Another Look at Sensitivity of Bayesian Networks to Imprecise Probabilities

Oscar Kipersztok
Mathematics and Computing Technology
Phantom Works, The Boeing Company
P.O.Box 3707, MC: 7L-44
Seattle, WA 98124
Oscar.kipersztok@boeing.com

Haiqin Wang*
Decision Systems Laboratory
Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA 15260
hawst9@pitt.edu

Abstract

Empirical study of sensitivity analysis on a Bayesian network examines the effects of varying the network's probability parameters on the posterior probabilities of the true hypothesis. One appealing approach to modeling the uncertainty of the probability parameters is to add normal noise to the log-odds of the nominal probabilities. However, the paper argues that differences in sensitivities found on true hypothesis may only be valid in the range of standard deviations where the log-odds normal distribution is unimodal. The paper also shows that using average posterior probabilities as criterion to measure the sensitivity may not be the most indicative, especially when the distribution is very asymmetric as is the case at nominal values close to zero or one. It is proposed, instead, to use the partial ordering of the most probable causes of diagnosis, measured by a suitable lower confidence bound. The paper also presents the preliminary results of our sensitivity analysis experiments with three Bayesian networks built for diagnosis of airplane systems. Our results show that some networks are more sensitive to imprecision in probabilities than previously believed.

1 INTRODUCTION

Sensitivity analysis is a method to investigate the effects of imprecision of a model's parameters on its output. For a probabilistic model, or a Bayesian network more specifically, performing a sensitivity analysis yields insight into the robustness of the network's performance in diagnosis or prediction under uncertainty in its probabilistic parameters. Basically, there are two approaches to sensitivity analysis: theoretical and empirical. The theoretical approach establishes a function expressing a posterior probability of interest in terms of the parameters under study (e.g., [Laskey, 1995, Castillo, Gutierrez, & Hadi, 1997, Kjærulff & van der Gaag, 2000]). The empirical methods examine the effects of varying the network's parameters on diagnostic or predictive performance (e.g., [Pradhan *et al.*, 1996, Coupé *et al.*, 1999]).

The common present belief, to a great degree based on a series of experiments in [Pradhan *et al.*, 1996], is that Bayesian networks are, on the average, insensitive to inaccuracies in the numeric value of their probabilities. Henrion *et al.* [1996] further elaborated one of the experiments and explored the possible explanations of the low sensitivity. In [Henrion *et al.*, 1996], the conclusions were drawn based on the average of the probabilities of the true diagnosis with simulated scenario cases run by imparting random noise on the nominal probabilities of known networks at increasing levels of uncertainty. The reported results differentiated between true-positive diagnosis cases and true-negatives, and between the effect of noise on the priors of conditional probabilities (also called link probabilities), leak probabilities, and prior probabilities.

This paper argues that differences in sensitivities found in [Henrion *et al.*, 1996] between true-positive and true-negative results may not be valid because the log odds-normal distribution, which is used to generate random noise on probabilities, may not be a suitable distribution in the range of standard deviations where the differences were observed. The presence of true-positive and true-negative biases in Bayesian network diagnosis results from misevaluation of the network by experts, and should be corrected once those biases are detected. Differences in network sensitivity due to noise on the different types of probabilities is a quantifiable random effect that depends on the distri-

*Summer 2000 intern at Mathematics and Computing Technology, Phantom Works, The Boeing Company, Seattle, WA 98124 bution used to model the added noise and possibly, on the topology of the network.

The paper also shows that comparing the average results of the simulated posterior probabilities to the nominal posterior probabilities may not be the most indicative measure of network sensitivity because information about the effect of the noise distribution variance is lost, especially when the distribution is very asymmetric as is the case at nominal values close to zero or one. It is in the variation of these posterior probabilities that imprecision in parameters is reflected. Although the difference in computed posteriors derived from noisy versus nominal probabilities is indicative of the sensitivity of the network, the partial ordering of the posterior probabilities is argued to be a more critical indicator of the outcome of the diagnosis. It is proposed then to assess the sensitivity of the network based on the effect that the uncertainty in probabilities has on the partial ordering of the probable causes, measured using a suitable lower confidence bound.

A series of experiments were designed to investigate the sensitivity of three Bayesian networks built for diagnosis of airplane systems, to imprecision in different type of probabilities: prior probabilities, conditional probabilities, and leak probabilities. We varied the probability parameters in the networks by introducing log-odds normal noise for the following range of standard deviations: 0.1, 0.25, 0.5, 0.8, and 1, respectively. The criterion we used to measure the sensitivity of the networks is a set of lower confidence bounds (50%, 80%, 90%, 95%, and 99%). Our results showed that generally, increasing noise level to the probabilities produced higher sensitivities in the tested networks. The results also suggested that prior probabilities turned out to be more influential parameters to diagnosis in our networks, compared with conditional probabilities and leak probabilities. In contrast to the common belief that Bayesian networks are generally insensitive to imprecise probabilities, our results showed that some networks can show significant sensitivity to imprecision in probabilities even with a small variance in the noise distribution. Our results agree with recent findings of high sensitivities reported by [Coupé *et al.*, 1999] in an empirical study using a Bayesian network from medical prognosis and treatment planning.

The paper is organized as follows: Section 2 elaborates our arguments related to the empirical approach to sensitivity analysis. Section 3 describes the sensitivity experiments conducted on three large production networks built for diagnosis of airplane systems. In Section 4, we give a brief conclusion about the results of our experiments.

2 LOG-ODDS NOISE AND MEASURES OF SENSITIVITY IN BAYESIAN NETWORK

The two main points of this paper are: a) the log-odds normal distribution, although it has appealing properties for modeling the noise of probabilities, it may not be valid for assessing network sensitivity for values of standard deviations greater than one, and b) the use of averages for comparing the posterior probabilities, derived from noisy probabilities, to the nominal posteriors may hide the effect of the variance of the noise distribution, especially for probability values near zero or one. We will deal with each point separately.

2.1 VALIDITY OF LOG-ODDS NORMAL DISTRIBUTION

The log-odds normal distribution is a suitable model for noise imposed on probabilities because the sampled probability remains in the [0,1] range and because it recognizes differences imparted by noise to probabilities near 0 or 1 versus those in the middle of the range near 0.5 [Henrion *et al.*, 1996]. However, in this paper we argue that this distribution may not be valid for standard deviations greater than one for the purpose of assessing network sensitivity to probability noise.

Equation 1 illustrates the probability density of the log-odds normal distribution:

$$Y = log\ \frac{p}{1-p} + \epsilon\ , \tag{1}$$

where $\epsilon \sim N(0, \sigma)$, transitions from unimodal to bimodal for values of $\sigma > 1$. A simplified equation for the distribution of the nominal probabilities with added noise, p', in terms of the nominal p and noise ϵ is

$$p' = \frac{1}{1 + (p^{-1} - 1)\ 10^{-\epsilon}}\ . \tag{2}$$

The relation between the noisy odds and the nominal odds is then:

$$\frac{1-p'}{p'} = \frac{1-p}{p}\ 10^{\epsilon}\ , \tag{3}$$

or $odds' = odd * 10^{\epsilon}$, where $odds' = \frac{1-p'}{p'}$ and $odds = \frac{1-p}{p}$.

This indicates that error introduced by the log-odds normal noise ϵ reflects the scale of the change of odds by a factor of 10^{ϵ}.

Table 1 shows values of p' computed from Equation 2 for various values of p and ϵ. Note that the values of ϵ correspond to values of σ in the standard normal distribution. For values of $\epsilon > 1$, the difference between the noisy and nominal probabilities increases rapidly for values of the nominal probability that are close to zero. The effect is also large but less pronounced for values of nominal probabilities in the mid-range towards p = 0.5.

The log-odds normal distribution is an adequate model of noise added to probabilities for values of $\sigma < 1$, where the distribution is unimodal. For values of $\sigma > 1$ the distribution becomes bimodal as shown in Figure 1. Using the distribution in that range to describe the noise on the priors is equivalent to considering an expert who assesses a prior probability, known to be near zero, and erring in judgment by such margin that the true prior probability may, in fact, be close to one! This is what the log-odds normal distribution implies for large values of σ.

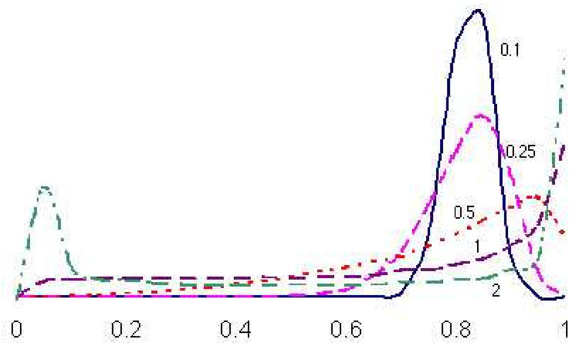

Figure 1: The log-odds normal distribution centered around the nominal prior of 0.8, for various values of σ

2.2 MEASURES FOR ASSESSING BAYESIAN NETWORK SENSITIVITY

In using Bayesian networks for diagnosis the partial ordering of probable causes resulting from the update of the posterior probabilities given a set of findings constitutes the diagnosis. While the most probable cause is often given the highest consideration, typically, in multiple fault-diagnostic systems it is a particular set of the top causes (e.g., the top five) and their partial ordering that is most informative. Since very seldom the diagnosis singles out a particular cause, the partial ordering provides guidance for subsequent actions. The effect that noise has on the posterior partial ordering of the causes is, therefore, a significant measure of the network sensitivity.

Table 2 shows the top five suspect parts selected from a Bayesian network diagnosis system representing a particular test case scenario of an airplane fault. It compares posterior probability from the nominal network with the average posteriors from one hundred noisy networks. The noise distribution used was the log-odds normal with $\sigma = 0.5$. Note that the deviation of the average posterior from the nominal does not appear to be substantial.

The top of Table 3 shows the average change in rank order for the five suspect parts of Table 2. This average reflects the average absolute-value change in rank for each part from its nominal rank that is due to noise added to the network probabilities. The average change in rank shows that noise with $\sigma = 0.5$ is expected on the average to affect a change in the rank of the top five suspect parts by approximately one ranking order. By itself, this is not a bad result considering the size of the variance of the distribution used. However, it is somewhat misleading.

Typically, for airplane diagnosis, the reliability estimates of most airplane parts is of order greater than 10^5 hours for the mean time between part failures. The corresponding prior probabilities are therefore approximately of order smaller than 10^{-5}. At such low probabilities the log-odds normal distribution is very asymmetric and the average rank does not adequately represent the effect that the noise imparts on the network. Shown at the bottom of Table 3 are lower-bound confidence estimates for confidence levels from 50% to 99%. The data show that for noise with $\sigma = 0.5$, there is a 50% chance that the ranking order of the parts could change by at least one position. For the most probable suspect part (i.e., Part 1), there is a 20% chance that it could drop by more than two ranks, a 10% chance that it may drop by more than three orders in rank, and a 1% chance that it may drop by more than four. For networks with high sensitivity to noise, the nominal diagnosis could advise the airplane maintainer to unleash a series of irrelevant actions that could result in unnecessary and costly delays and cancellations.

This analysis, we believe, is more representative of the sensitivity of the network due to noise in the network probabilities. The remainder of the paper will present data compiled from several airplane diagnosis networks under various test scenarios, and will distinguish the network sensitivity to noise contributions from prior probabilities, conditional probabilities, and leak probabilities.

Table 1: Values of the noisy p' computed from Equation 2 for different values of the nominal p and noise ϵ.

Values of ϵ	**Values of p and percentage of $(p'-p)/p$**									
	0.0001	%	0.01	%	0.1	%	0.25	%	0.5	%
0.1	0.00013	26	0.013	26	0.12	23	0.30	18	0.56	11
0.3	0.00020	100	0.020	98	0.18	81	0.40	60	0.67	33
0.5	0.00032	216	0.031	210	0.26	160	0.51	105	0.76	52
0.7	0.00050	401	0.048	382	0.36	258	0.63	150	0.83	67
1	0.00010	899	0.092	817	0.53	426	0.77	208	0.91	82
3	0.09092	90817	0.91	8999	0.99	891	1.00	299	1.00	100
5	0.90910	908999	1.00	9890	1.00	900	1.00	300	1.00	100

Table 2: The nominal posteriors of the top five suspect parts from an airplane diagnosis compared to the average from one hundred noisy posteriors (log-odds normal, $\sigma = 0.5$).

	Part 1	**Part 2**	**Part 3**	**Part 4**	**Part 5**
Nominal posterior	0.40	0.29	0.11	0.07	0.07
Average posterior	0.35	0.28	0.15	0.07	0.07
Standard Deviation	0.24	0.23	0.15	0.07	0.09

Table 3: Lower confidence bounds and average changes of the ranks for the five most probable causes.

	Part 1	**Part 2**	**Part 3**	**Part 4**	**Part 5**
Average rank change	1.02	1.05	1.26	1.21	1.15
Standard deviation	1.19	0.93	0.90	0.83	1.06
50th percentile	1	1	1	1	1
80th percentile	2	2	2	2	2
90th percentile	3	2	2	2	2.9
95th percentile	3	3	3	2	3
99th percentile	4	3.99	3	3	4

3 SENSITIVITY EXPERIMENT

3.1 MEASURE OF DIAGNOSTIC PERFORMANCE

As indicated in Section 2.2, average posterior probabilities may not be an adequate measure to assess sensitivity of Bayesian networks with respect to diagnosis, especially when the probability distribution is extremely skewed by adding in the log-odds normal noise. Instead, lower confidence bounds on rank changes of the diagnosis recommended by a partial-ordered list of suspect parts, better reflect the effect that random noise has on the network.

In our experiments, we use lower confidence bounds for 0.50, 0.80, 0.90, 0.95, 0.99 percentiles of the diagnosis ranks over test cases to quantify diagnostic performance. Average and standard deviation of rank changes are also calculated for comparison.

3.2 NETWORKS AND TEST CASES

We used three large networks built for diagnosing three major airplane systems. A number of test case scenarios were defined for each network. These scenarios represented real-life cases encountered during routine airplane maintenance procedures. Each test case constitutes a set of findings, used as inputs to the networks, which do not necessarily isolate the failed parts with certainty, but rather generate a ranked list of the most likely suspect parts. The ranked list of parts is what constitutes the diagnosis given a particular test case scenario. For illustration purposes and without loss of generality we denote the three networks as Net 1, Net 2, and Net 3. The airplane parts are also denoted by numbers associated with their posterior ranking order, i.e., Part 1, Part 2, etc.

3.3 EXPERIMENTAL DESIGN

We tested with three networks built for diagnosing airplane part failure. For each network, we first classified the nodes into different sets according to their probability types: prior, conditional and leak. To generate a noisy network, we added noise to each set of nodes independently for a given level of noise and scenarios. Each scenario was run one hundred times with the same noise distribution for each set of nodes. A noisy network was generated in each run. The total number of networks used in our experiment were 34503, consisted of 3 types of probability * 5 levels of noise * 100 runs * (3+5+15) scenarios, plus 3 original networks without noise.

The test began with a diagnosis on the nominal network given the findings defined in the scenario. For this network, the nominal partial ordering of the recommended failed parts was generated. The rank of each probable failed-part was recorded according to the partial ordering. Under the same situation, (i.e., the same set of nodes, the same noise distribution, and the same scenario), the noisy networks were used to compute the noisy rank changes of the diagnosed failed-parts from the rank changes computed with the nominal network. The effect of noise was assessed by computing statistics on the rank changes, such as average and standard deviation of rank changes, and 0.50, 0.80, 0.90, 0.95, 0.99 percentiles of lower confidence bounds.

3.4 RESULTS

Figure 2 plots the average rank changes over one hundred cases across different scenarios of the most probable failed parts in Net 3 affected by five levels of prior noise. As expected, performance degrades as the noise increases. Note that the rank of the most probable failed-part drops, on the average, about one position when noise is distributed with $\sigma(or std) = 0.1$, and it drops about two positions when noise is distributed with $\sigma = 1.0$.

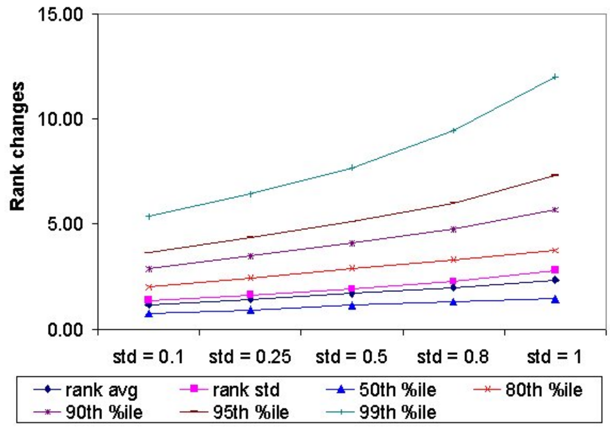

Figure 2: Rank changes of the most probable failed parts in Net 3 based on 100 run cases across different scenarios and prior noise.

Since with $\sigma = 1.0$ the most probable failed part will change on the average almost three rank positions, it may look as if the diagnosis performance is robust and insensitive to the imprecise prior probabilities. However, looking at the lower confidence bounds, Figure 2 indicates that there is a 90% chance that the most probable failed part will stay within the top five rank positions for noise with $\sigma < 0.5$. Conversely, with $\sigma >= 0.5$, there is a 90% chance that the most probable failed part will disappear from the top five recommended parts given by the diagnosis, which could possibly result in incorrect diagnosis by the network.

Figure 3 illustrates the rank changes of the top five most probable failed parts in Net 3 when the prior noise is distributed with $\sigma = 1.0$. From the chart, we see that 0.50 percentile lower bound for the five parts are smaller than rank average, further indicating the asymmetry of the noise distribution.

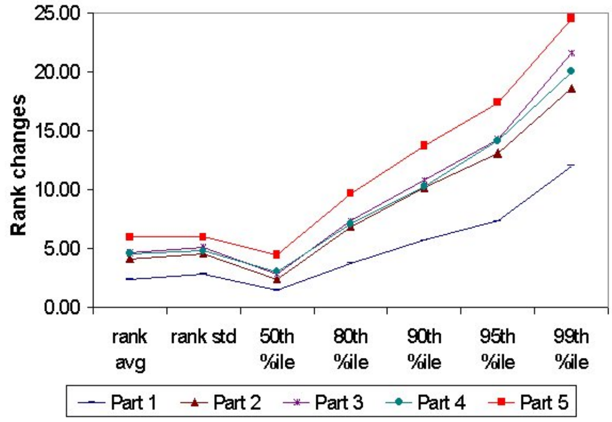

Figure 3: Rank changes of the top five most probable failed parts in Net 3 based on 100 run cases across different scenarios and prior noise $\epsilon \sim N(0, 1.0)$.

Also note the high standard deviations of the rank. This illustrates that the sensitivity of the noisy networks varies greatly with different scenarios. The noisy network may be pretty robust for some of the observations, but may be quite sensitive to others. Therefore, different scenarios play an important role in testing sensitivity of Bayesian networks.

The effect of noise on conditional probabilities and on leak probabilities is much smaller than that on prior probabilities for all of the three networks in our experiments. As shown in Figure 4, the average rank changes are smaller than 1 even when the conditional noise is distributed with $\sigma = 1.0$. The 0.99 percentile lower bounds are all smaller than 4. Therefore, in 99 percent of the time, the top five most probable failed-parts would stay in the top positions in the partial ordering given by the diagnosis.

As was the case with noise added to prior probabilities, the same trend is observed with conditional probabilities when noise level increasing. Namely, when the noise added to the conditional probability tables becomes higher, the network becomes more sensitive, as a result, the diagnosis capability of the networks degrades.

Figure 5 shows the rank changes of the most probable failed part in Net 1 based on one hundred run cases across different scenarios and different prior noise

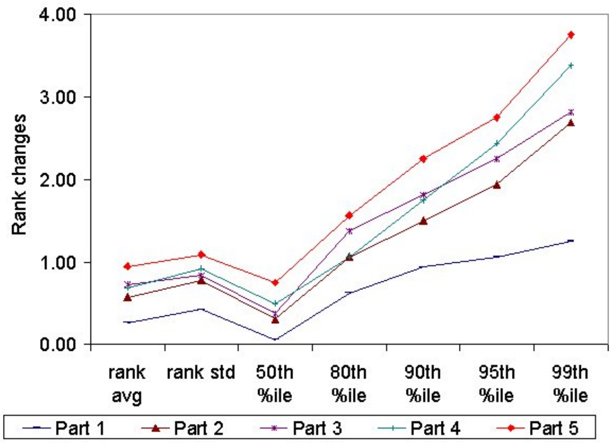

Figure 4: Rank changes of the top five most probable failed parts in Net 3 based on 100 run cases across different scenarios with CPT noise $\epsilon \sim N(0, 1.0)$.

distributions. The rank changes in Net 1 are much smaller than the rank changes found in Net 3, which indicates that different networks may have a different degree of sensitivity to imprecise probabilities. However, the rank changes in Net 2 were close to those of Net 1.

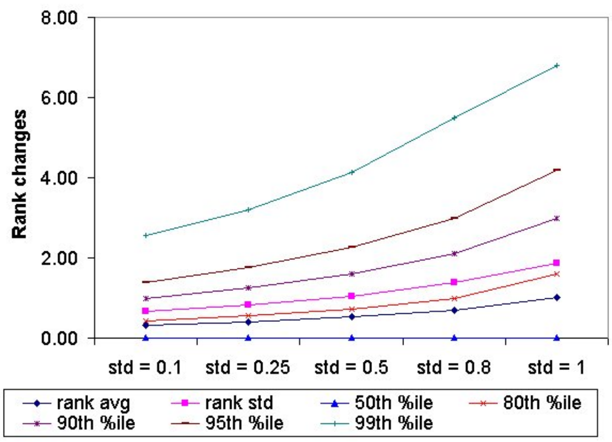

Figure 5: Rank changes of the most probable failed part in Net 1 based on 100 runs cases across different scenarios and prior noise.

4 CONCLUSION

We argue in this paper that the log-odds normal distribution is valid as a model for sensitivity analysis only in the range of standard deviations where the distribution is unimodal. The paper also shows that using average posterior probabilities as criterion to measure the sensitivity may not be the most indicative, especially when the distribution is very asymmetric as is the case at nominal values close to zero or one. It is proposed, instead, to use the partial ordering of the most probable causes of diagnosis, measured by a suitable lower confidence bound on the change in the rank order. Preliminary results of our sensitivity analysis experiments were shown with three Bayesian networks built for diagnosis of airplane systems. Our results showed that some networks are more sensitive to imprecision in probabilities than previously believed.

Acknowledgements

The authors would like to acknowledge the following individuals who have contributed in various important capacities to the work presented in this paper. Very special thanks go to Cathy Kitto, Susan Chew, Nick Walker, Andrew Booker, Paul Jackson and John Bremmer.

References

[Castillo, Gutierrez, & Hadi, 1997] Castillo, E. F.; Gutierrez, J. M.; and Hadi, A. S. 1997. Sensitivity analysis in discrete Bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 27(4):412–423.

[Coupé *et al.*, 1999] Coupé, V. M.; Peek, N.; Ottenkamp, J.; and Habbema, J. D. F. 1999. Using sensitivity analysis for efficient quantification of a belief network. *Artificial Intelligence in Medicin* 17:223–247.

[Henrion *et al.*, 1996] Henrion, M.; Pradhan, M.; Del Favero, B.; Huang, K.; Provan, G.; and O'Rorke, P. 1996. Why is diagnosis using belief networks insensitive to imprecision in probabilities? In *Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence (UAI–96)*, 307–314. San Francisco, CA: Morgan Kaufmann Publishers.

[Kjærulff & van der Gaag, 2000] Kjærulff, U., and van der Gaag, L. C. 2000. Making sensitivity analysis computationally efficient. In *Uncertainty in Artificial Intelligence: Proceedings of the Sixteenth Conference (UAI-2000)*, 317–325. San Francisco, CA: Morgan Kaufmann Publishers.

[Laskey, 1995] Laskey, K. B. 1995. Sensitivity analysis for probability assessments in Bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics* 25(6):901–909.

[Pradhan *et al.*, 1996] Pradhan, M.; Henrion, M.; Provan, G.; del Favero, B.; and Huang, K. 1996. The sensitivity of belief networks to imprecise probabilities: An experimental investigation. *Artificial Intelligence* 85(1-2):363–397.