The causal L-graph consists of a set of nodes, one for each subjective probability of type $E_{i,j}$. Each two nodes $E_{i,j}$ and $E_{k,l}$ may or may not be connected, as follows. If the conditional likelihood-ratio $E_{i,j,k,l}$ is known, the two nodes are connected by an arc; Otherwise, the nodes are left unconnected. When the arcs set of an L-graph is null, the graph is said to be empty. A *diagnostic* L-graph is exactly the same as a causal L-graph, except that it is built from H-type rather than E-type nodes and arcs.

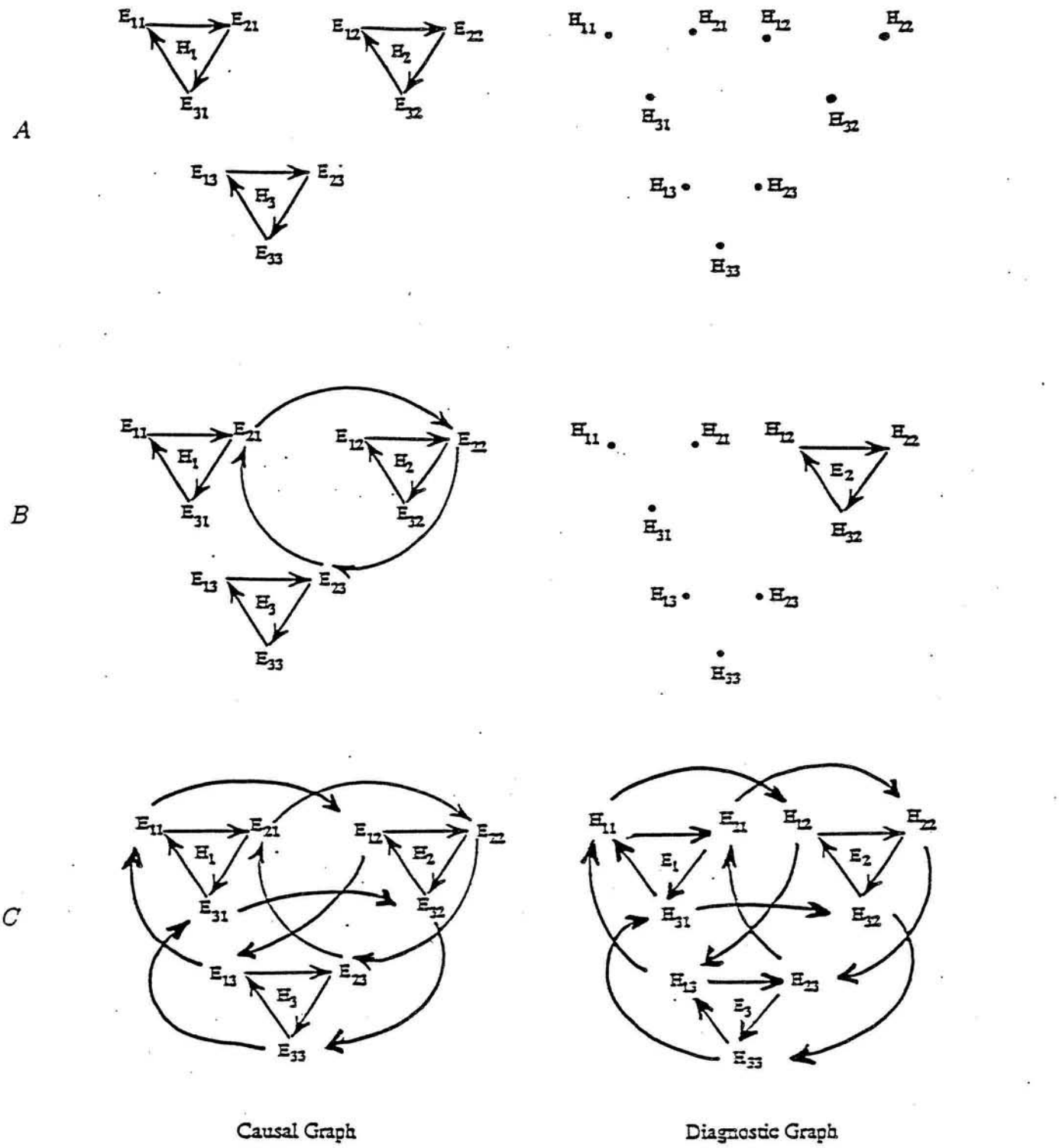

Figure 5: Three "snapshots" from a two-way elicitation process designed to determine the P element of a $< \{e_1, e_2, e_3\}, \{h_1, h_2, h_3\}, P >$ model. The elicitation process begins with empty causal and diagnostic L-graphs and terminates when both graphs are fully connected. In each graph, the nodes are clustered according to their second index, which also determines the cluster number. In the causal (diagnostic) graph, cluster number k contains information about the relative likelihoods of the e_i's (h_i's) under the background assumption that h_k (e_k) is present.

To illustrate, consider the L-graph depicted in figure 5-b, which refers to a 3×3 model in which $H = \{h_1, h_2, h_3\}$ and $E = \{e_1, e_2, e_3\}$. The graph contains 9 nodes, one for each conditional probability $E_{i,j}$. The nodes are arranged spatially in clusters, according their second index. Some of the nodes are connected by arcs, and the arcs's labels are set to the conditional likelihood-ratios associated with their end-nodes. For example, the label of the arc that connects $E_{2,1} = P(e_2|h_1)$ and $E_{2,2} = P(e_2|h_2)$ is $E_{2,1,2,2} = P(e_2|h_1)/P(e_2|h_2)$. The special connectivity in this particular example will be discussed shortly. We now turn to describe some general properties of L-graphs.

Definition 5: An L-graph is said to be *connected* if every two nodes in the graph are connected by a path.

Definition 6: An L-graph is said to be *fully-connected* if every two nodes in the graph are connected by an arc.

Definition 7: Let x and y be two connected nodes in an L-graph. The *intensity* of the path between x and y is taken to be the product of all the labels of the arcs along the path.

Lemma 1: If x and y are two connected nodes in an L-graph, then all the paths between x and y have the same intensity.

Lemma 2: If an L-graph is connected, it is also fully-connected.

Lemma 3: Let $E_{i,j}$ and $E_{i,k}$ be two nodes in a causal L-graph, and let $H_{j,i}$ and $H_{k,i}$ be the two nodes in the respective diagnostic L-graph with their indices reversed. It follows that the labels of the arcs $(E_{i,j}, E_{i,k})$ and $(H_{j,i}, H_{k,i})$ are proportional to each other. In particular, $E_{i,j,i,k} = H_{j,i,k,i} \cdot O_{k,j}$, where $O_{k,j} = P(h_k)/P(h_j)$ is the prior odds favoring hypothesis h_k on hypothesis h_j.

We sketch the proofs as follows. Since the *intensity* of a directed path is the product of all the arc labels along the path (def. 7), we immediately get from the arcs's definition that this is a telescopic product. For example, the intensity of the path $((E_{3,1}, E_{1,1}), (E_{1,1}, E_{2,1}), (E_{2,1}, E_{2,2}), (E_{2,2}, E_{3,2}))$ is:

$$\frac{P(e_3|h_1)}{P(e_1|h_1)} \cdot \frac{P(e_1|h_1)}{P(e_2|h_1)} \cdot \frac{P(e_2|h_1)}{P(e_2|h_2)} \cdot \frac{P(e_2|h_2)}{P(e_3|h_2)} = \frac{P(e_3|h_1)}{P(e_3|h_2)} \tag{14}$$

In other words, the topology of a path has no impact on its intensity. Also, the intensity of a loop is 1, and thus loops do not change the intensity either. It follows that in an L-graph, all the paths that connect the same pair of nodes have the same intensity (lemma 1). Now, let $E_{i,j}$ and $E_{k,l}$ be two arbitrary nodes in a *connected* L-graph (def. 6). Since the two nodes are connected by a path, we can compute the intensity of that path. This intensity will characterize *all* the paths that connect $E_{i,j}$ and $E_{k,l}$, including the single-step path $(E_{i,j}, E_{k,l})$. Said otherwise, the fact that two nodes are connected by a path implies that their respective conditional likelihood-ratio can be computed, and thus that they can be connected by an arc (def. 4). It follows that a connected graph is fully-connected (lemma 2).

Lemma 3 is the likelihood-ratio version of Bayes rule, stated in the language of L-graphs. The proof follows immediately from formulas (2), (12), and (13), and the definition of $O_{i,j}$ in lemma 3. The lemma implies that the causal and diagnostic L-graphs of any given model $< E, H, P >$ mirror each other. Specifically, every inter-cluster arc $E_{i,j,i,k}$ in the former is proportionally related to an intra-cluster arc $H_{j,i,k,i}$ in the latter. For example, in figure 5-b, the cross-triangle circuit that connects $E_{2,1}$, $E_{2,2}$, and $E_{2,3}$ in the diagnostic graph is the dual image of the within-triangle circuit that connects $H_{1,2}$, $H_{2,2}$, and $H_{3,2}$ in the diagnostic graph.

Taken together, lemmas 1-3 imply that the L-graphs of a model $< E, H, P >$ provide a convenient way to organize the process through which P is elicited from a human expert. It's important to note however that neither the expert nor the knowledge engineer need ever see these graphs – they are merely a graphical means to explain and study the dynamics of the elicitation process. In a nutshell, the process begins by setting up the empty causal and diagnostic L-graphs associated with E and H, respectively, leaving all the arcs in both graphs unspecified (see figure 5-a, and assume that the arcs on the left are not there). Subsequently, when we estimate or compute a particular conditional likelihood-ratio, we draw its respective arc in the relevant graph. Therefore, the elicitation procedure commences with a pair of empty L-graphs and terminates when both graphs are fully-connected. Specifically, we have the following steps:

1. Given: an $n \times m$ model $< E, H, P >$, where P is unknown.
2. Use the elements of E and H to draw empty causal and diagnostic L-graphs, respectively;
3. Using a series of m one-way elicitation procedures, connect all the nodes within *each* cluster in the causal L-graph;
4. Using a *single* one-way elicitation procedure, connect all the nodes within *one* cluster in the diagnostic L-graph;
5. Connect all the nodes in the causal graph;
6. Connect all the nodes in the diagnostic graph.

The remainder of this section demonstrates the procedure in the case of a 3×3 model of the form $< \{e_1, e_2, e_3\}, \{h_1, h_2, h_3\}, P >$. The extension to $n \times m$ models is straightforward.

The procedure begins by constructing the causal and diagnostic graphs of the model, as depicted in figure 5-a (the spatial clustering of the nodes according to their second index is not necessary, and is done here to promote clarity). In step 2, the knowledge engineer and the expert focus on the model's causal graph, considering one cluster of nodes at a time. Specifically, by successively fixing the background hypothesis on the values h_i, $i = 1, \ldots m$, the knowledge engineer administers m one-way elicitation procedures with h_i being the conditioning proposition. Taken in isolation, each one of these elicitation procedure is precisely the same as the procedure that we have described in section 4. Therefore, when step 2 is completed, each cluster $k, k = 1, \ldots, m$ in the L-graph is characterized by the output of its respective one-way elicitation procedure, which can be arranged as a vector of the form[3]:

$$(\frac{P(e_2|h_k)}{P(e_1|h_k)}, \frac{P(e_3|h_k)}{P(e_2|h_k)}, \ldots, \frac{P(e_n|h_k)}{P(e_{n-1}|h_k)}) \tag{15}$$

Or, using our L-graph notation:

$$(E_{2,k,1,k}, E_{3,k,2,k}, \ldots, E_{n,k,n-1,k}) \tag{16}$$

[3]In L-graphs, the term "cluster k" refers to all the nodes $E_{i,k}$ (or $H_{i,k}$) whose second index is k.

For any given k, this vector gives the labels of all the arcs that connect all the nodes within the kth cluster. Thus, when step 2 is completed, all the nodes within each cluster in the causal graph can be connected, leading to the three triangles in figure 5-a.

In step 3, the knowledge engineer and the expert shift their attention to the model's diagnostic graph. Their first action is to select one particular cluster of H nodes. The selection rationale should follow the guidelines given in section 3. For example, let e_k and e_j stand for *chest ache radiating to the left arm* and *swollen ankles*, respectively. If the expert can retrieve from his experience more e_k cases than e_j cases, the knowledge engineer should focus on the kth cluster in the diagnostic graph. In any event, it's important to note that *any* one cluster will do. Having selected one such cluster, the knowledge engineer proceeds to administer a diagnostic, or "forward reasoning," elicitation procedure, with e_k being the conditioning proposition. That is, the questions will follow the format: "Suppose that a patient suffers from a certain symptom e_k. To what extent is h_i more likely than h_j?" Note that even though the direction of the questions is reversed compared to the one-way elicitation procedure described in section 4, both procedures follow exactly the same structure. Algebraically, the diagnostic procedure (associated with the kth cluster) yields a single output vector, as follows:

$$\left(\frac{P(h_2|e_k)}{P(h_1|e_k)}, \frac{P(h_3|e_k)}{P(h_2|e_k)}, \ldots, \frac{P(h_m|e_k)}{P(h_{m-1}|e_k)}\right) \tag{17}$$

Or, using our L-graph notation:

$$(H_{2,k,1,k}, H_{3,k,2,k}, \ldots, H_{m,k,m-1,k}) \tag{18}$$

As we have argued following formula (16), the availability of this vector enables us to connect all the nodes in the kth diagnostic cluster. This is illustrated in the right side of figure 5-b for $k = 2$.

In Stage 4 of the process, and assuming $k = 2$, we invoke lemma 3 and vector (18) to carry out the following calculations:

$$E_{2,i,2,j} = H_{i,2,j,2} \cdot O_{j,i} \tag{19}$$

By varying the index i over 1,2,3 and the index j over 2,3,1, formula (19) is applied three times, yielding the arc labels $E_{2,1,2,2}$, $E_{2,2,2,3}$, and $E_{2,3,2,1}$. Following our graph conventions, these labels enable us to construct the across-triangle circuit that connects all the $E_{2,i}$ nodes in the left side of figure 5-b, $i = 1, 2, 3$. As a result, every two nodes in the diagnostic graph become connected by a path. Invoking lemma 2, we proceed to compute the labels of *all* the arcs that connect *all* the nodes in the diagnostic graph. Said otherwise, the causal graph becomes fully connected.

In step 5, lemma 3 is used in an inverted fashion (and with different indices): now that we have computed all the arc labels in the causal graph, we use lemma 3 to compute all the (yet unknown) arc labels in the diagnostic graph. As a result, both graphs become fully connected. This completes the elicitation procedure.

We note in closing that the elicitation procedure is somewhat of an overkill, in the sense that it produces (in the way of connectivity) more likelihood-ratios than would normally be necessary, and some of these ratios are more "interesting" than others. For example, inter-cluster ratios like $e_{i,k,j,k} = P(e_i|h_k)/P(e_j|h_k)$ are not particularly useful for most belief revision algorithms. At the same time, some intra-cluster ratios such as $e_{i,j,i,k} = P(e_i|h_j)/P(e_i|h_k)$ are very relevant, as they measure the extent to which observation e_j serves to discriminate between the competing hypotheses h_j and h_k. These are the famous *weights of evidence* parameters that can be found in the writings of Good (1950), Peirce (1956), and Minsky & Selfridge (1961).

Estimating prior rates: Stages 5 and 6 of the elicitation procedure assume that the prior likelihood ratios of the hypotheses are known. In some situations, e.g. when the h_i's represent a set of well-documented diseases, the prevalence of the diseases in the general population are available from textbook information and field records. In other situations, the base-rate ratios will have to be elicited from a domain expert. This can be done through an unconditional version of the one-way elicitation procedure described in Section 4. That is, the expert will be asked to compare the relative likelihoods of $\frac{1}{2} \cdot n \cdot (n-1)$ hypotheses, forming an unconditional likelihood matrix. The maximal eigenvector of the matrix will then be taken to be a ratio-estimate of the desired base-rate likelihood vector.

6 Conclusion

Our elicitation approach relies heavily on the eigenvector method, the cornerstone of Saaty's "Analytic Hierarchy Process." As Harker and Vargas (1987) put it, the AHP framework is designed to cope with intuitive, rational, and irrational judgement, with and without certainty. It is thus natural, in our opinion, to attempt to apply it to the problem of eliciting degrees of support, where rational knowledge is often combined with intuitive guts feeling and inconsistent judgement. In general, we concur with Fischhoff et al (1980) that it is inappropriate to "think of a person's opinion about a set of events as existing within that person's head in a precise, fixed fashion, just waiting to be measured." Yet asking experts to provide numeric degrees of support and then plugging them verbatim into a knowledge-base is a common practice among many knowledge engineers. We believe that the elicitation problem merits a more rigorous treatment, and this paper provides a step in that direction.

We conclude with some comments on limitations and future research directions. The belief revision process that our elicitation methods seek to support is based on several restrictive assumptions regarding the probabilistic backdrop of the $< E, H, P >$ model. First, given any one hypothesis $h \in H$, any two observations $e_i, e_j \in E$ are assumed to be conditionally independent with respect to h. Second, the hypotheses set H is assumed to be exhaustive and mutually exclusive. Finally, it is implicitly assumed that a "noisy-or" relationship exists between H and E. That is, it is assumed that any one hypotheses $h \in H$ can cause any one of the $e \in E$ with probability $P(e|h)$.

The above assumptions are quite restrictive, and trying to relax them is an important challenge that goes beyond the scope of this research. Moreover, the reader should understand that the assumptions are not a limitation of the elicitation procedures presented in this paper; rather, they are inherent in the representations that we seek to support, namely belief networks and influence diagrams. That said, it is important to note that many real life diagnostic problems are characterized by these assumptions, and operational systems that were built under them, e.g. Pathfinder (Heckerman, 1991) and QMR-DT (Shwe, 1991), were shown to perform extremely well in the field.

The research program that we have undertaken will not be complete until our elicitation procedures will be tested, either in controlled experiments or in the field. To construct a

laboratory elicitation experiment, one can expose a group of subjects to many instances drawn from a certain $< E, H, P_0 >$ model (where P_0 is known to the experimenter but not to the subjects), and then go on to elicit P. For example, H can range over the hypotheses "no risk," "moderate risk," and "high risk," and E can contain a variety of financial and business data about companies that seek commercial loans. Using P_0, the experimenter can generate (through computer simulation) many examples of "companies" drawn from the $< E, H >$ space and present them to the subject, one at a time. At some point, the experimenter can stop the training session and proceed to elicit P. By comparing P to P_0 (where the latter is the probability vector that generated the examples presented to the subject during training), the experimenter can obtain a measure of external validity for a variety of different elicitation techniques. A general methodology for comparing the validity of "competing" belief languages was described in Schocken and Wang (1993), and we intend to use it in the near future to test the elicitation procedures described in this paper. Since these procedures are quite general and "off-the-shelf," we hope that other researchers and practitioners will put them to the test in the context of building belief networks and influence diagrams.

Appendix: A Graphical Justification of the Eigenvector Method

This appendix justifies the mathematical background of the one-way elicitation procedure described in section 4. The justification is based on a 3×1 model $< \{e_1, e_2, e_3\}, h, P >$ in which a single hypothesis h manifests itself through three relevant observations, or pieces of evidence. The P symbol represents a subjective ratio-scale estimate of the "true" probabilities $P_0(e_1|h)$,$P_0(e_2|h)$, and $P_0(e_3|h)$. In what follows, we'll refer to these probabilities through the vector notation $P_0 = (P_0^1, P_0^2, P_0^3)$. In figure 6, this vector is the 3-dimensional point P_0. Figure 6, which plays a central role in this appendix, appears at the end of the paper.

The goal of the elicitation procedure is to elicit a vector, say P, so that $P = \alpha \cdot P_0$ for some scalar α. In terms of the figure, the goal is to determine one point – any point – that lies on the ray that goes through the points $(0, 0, 0)$ and P_0. Said otherwise, the goal of the elicitation procedure is to estimate the *direction* of P_0 from the origin.

To help us in this task, we consult a human expert who knows something about the domain $< E, h, P >$. After asking the expert to assume that the hypothesis h obtains, we ask him to estimate (1) the degree to which e_1 is more likely than e_2, (2) the degree to which e_1 is

more likely than e_3, and (3) the degree to which e_2 is more likely than e_3. The expert's responses – P_2/P_1, P_2/P_3 and P_3/P_1 – determine the slopes of the rays marked A, B, and C in the figure. Taken together, the three slopes define an infinite series of proportional boxes anchored in $(0,0,0)$. If the expert were a perfect estimator, the far vertex of one of these boxes (the vertex which is opposite to the $(0,0,0)$ vertex) would coincide exactly with P_0, as is the case in the figure. However, if one or more of the human-specified slopes were somewhat off target, the far vertex of the box will also be off target with respect to P_0.

We note is passing that the case of $n = 3$ is somewhat deceiving because the number of pair-wise comparisons between three objects also happens to be three. In the general case, the goal of the elicitation procedure is to simplify the task of estimating an n dimensional point by administrating a series of $\frac{1}{2} \cdot n \cdot (n-1)$ two-dimensional questions; in terms of the geometry, the method seeks to pinpoint the direction of a single point in an n-dimensional space by drawing rays (lines that extend from the origin) on any one of the space's $\frac{1}{2} \cdot n \cdot (n-1)$ orthogonal planes.

Now, the slopes that the expert specifies are not independent of each other. For example, the slope P_2/P_1 can be drawn according to the number that the expert supplied, but it can also be drawn according to the product of the expert-supplied slopes P_1/P_3 and P_3/P_2. Using this rationale, it can be seen that the expert's answers end up specifying not three slopes, but actually nine slopes, or three "versions" of the same slope for each one of the space's three orthogonal planes. If the expert's estimates were consistent, the three versions of each slope would coincide. But this is a very unlikely to happen: in reality, we will end up having three different (but hopefully not *too* different) versions of the rays A, B, and C.

If we plug the expert's answers into a 3×3 likelihood matrix, a similar analysis occurs. Although the expert is asked to specify only the three entries above the matrix's diagonal, he ends up specifying three 3-dimensional vector columns (because the matrix is reciprocal). Each one of these vectors can be seen as a different subjective attempt to pinpoint the direction of the ray that extends from $(0,0,0)$ to the elusive target P_0. In the figure, these vectors are denoted V_1, V_2, and V_3. Taken together, the three vectors define a 3-dimensional ellipsoid, or a football, in the (P_1, P_2, P_3) space.

Now, it's important to understand that the holy grail of this process – the point P_0 – does

not exist ex-ante. In other words, it is not an objective probability vector that we are after. Rather, we are trying to put the finger on a subjective probability vector that exists in the expert's head. The entire objective of the elicitation procedure is to help the expert *express* this vector through a series of 2-dimensional, pairwise comparison questions.

Because of the algebraic properties of the likelihood matrix that these questions induce, and because of the "transitive redundancy" of the questions, the expert ends up producing three different estimates of the desired vector (V_1, V_2, and V_3). With that in mind, it is reasonable to try to synthesize a point that lies *in the general direction* of these vectors. Graphically, we wish to determine the principle axis of the 3-dimensional ellipsoid defined by V_1, V_2, and V_3. As it turns out, this direction is given by the eigenvector associated with the largest eigen value that characterizes the three vectors.

Taking this graphical analysis one step further, we can also shed light on the extent of the inconsistency that the expert displays. If the expert were perfectly consistent, the "football" defined by V_1, V_2, and V_3 would be completely deflated, amounting to a subset of the ray defined by $(0,0,0)$ and P_0. Said otherwise, the inconsistency of the expert can be measured in terms of the football's volume: the greater the volume, the greater the inconsistency. As the figure indicates, this volume is determined by the eigen vectors associated with the second and the third largest eigen values associated with V_1, V_2, and V_3. Since the football is not symmetric, some of its (non-primary) axes will be longer than others; these axes literally point at the expert's answers (2-dimensional orthogonal planes) that are mostly responsible for the inconsistency. We believe that this analysis provides a much better estimate of the expert's inconsistency and ways to resolve it compared to the heuristic way in which inconsistency is presently treated in the standard AHP model.

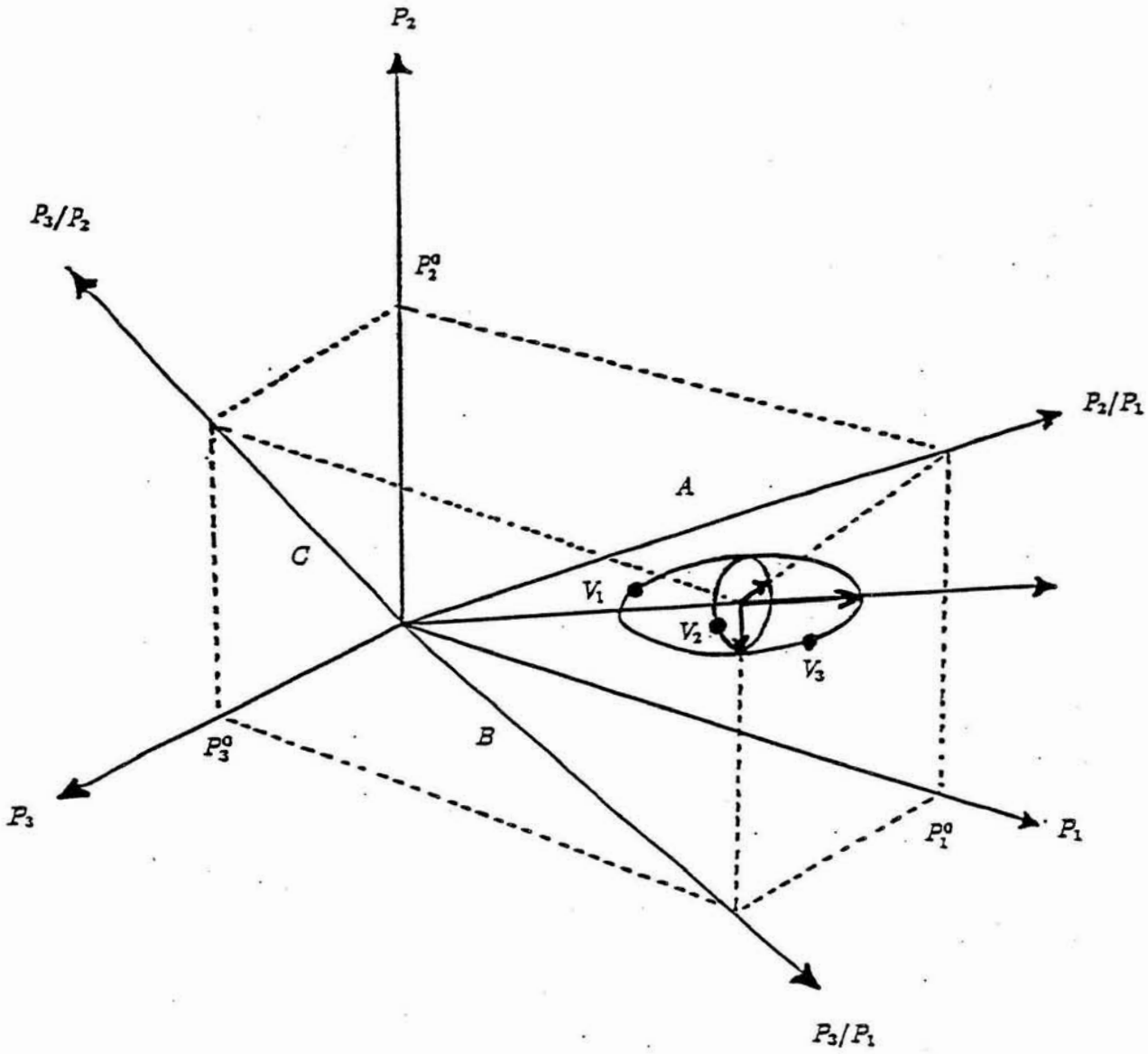

Figure 6: A pictorial depiction of the process through which the P element of a $< \{e_1, e_2, e_3\}, \{h\}, P >$ model is determined. The goal of the elicitation procedure is to construct a ratio-scale estimate of the point $P_0 = (P_0^1, P_0^2, P_0^3)$, which represents the expert's implicit beliefs about the relative conditional likelihoods of the three observations. Since this n-dimensional point is not readily available (from a cognitive standpoint), the expert is asked to "plot" three 2-dimensional rays on the orthogonal planes of the (P_1, P_2, P_3) space. More specifically, the expert is given "three trials" for each ray. Together, these trials determine the vectors V_1, V_2, and V_3. The expert perception of P_0 is then taken to be a point which lies on the principle axis of the ellipsoid that the three vectors specify. The direction of this axis is the maximal eigenvector that characterizes the matrix $[V_1, V_2, V_3]$. The expert's inconsistency is given by the volume of the ellipsoid, which is determined by the eigenvectors of the second and third largest eigen values of the same matrix.

References

[1] B.G. Buchanan and E.H. Shortliffe. Uncertainty and evidential support. In B.G. Buchanan and E.H. Shortliffe, editors, *Rule-Based Expert Systems*, pages 217–219. Addison-Wesley, 1984.

[2] G.F. Cooper. The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence*, 42(2-3):393–402, March 1990.

[3] J.G. de Graan. Extensions of the multiple criteria analysis method of t.l. saaty,. Technical report, EURO IV, Cambridge University, 1980.

[4] R.O. Duda and E.H. Shortliffe. Expert systems research. *Science*, 220:261–268, 1981.

[5] H. J. Einhorn and R. M. Hogarth. Decision making: Going forward in reverse. *Harvard Business Review*, January-February:66–70, 1987.

[6] B. Fischhoff and R. Beyth-Marom. Hypothesis evaluation from a bayesian perspective. *Psychological Review*, 90(3), 1983.

[7] B. Fischhoff, P. Slovic, and S. Lichtenstein. Knowing what you want: Measuring labile values. In T.S. Wallsten, editor, *Cognitive Processes in Choice and Decision Behavior*. Hillsdale, N.J.: Erlbaum Associates, 1980.

[8] I.J. Good. *Probability and the Weighing of Evidence*. New York: Hafner, 1950.

[9] D.L. Hamilton. A cognitive attributional analysis of stereotyping. In L. Berkowitz, editor, *Advances in Experimental Social Psychology (Vol. 12)*. New York: Academic Press, 1979.

[10] P.T. Harker and L.G. Vargas. The theory of ratio-scale estimation: Saaty's analytic hierarchy process. *Management Science*, 33(11):1383–1403, 1987.

[11] D. Heckerman, E. Horvitz, and B. Nathwani. Toward normative expert systems: the pathfinder project. Mthods of Information in Medicine, Forthcoming.

[12] M. Henrion, J.S. Breese, and E.J. Horvitz. Decision analysis and expert systems. *AI Magazine*, Winter:64–91, 1991.

[13] R.A. Howard and J.E. Matheson. Influence diagrams. In R.A. Howard and J.E. Matheson, editors, *Reading on the Principles and Applications of Decision Analysis*, pages 721–762. Strategic Decisions Group, Menlo Park CA, 1981.

[14] D. Kahneman, P. Slovic, and A. Tversky. *Judgement Under Uncertainty: Heuristics and Biases.* Cambridge University Press, 1982.

[15] S. Kotz and D.F. Stroup. *Educated Guessing.* Marcel Dekker, 1983.

[16] D.H. Krantz. A theory of magnitude estimation and cross-modality matching. *J. Math. Psychology*, 9(2):168–199, 1972.

[17] S. Lichtenstein and J. R. Newman. Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychonomic Science*, 9:563–564, 1967.

[18] M. Minsky and O. Selfridge. Learning in random nets. In C. Cherry, editor, *Information Theory*, pages 335–347. London: Butterworths, 1961.

[19] J. Pearl. Fusion, propagation and structuring in belief networks. *Artificial Intelligence*, 29:241–288, 1986.

[20] C.S. Peirce. The probability of induction. *The World of Mathematics*, 2:1341–1354, 1956.

[21] K.R. Popper. Corroboration, the weight of evidence. In *The Logic of Scientific Discovery*, pages 387–419. Scientific Editions, New York, 1959.

[22] T.L Saaty. *The Analytic Hierarchy Process.* New York: McGraw-Hill, 1980.

[23] L.J. Savage. *The Foundations of Statistics.* Wiley, 1954.

[24] S. Schocken and Y.M. Wang. An experimental comparison of two rule-based belief languages. *Information Systems Research*, 4(4), 1993.

[25] R.D. Shachter and D.E. Heckerman. Thinking backward for knowledge acquisition. *AI Journal*, Fall:55–61, 1987.

[26] G. Shafer. Probability judgement in artificial intelligence and expert systems. *Statistical Science*, 2(1):3–16, 1987.

[27] M. Shwe, B. Middleton, D. Heckerman, M. Henrion, E. Horvitz, and et al Lehmann, H. Probabilistic diagnosis using a reformulation of the internist-1/qmr knowledge base: 1. the probabilistic model and inference algorithms. *Methods of Information in Medicine*, 30:241–255, 1991.

[28] S.S. Stevens. Measurement, psychophysics, and utility. In C.W. Churchman and P. Ratoosh, editors, *Measurements, Definitions and Theory*. Wiley, New York, 1959.

[29] S.S. Stevens and E. Galanter. Ratio-scale and category-scale for a dozen perceptual continua. *Experimental Psychology*, 54:377–411, 1964.

[30] A. Tversky and D. Khaneman. Judgement under uncertainty: Heuristics and biases. *Science*, 185:1124–1131, 1974.

[31] E. Velez-Garcia, J. Durant, R. Gams, and A. Bartolucci. Results of a uniform histopathological review system of lymphoma cases: a ten-year study of the south-eastern cancer study group. *Cancer*, 52:675–679, 1983.

[27] M. Shwe, B. Middleton, D. Heckerman, M. Henrion, E. Horvitz, and et al Lehmann, H. Probabilistic diagnosis using a reformulation of the internist-1/qmr knowledge base: 1. the probabilistic model and inference algorithms. *Methods of Information in Medicine*, 30:241–255, 1991.

[28] S.S. Stevens. Measurement, psychophysics, and utility. In C.W. Churchman and P. Ratoosh, editors, *Measurements, Definitions and Theory*. Wiley, New York, 1959.

[29] S.S. Stevens and E. Galanter. Ratio-scale and category-scale for a dozen perceptual continua. *Experimental Psychology*, 54:377–411, 1964.

[30] A. Tversky and D. Khaneman. Judgement under uncertainty: Heuristics and biases. *Science*, 185:1124–1131, 1974.

[31] E. Velez-Garcia, J. Durant, R. Gams, and A. Bartolucci. Results of a uniform histopathological review system of lymphoma cases: a ten-year study of the south-eastern cancer study group. *Cancer*, 52:675–679, 1983.