MSc Program
Engineering Management

Monitoring, Validation and diagnosis data processing using intelligence Techniques

A Master's Thesis submitted for the degree of
"Master of Science"

supervised by
Em.O.Univ.Prof. Dipl.-Ing. Dr.h.c.mult. Dr.techn. Peter Kopacek

Chaminda Fernando, BSc

Vienna, March 2014

Affidavit

I, **Chaminda Fernando, BSc**, hereby declare

1. that I am the sole author of the present Master's Thesis, "MONITORING, VALIDATION AND DIAGNOSIS DATA PROCESSING USING INTELLIGENCE TECHNIQE", 73 pages, bound, and that I have not used any source or tool other than those referenced or any other illicit aid or tool, and
2. that I have not prior to this date submitted this Master's Thesis as an examination paper in any form in Austria or abroad.

Vienna, 11.03.2014

Signature

TABLE OF CONTENTS

ABSTRACT

This master thesis explains applying statistical and artificial intelligence techniques (Bayesian Networks) to validate and diagnose data processing systems. Complex processes have many processing variables and operators are challenged with monitoring, controlling, diagnosing and analysing current states of processes. He also has to take appropriate actions when it is needed. This thesis may help for operators who maintain such a complex systems.

The verification regime of the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO) is designed to detect events around the world. The International Monitoring System (IMS) consists of facilities around the world that equipped with seismometers that convert ground motion into electric voltage. International Data Centre (IDC) is located at the headquarters of the CTBTO and it receives data continuously from these stations for processing. Detection and Feature Extraction (DFX) and Global Association (GA) are two most important applications in the data processing pipeline. The primary function of DFX application is to identify detections and to measure features from waveforms. GA application reads detections and amplitude data for a pre-defined time interval and forms set of associations using an exhaustive search algorithm.

Hypothesis testing and estimation are used to reach conclusions about a population by examining a sample of that population. Hypothesis testing is widely used in medicine, dentistry, health care, biology and other fields as a means to draw conclusions about the nature of populations. Hypothesis testing is to provide information in helping to make decisions. The administrative decision usually depends on a test between two hypotheses and decisions are based on the outcome. The null hypothesis (H_0), stated as the null, is a statement about a population parameter, such as the population mean, that is assumed to be true. We will test whether the value stated in the null hypothesis is likely to be true. An alternative hypothesis (H_A) is a statement that directly contradicts a null hypothesis by stating that that the actual value of a population parameter is less than, greater than, or not equal to the value stated in the null hypothesis. Level of significance refers to a criterion of judgment upon which a decision is made regarding the value stated in a null hypothesis. The criterion is based on the probability of obtaining a statistic measured in a sample if the value stated in the null hypothesis were true. In behavioural science, the criterion or level of significance is typically set at 5%.

When we decide to reject the null hypothesis, the decision can be correct or incorrect. If the incorrect decision is to reject a true null hypothesis, this decision is an example of a Type I error. Other option is to retain a false null hypothesis. This decision is an example of a Type II error.

Artificial Intelligence (AI) section of this master thesis focuses on Bayesian network. AI systems have to cope with uncertainty and they have to deal with incomplete evidence leading to conclusion through its short of knowledge. This fallible conclusion is called non-monotonic reasoning. Bayesian reasoning is a kind of probabilistic reasoning. There are mainly two types of probabilities. *Prior Probability*: This Probability is also popularly known as unconditional probability. *Posterior Probability*: This type of probability is also known as conditional probability. Bayesian networks are successfully applied to a variety of applications such as machine diagnosis, robotics, data mining and natural language interpretation and planning.

A Bayesian network is used to model domains containing uncertainty in some manner. It is a graphical model that shows probabilistic relationships among a set of variables. It also consists of directed acyclic graphs (DAGs) and the links represent informational or causal dependencies among the variables. The conditional probability table of a node contains probabilities of the node being in a specific state given the states of its parents. Master thesis explains how to use Bayesian networks as a diagnostic support tool for DFX and GA application failures.

In Bayesian networks, there are four inferences. (1) Backward inferences, which is also called diagnostic inference (from effects to causes) (2) Forward inferences, which is also called predictive inference (from causes to effects) (3) Intercausal inferences, which is also called explaining away (between parallel variables). The inference reasons about the mutual causes (effects) of a common effect (cause). (4) Mixed inferences, which is also called combined inferences. It does not fit exactly into one of the types described above. These four inferences are demonstrated using simple version of DFX failures Bayesian network.

There are models in Bayesian networks, which are single-connected, multiple connected, or event looped networks. To solve more complex network, it is necessary to simplify it into Polytree. Polytree at most have one path between any pair of nodes; hence they are also referred to as singly-connected networks. Polytree algorithm can be applied to Polytrees. It is a basis for a more general class of algorithms, known as conditioning algorithms, which apply to arbitrary Bayesian networks.

1 INTRODUCTION

Statistics is the mathematical science which analyses collected data and then interpret and presented in a detail format for decision making process. Statistic is important field of study in human history because it helps to understand current situation and also to predict the future. Some areas which use statistical methods are the medical, biological and social sciences, economics, finance, marketing research, manufacturing and management, Meteorological centres , research institutes and many more.

Hypothesis testing is one key part of statistics and it is used to determine probability of given hypothesis is true for observed data. Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution. Hypothesis testing is commonly used in scientific research to justify the final conclusion.

Artificial Intelligence (AI) has become important in science and engineering fields in last few decades. Artificial Intelligent is the study of systems that act in a way that to any observer would appear to be intelligent. AI solves problems the way that human beings solve the complex problems. After World War II, lot of software applications and machines were created to perform difficult intellectual tasks. There are lot of other fields which contribute to development of AI and some of them are Philosophy, Mathematics, Computer Engineering, Control theory.

Artificial Intelligence is commonly used in many applications in 21st century. Fuzzy logic is used in washing machines, cars and elevators. Robots are used to perform difficult tasks. Computer games and intellectual games such as chess apply AI concepts. Reasoning with uncertainty is important for field of Artificial Intelligence and probability theories are used to uncertainties. Bayesian networks uses for diagnosing in computer systems, health and many other fields. There will be a comprehensive description about Bayesian network and calculations of different inferences in this thesis.

Data processing system at International Data Centre (IDC) processes data from seismic, infrasound and hydroaccustic stations. This data processing system is subjected to various changes and the thesis will describe how to validate these changes using hypothesis testing.

Detection and Feature Extraction (DFX) and Global Association (GA) applications are two processing subsystems of IDC. The primary functions of DFX are to make detections and to measure features from waveforms. DFX processes data from all three waveform technologies (seismic, hydroacoustic, and infrasonic). GA reads arrival and amplitude data for pre-defined time interval and forms set of associations and then these associations create an event.

This thesis describes how to implement a Bayesian diagnosis support tool for DFX and GA applications.

This master thesis will cover the following sub topics in four major chapters:

- Null hypothesis & Statistics: Average and Standard Deviation, null hypothesis and its related theories, P-value, Z-statistics and its examples, TYPE I and TYPE II Errors, Two-sample t-test.
- Reasoning with Uncertainty: Inductive, Abductive and Deductive Reasoning. Then, it will describe about Bayesian probability, Bayesian network and believe network. Markov blanket and d-separation, generic algorithm for Bayesian network queries.
- IDC processing system: seismology, CTBTO IMS network, automatic processing pipeline at IDC, DFX and GA application
- Automatic data processing validation and diagnosis: Null hypothesis for processing changes validation, Bayesian network as DFX and GA diagnosis supporting tool.

2 PROBLEM FORMULATION

Complex data processing system has many processing variables and database tables. These parameters are used to configure the system. Their values are subjected to changes and modifications. System operates face difficulties in validating these changes and come to conclusions within a short period. Hypothesis testing application will provide additional support for validating these changes.

These complex systems usually consist of many sub-systems and processing of these sub systems depend on inputs of other minor systems. In a case of a failure in sub-systems, it will take more time and effort to diagnosis causes of failures. Bayesian diagnosis supporting tool will provide additional assist to find out possible causes of failures and system operators can take necessary prompt actions using results of diagnosis tool.

3 TESTING HYPOTHESES

Statistic is the art and science of collecting and analyzing data and understanding the nature of variability. Mathematics, especially probability, governs the underlying theory, but statistics is driven by applications to real problems. (Chihara and Hesterbery, 2011)

3.1 Average and Standard deviation

3.1.1 Population and Sample

In analyzing data, we need to determine whether the data represents a population or a sample. The population must be fully defined so that those to be included and excluded are clearly spelt out. For example, all the earthquakes, which have magnitude greater than 5 and taken place in Asia last year.

A sample is a subset of a population, containing the objects or outcomes that are actually observed. A simple random sample (SRS) of size n consists of n items from the population and items are chosen in such a way that every set of n individuals has an equal chance to be the sample actually selected.

3.1.2 Mean and Standard Deviation

The mean or average is obtained by dividing the sum of observed values by the number of observations, n. It can be considered a good estimate for predicting subsequent data points. The formula for the mean is given below as equation:

$$\bar{X} = \frac{\sum_{i=1}^{i=n} X_i}{n}$$

The standard deviation gives an idea of how close it is the entire set of data to the average value. Data sets with a small standard deviation have tightly grouped, precise data. Data sets with large standard deviations have data spread out over a wide range of values. The formula for standard deviation is as follows:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{i=n}(X_i - \overline{X})^2}$$

3.2 Testing Hypothesis

Hypothesis testing is a method for testing hypothesis about a parameter in a population, using data measured in a sample.

The method of hypothesis testing can be summarized in four steps:

- Identify a hypothesis, which should be tested
- Select a criterion upon which that the claim being tested is true or not
- Select a random sample from the population and measure the sample mean
- Compare the observe in the sample to what expect to observe if the claim to be tested is true

3.2.1 Null and Alternative Hypothesis

Statistical problems involved a parameter θ, whose value is unknown but must lie in a certain parameter space Ω, that can partitioned into two disjoint subsets Ω_0 and Ω_1, and that the statistician must decide whether the unknown value of θ line in Ω_0 or in Ω_1.

H_0 is denoted the hypothesis that $\theta \in \Omega_0$ and H_1 denotes the hypothesis that $\theta \in \Omega_1$. Since the subsets Ω_0 and Ω_1 are disjoint and $\Omega_0 \cup \Omega_1 = \Omega$, exactly one of hypotheses H_0 and H_1 must be true. The statistician must decide whether to accept the hypothesis H_0 or to accept the hypothesis H_1. A problem of this type, in which there are only two possible decisions, is called a problem of *testing hypotheses*. If the statistician makes the wrong decision, he typically must suffer a certain loss or pay a certain cost. In many problems, he will have an opportunity to take some observations before he has to make his decision, and the observed values will provide him with information about the value of θ. A procedure for deciding whether to accept the hypothesis H_0 or to accept the hypothesis H_1 is called a test procedure or simply a test.

In most problems, however, the two hypotheses H_0 and H_1 are treated quite differently. To distinguish between them, the hypothesis H_0 is called the *null hypothesis* and the hypothesis H_1 is called the *alternative hypothesis*. One way of describing the decisions available to the statistician is that he may accept either H_0 and H_1. However, since there are only two possible decisions, accepting H_0 is equivalent to rejecting H_1, and accepting H_1 is equivalent to rejecting H_0.

(DeGroot, 1986)

The *null hypothesis*, denoted H_0, is a statement about a population parameter (population mean), that is assumed to be true.

$H_0 : \mu = 1.75$ Mean height of students in the university is 1.75m

The *alternative hypothesis,* denoted H_a, is a statement that directly contradicts a null hypothesis by stating that that the actual value of a population parameter is less than, greater than, or not equal to the value stated in the null hypothesis.

$H_a : \mu \neq 1.75$

3.2.2 Significance Level

The significance level is the criterion used for rejecting the null hypothesis in hypothesis testing. It refers to make a decision regarding the value stated in a null hypothesis. Traditional experimenters use either the 0.05 or 0.01 level, although the choice of levels is largely subjective. The lower the significance level, the more the data must diverge from the null hypothesis to be significant. Therefore, the 0.01 level is more conservative than the 0.05 level.

3.2.3 P-Value

When testing a null hypothesis against an alternative hypothesis using a dataset, the two hypotheses specify two statistical models for the process sample data. The alternative hypothesis may be true if the null hypothesis is false. The alternative hypothesis cannot be proved that it is true but it may be possible to demonstrate that the alternative is much more plausible than the null hypothesis given the data. This experiment is usually carried out using a probability and it is P-value and it will strength of the evidence against the null hypothesis in favour of the alternative hypothesis.

The P-value is often incorrectly interpreted as the probability that the null hypothesis is true. One can interpret "the probability that the null hypothesis is true" using subjective probability, a measure of one's belief that the null hypothesis is true. Then, calculate this subjective probability can be calculated by specifying a prior probability (subjective belief before looking at the data) that the null hypothesis is true, and then uses the data and the model to update one's subjective probability. This is called the Bayesian approach because Bayes' Theorem is used to update subjective probabilities to reflect new information.

When reporting a P-value to persons unfamiliar with statistics, it is often necessary to use descriptive language to indicate the strength of the evidence.

$P > 0.10$	No evidence against the null hypothesis. The data appear to be consistent with the null hypothesis.
$0.05 < P < 0.10$	Weak evidence against the null hypothesis in favour of the alternative
$0.01 < P < 0.05$	Moderate evidence against the null hypothesis in favour of the alternative.
$0.001 < P < 0.01$	Strong evidence against the null hypothesis in favour of the alternative.
$P < 0.001$	Very strong evidence against the null hypothesis in favour of the alternative.

Table 3.1 P value and its corresponding descriptive language

3.2.4 Z statistic

The test statistic for a one–independent sample z test is called the z statistic. The z-statistic converts any sampling distribution into a standard normal distribution. The z statistic is therefore a z transformation. The solution of the formula gives the number of standard deviations, or z-scores, that a sample mean falls above or below the population mean stated in the null hypothesis.

$$\text{Z-score} = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

$\overline{X}$ - sample mean

μ_0 - population mean

n - sample count

σ - Standard deviation of sample

3.2.5 One-Sided and Two-Sided Tests

When null hypothesis (H_0) specifies a single value for μ (sample mean), both tails contribute to the P-value, and the test is said to be a two-sided or two-tailed test.

When H_0 specifies only that μ is greater than or equal to, or less than or equal to a value, only one tail contributes to the P-value, and the test is called a one-sided or one-tailed test

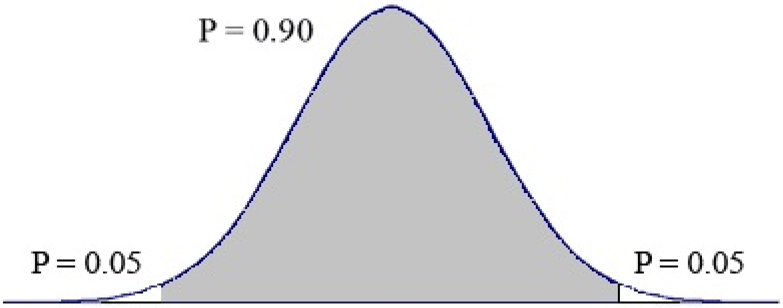


Fig. 3.1 probability distribution for two-sided test

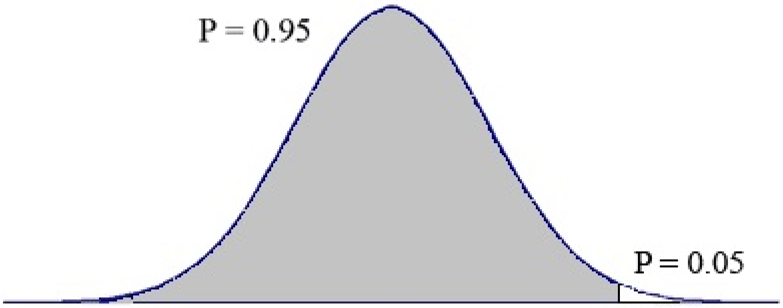


Fig. 3.2 probability distribution for one- sided test

<u>Simple example to illustrate above topics</u>

$H_0 : \mu \geq 1.75$ Mean height of students in the university is 1.75m

Alternative hypothesis; $H_a : \mu < 1.75$

Let’s assume that a random sample of students’ height has been taken with following values:

Sample size (n) = 100

Mean of Sample $(\overline{\mathrm{X}}) = 1.74$

Standard deviation $\sigma = 0.24$

Assumptions: Sample is normally distributed with given standard deviation

$$\text{Z-score} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

$$= \frac{1.74 - 1.75}{0.24/\sqrt{100}}$$

$$= -0.417$$

It is one-tailed test; P value is 0.3383

There is no evidence against null hypothesis. The data appear to be consistent with the null hypothesis.

3.3 TYPE I and TYPE II Errors

With and hypothesis test, there are two errors we can make – to reject H_0 when it holds, or to fail to reject it when it does not hold.

A Type I error occurs if we reject the null hypothesis when it is true. A Type II error occurs if we do not reject the null hypothesis when the alternative is true.

Type I Errors

To get an idea of types of errors in a hypothesis test, we look at this courtroom setting. Suppose John Doe is on trial for murder. In the United States, accused are considered “innocent until proven guilty,” and the proof must be “beyond a reasonable doubt.” This corresponds to

H_0 : John Doe is innocent

H_A : John Doe is guilty

Unless the evidence strongly shows otherwise, we accept the null hypothesis.

Jury Decision	**Truth**	
	Innocent	Guilty
Guilty	Type I error	Correct
Not Guilty	Correct	Type II error

Table 3.2 Type I and Type II error

Which error is more serious, convicting an innocent person (Type I) or freeing a guilty person (Type II)? Our justice system sidesteps this question; it holds that convicting an innocent person is bad, and the probability of a wrongful conviction must be small. The severity of a Type II error doesn't really enter the picture.

Similarly, in the classical approach to hypothesis testing, we do not adjust critical values to balance the two kinds of errors, taking into account their relative severity. Instead, we set thresholds to limit the probability of a Type I error to a pre-specified value. (Ex 5 %).

(Chihara and Hesterbery, 2011)

Increasing sample size is an obvious way to reduce both types of errors for either the justice system or a hypothesis test. An increase of sample size narrows the distribution because the distribution represents the average of the entire sample instead of just a single data point. In hypothesis testing the sample size is increased by collecting more data.

3.4 Two-Sample t-TEST for means

t-distribution is a family of continuous probability distributions that arises when estimating the mean of a normally distributed population in situations where the sample size is small and population standard deviation is unknown.

Let $X_1, X_2, X_3, \ldots\ldots, X_{n1} \sim N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, Y_3, \ldots\ldots, Y_{n2} \sim N(\mu_2, \sigma_2^2)$ be two independent random samples with sample means and standard deviation $\bar{X}$, S_1, $\bar{Y}$, S_2 respectively;

To test,

$$H_o : \mu_1 = \mu_2 \text{ versus } H_A : \mu_1 \neq \mu_2$$

We form the test statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

If the null hypothesis is true, then T has approximate a t distribution with degrees of freedom is given by equation :

$$v = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\left(S_1^2/n_1\right)^2/(n_1 - 1) + \left(S_2^2/n_2\right)^2/(n_2 - 1)}$$

The P- value is the probability that chance alone would produce a test statistic as extreme as or more extreme than the observed value if the null hypothesis is true.

<u>An example for illustrate two – sample t-test for means.</u>

Birth weights data for smoking and nonsmoking mothers in North Carolina

	Non-smoking mothers	Smoking mothers
Mean weight	3472g	3257g
Std. Deviation	479g	520g
No of Babies	898	111

Table 3.3 Sample details (Mean weight, Std. Deviation, No of Babies)

Is the observed mean difference in the mean weights of $\bar{x}_1 - \bar{x}_2 = 215g$ easily explained by chance, or is there a real difference in the mean weights of North Carolina babies born to non-smoking and smoking mothers ?

Let μ_1 and μ_2 denote true mean weight of babies born to non-smoking and smoking mothers, respectively. We consider the hypotheses

$$H_o : \ \mu_1 = \mu_2 \ \text{versus} \ \ H_A : \ \mu_1 \neq \mu_2$$

Assuming that the distribution of weights is normal for babies born to both non-smoking and smoking mothers, then the statistic has approximately a t-distribution

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

If the null hypothesis is true $(H_o) : \ \mu_1 - \mu_2 = 0$;

$$t = \frac{(3472 - 3257) - (0)}{\sqrt{479^2/898 + 520^2/111}} = \frac{215}{51.88} = 4.144$$

Degree of freedom is

$$v = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\left(S_1^2/n_1\right)^2/(n_1 - 1) + \left(S_2^2/n_2\right)^2/(n_2 - 1)}$$

$$v = \frac{\left(479^2/898 + 520^2/111\right)^2}{\left(479^2/898\right)^2/(898-1) + \left(520^2/111\right)^2/(111-1)} = 134.011$$

If the null hypothesis is true, then the chance of obtaining a statistic as extreme as 4.14 is

$$P(t \geq 4.14) = 0.00003$$

Thus, if there really is no difference in mean weights, then the samples we obtained are rare-random chance alone would give a test statistic that large, less than 3 out of a 100,000 times.

(Chihara and Hesterbery, 2011)

There is very strong evidence against the null hypothesis in favour of the alternative hypothesis. Thus, we conclude that babies born to non-somking mothers do weigh, on average, more than babies born to smoking mothers.

4 REASONING WITH UNCERTAINTY

The "reasoning with uncertainty" (or "reasoning under uncertainty") is a research field in Artificial Intelligence(AI) and it has been focused on the uncertainty of truth value. This is a large and active subfield of AI research. There are mainly three different approaches for representing uncertainty, are considered. Two of them represent two different forms of quantitative uncertainty; that is where we attempt to give numerical values expressing the degree to which we are uncertain about pieces of knowledge. At the third approach, truth maintenance systems, is just one example of a non-monotonic reasoning system, that is one where adding new items of knowledge may cause conclusions we had previously drawn to become invalid.

4.1 Reasoning

(ELA KUMAR (2008))

Reasoning, in simple terms means deriving conclusion from the available set of data and information. We are called intelligent when we are able to draw conclusions from the given information. Hence, for a human being to be intelligent, it is necessary to have ability to reason. However, in real world, there are many situations where we are required to draw conclusions from incomplete and uncertain evidences. For example, if the available information is:

Birds can fly.

Yamu is a bird.

From this information, an obvious conclusion would be that Yamu could fly. However, this conclusion is based on the most likely characteristics of birds. We often draw conclusions based on assumptions we make that are inclined towards most likely characteristics of the situation or object under consideration and also, our beliefs about real world situations. If some information is withdrawn or some new information is added, our conclusions would change. For example, in above mentioned case, if one or more information is added that Yamu was an Ostrich, our conclusion would be exactly opposite of what had been earlier. Our aim is to understand that we have to deal with many such situations in everyday life that is full of uncertainties. Such types of situations are called uncertain situations and reasoning with these situations is known as reasoning with uncertainty.

An uncertain situation requires representation of uncertain knowledge. Consider the following examples of certain and uncertain events:

Certain events:

- Earth revolves around Sun.
- The states of a chess game.

Uncertain events:

- If it cloudy, it will rain.
- If the weather is sunny, it will not rain.
- If a patient is vomiting, he is suffering from cholera.

The representation of uncertain knowledge requires attachment of an additional factor indicating the correctness of knowledge. This additional conceptual factor is called "degree of belief". The value of this factor varies between 0 and 1. It can take any fractional value in this range. Thus, the uncertain situation is represented by attaching a degree of belief factor e.g. in medical diagnosis, we observe some symptoms in the patient, but if those symptoms are present, still it cannot be guaranteed that a particular disease is present. Only with some degree of belief, it can be said that particular disease is present. The degree of belief is a conceptual factor indicating the degree of correctness of diagnosis. This belief factor is also related with probability theory. However, in the theory of probability, the total possible outcomes are defined. But in real world problems, there may be situations where even total outcomes are not defined.

In representing uncertain knowledge, we define certain terms as follows:

Evidence: It is the observations obtained in real world.

Belief: It is any meaningful and coherent expression that can be represented. Hence it can be true or false. Belief represents just observer's view about any incidence. At the time of defining the face, nothing can be said about the truthiness of belief.

Hypothesis: It is a justified belief that is supported by some evidence.

AI means building intelligent systems to solve real world problems. Intelligent systems provide solutions on the basis of facts and rules stored in the knowledge base. These fact and rules are often incomplete and hence uncertain. AI systems are required to reason with this uncertain knowledge or information. Thus, reasoning is the process by which we use available knowledge, in whatever quantity or of whatever quality, to draw conclusions or to infer about a new event. Without this ability, the AI system will simply be considered as "information system" giving answers based on lookup table. The three basic types of reasoning methods are:

- Inductive reasoning
- Abductive reasoning
- Deductive reasoning

4.1.1 Inductive Reasoning

Inductive reasoning is based on the generalizations of the previous experiences about the problem. e.g. a person, how haven't seen any black pigeo**n**s, can come to conclusion such that

“ All pigeo**n**s are completely white ”

Inductive reasoning can be risky because conclusion may be not correct. However, inductive reasoning is used in machine learning AI applications.

4.1.2 Abductive reasoning

In the abductive reasoning, we will come to final conclusion, only looking after back through the chain of events to perform reasoning. For example;

I went to market yesterday.

There are potatoes at my home.

Conclusion: I bought potatoes from market yesterday

We believe that most plausible conclusion is also the correct one in Abductive reasoning. As a result, conclusion may not necessary true for every interpretation.

4.1.3 Deductive reasoning

Deductive reasoning originates from the philosophy and mathematics and is the most obvious form of reasoning. It is worked on the standard logic. It is kind of explicit reasoning.

All shops are closed on holidays.

Sunday is a holiday.

Conclusion: All shops are closed on Sunday.

4.2 Dealing with Uncertain Situations

(ELA KUMAR (2008))

For certain situations, where the knowledge base stores consistent information, all new knowledge that is added to it is bound to be consistent with the previous knowledge. Such type of reasoning is known as *monotonic reasoning* and in systems using this, the size of knowledge base always increases. In monotonic reasoning, whenever some conclusion is drawn as true, it remains true under all circumstances. However, in real life, all inferences do not necessarily be considered correct under all circumstances.

What does it mean in simple terms can be viewed as logical reasoning cannot be a realistic presentation of real world. On the contrary, intelligent beings are required to make decisions and function in a world full of uncertainties. However, while reasoning with uncertain knowledge conclusion is drawn based on what is most likely to be true. Following approaches are followed for this type of reasoning:

- Nonmonotonic reasoning: In this type of reasoning, the rules of inference are extended to make it possible to reason with incomplete information. The systems using this method show the property that at any given point of time, a statement is either believed to be true, believed to be not true or not believed to be true or not true.
- Probabilistic reasoning: These are also known as statistical methods of reasoning. In these method, the results are not in the form of TRUE or FALSE but some numeric value is assigned to them that is a measure of certainty of those events to be true under given circumstances. Some of the methods using probabilistic reasoning methods are:
  - Bayesian belief networks
  - Reasoning with certainty factors
  - Dempster Shaffer theory
  - Fuzzy reasoning

4.3 Bayesian Reasoning

(ELA KUMAR (2008))

Bayesian reasoning is a kind of probabilistic reasoning, introduced by Thomas Bayes in eighteenth century that is based on formal probability theory and is used in several areas of research including pattern recognition and classification. Assuming a random sampling of events, Bayesian theory devises the calculation of complex probabilities from previously known results. Probability is of two types

- *Prior Probability*: This probability is also popularly known as unconditional probability. It is probability assigned to an event in the absence of knowledge supporting its occurrence or absence. i.e. the probability of the event prior to any evidence supporting or negating the occurrence of that particular event. The prior probability of an event is represented as P(event)
- Posterior Probability: This type of probability is also known as conditional probability. It is probability of an event after evidence, i.e. the probability when some evidences supporting or negating the outcome are known. Posterior probability is symbolized as P(event | evidence).

Bayes Theorem is based on the theory of conditional probability. Let's explain the concept by an example of medical diagnosis.

p = number of sick persons or number of patients.
n = total number of persons in the domain
d = set of persons actually having disease d
s = set of persons having symptoms of disease d
d ∩ s = set of persons actually having diseases d and symptoms s both

Unconditional probability of having disease:

$$P(d) = {}^{p}/_{n}$$

Thus, the conditional probability of persons having disease d with symptoms s would be represented as:

$$P(d \mid s) = {}^{d \,\cap\, s}/_{s}$$

Let's consider that total 1000 people are present in the domain out of which 100 persons are sick. Out of 100 patients, 60 persons had high fever that this symptom of malaria. Further investigation found that, only 20 persons had malaria.

Conditional probability of persons having malaria:

$$P(\,d) = {}^{p}\!/_{n} = \frac{100}{1000} = 0.10$$

Unconditional probability of persons having malaria:

$$P(\,d \mid s) = {}^{d\,\cap\,s}\!/_{s} = \frac{20}{60} = 0.33$$

It can be calculated also using probabilities values:

$$P(\,d \mid s) = {}^{P(d\,\cap\,s)}\!\Big/_{P(s)} \quad ----(1)$$

$$P(\,d \mid s) = {}^{(\frac{20}{1000})}\!\Big/_{(60/1000)} = 0.33$$

We can have an equivalent relationship for conditional probability of persons having symptoms s with disease d

$$P(\,d \mid s) = {}^{P(d\,\cap\,s)}\!\Big/_{P(s)} \quad ----(2)$$

Substituting this result in the equation for (1) and (2);

$$P(\,d \mid s) = {}^{P(\,s \mid d) * P(d)}\!\Big/_{P(s)}$$

This equation is known as Bayes theorem.

We present here the most important finding of probability theory, the general form of Bayes theorem. Taking reference of the above discussion, the disease would now be called hypothesis (H) and symptom would be called evidence (E). Substituting these notations;

$$P(\,H \mid E) = {}^{P(\,E \mid H) * P(H)}\!\Big/_{P(E)}$$

This is most simple view of Bayes theorem. It can be applied assuming real world phenomena as simple and straight forward. However, real world situations are complex and tedious where we have to deal with multiple hypothesis and multiple evidences. These set of exhaustive and mutually exclusive hypotheses can be represented as H_j (j =1 to n, n is number of hypotheses)

Now, if we generalize the Bayes theorem for multiple hypothesis and multiple evidences:

$$P(\, H_i \mid E) = \frac{P(\, E \mid H_i\,) * P(H_i)}{\sum_{j=1}^{n} P(\, E \mid H_i\,) * P(H_i)}$$

This is generalized version of Bayes theorem, where

$P(\, H_i \mid E\,)$: the probability that H_i is true given evidence E

$P(\, H_i\,)$: the probability that H_i is true overall

$P(\, E \mid H_i\,)$: the probability of observing E when H_i is true n is the number of hypotheses

4.4 Bayesian and Believe Network

Bayes theorem can help in reasoning, especially when case simply contains single disease and single symptom, because in these cases many numbers are not needed on the right-hand side in the equation of Bayes theorem. In real world situations, the Bayes theorem in this case would like:

$$P(d \mid s_1 \,\&\, s_2 \,\&\, \ldots\ldots \,\&\, s_n) = (P\,(s_1 \,\&s_2 \,\&\, \ldots\ldots \&s_n \;\mid d)\,) * P(d)/\; P(s_1 \,\&s_2 \,\&\, \ldots\ldots \&s_n)$$

However, problem starts when reasoning is required to be done about possible disease from among set of multiple diseases and multiple symptoms. Let us consider multiple diseases d_m from set of diseases D and multiple symptoms S_n from the set of symptoms S. In the case, we would require (m * n) posterior probabilities and (m + n) prior probabilities. In actual real world situations, we hardly have single disease and single symptom.

Now suppose there are m number of diseases in D and if we want to use Bayes theorem to calculate the probability of a patient having a particular disease out of possible m, if he has n number of symptoms, we would require about ($m * n^2$) conditional probabilities + n^2 symptoms probabilities + m disease probabilities to complete right-hand side of the Bayes theorem.

(ELA KUMAR (2008))

As described above, it is required large number of probabilities due to combining effect of hypotheses to use Bayesian theorem and calculations also become difficult in these situations. Human would use their heuristics and intuition to separate evidence and hypotheses that are not dependent.

In combine probability, there are possible two scenarios depending on events are independent or not. Combined probability of two events occurring if they are independent;

$P(A\&B) = P(A) * P(B)$

If they are not independent, their combined probability is ;

$P(A\&B) = P(A) * P(B|A)$

With knowledge of combine probabilities, we can draw a graphical model to show interdependence of various parameters. In these graphical diagrams, propositional variables are represented as nodes and the causal influences or dependencies among nodes are represented by arcs. These graphical diagrams are called as Bayesian networks. Bayesian belief network reduce many constraints of the full Bayesian model. It is not necessary to build large joint probability tables in which the probabilities of all possible combinations of events and evidences are listed. After analysing all events, experts can collect obtain probabilities of events which are only dependent events.

Let’s look at Bayesian network example for water sprinkler-rain problem from Pearl (1998) in figure 4.1

In this problem, grass wet being depends upon water either from rain or from sprinkler. The water from sprinkler system or from rain depends on cloudy weather condition and they are not independent. This Bayesian network is an example for multiply connected belief network. To convert this directed acyclic graph to undirected acyclic graph, it is required a mechanism to transfer probabilities correctly. Clique triangulation method is one of algorithms for doing such computations.

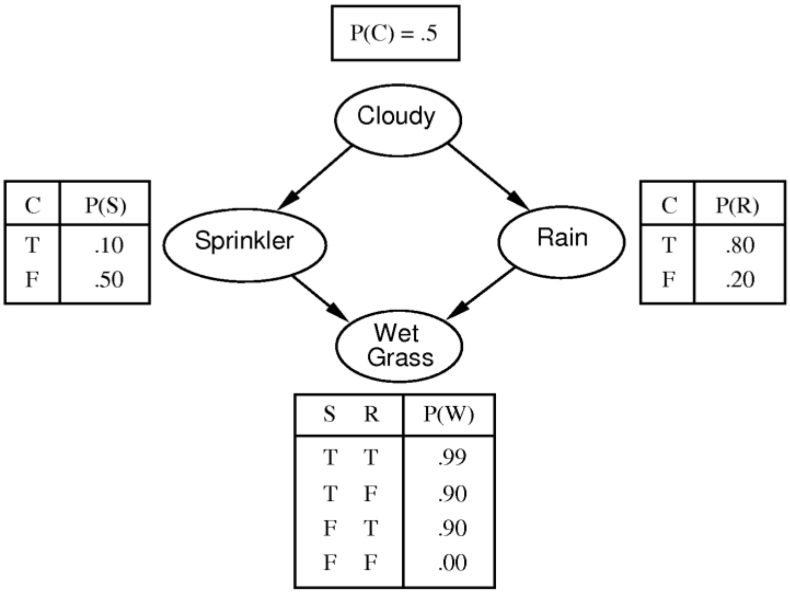


Fig. 4.1 Bayesian probabilistic network in multiply connected, probability dependencies are located next to each node

The algorithm to build a junction tree:

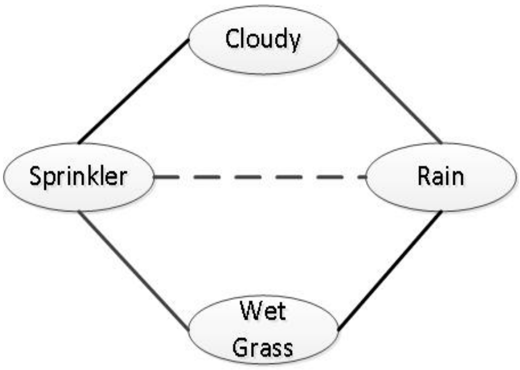


Fig. 4.2 Triangulated structure from Bayesian network (ELA KUMAR (2008))

1. Make all directed links are replaced with undirected links in the belief network
2. Draw links between all parents for any node which doesn't include. E.g. dashed link between sprinkler and rain in the figure

3. Check the network and ensure that all the cycles have only three nodes. If not, add further links to reduce all cycles to have maximum three nodes at each cycle. This process is called triangulation.
4. Create a Junction tree from the resulting triangulated structure. This is done by finding the maximal cliques. The variables in these cliques are put into junctions and the junction tree is created by connecting any two junctions that share at least one variable. e.g. "R, W" rectangular box reflect the variables which share above and below nodes.

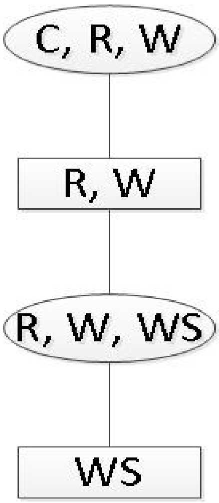


Fig. 4.3 A junction tree for Bayesian probabilistic network (ELA KUMAR (2008))

4.5 Markov blanket

Let us consider a set of random variables {X_1 ,….., X_n}. As before, by a model M we mean a set of edges between nodes that represent the random variables {X_1,…….,X_n} or {X_1[1],…..,X_n[1],X_1[2],….,X_n[2]} for the Bayesian Network (BN) or Dynamic Bayesian Network (DBN) case, respectively. For the DBN case, we allow only connections between time-slices, leaving the intra-graph empty. We start by introducing the parents set for node X_i.

$$Par(X_i) = \{ X_j : (X_j, X_i) \in M \}$$

The edges which start in nodes taken from Par(X_i) and end in node X_i surely belong to the common set of relationships, because they directly explain how the node X_i is influenced. The node X_i is conditionally independent of any other node conditioned on the parents set of X_i.

The explanation of impact on the behaviour of the node X_i is complete. Although the dependency goes even further to parents sets of X_i's parents and so on, the impact gets less and less important with every step so we narrow our interest down to the direct dependencies set. Beside edges that explain somehow the behaviour of the node X_i we can also look in the opposite direction and ask how the node X_i influences other nodes in the structure. These dependencies can be expressed as the set of edges which start at the node X_i and end in a node taken from children set of node X_i. We define the children set of node X_i as follows.

$$Ch(X_i) = \{ X_j : (X_i, X_j) \in M \}$$

Now we formed the set of edges that somehow make up a coherent system that describes the behaviour of the node X_i and its impact on other nodes. The Markov Blanket in a BN for node X_i which we denote by MB(X_i) is a set of nodes composed of X_i's parents, its children and parents of its children. Formally the definition of Markov Blanket in a BN, or more general in a graph, is as follows.

(Tomasz Ku laga, 2006)

$$MB(X_i) = Par\ (X_i)\ \cup Ch(X_i)\ \cup \bigcup_{Y\ \in Ch(X_i)} Par(Y)$$

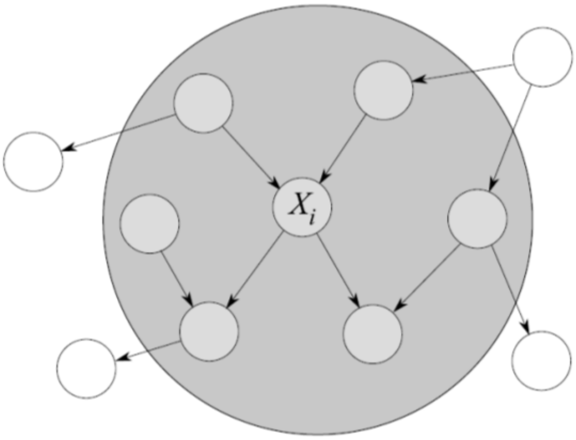


Fig. 4.4 Markov Blanket

4.5.1 d-separation

d-separation is a graphical test of independence between variables in a directed acyclic graph. This is a very useful tool for working with Bayesian networks. Given two sets of variables A and B, we test if they are independent conditioned on a set Z of variables by checking all paths between each variable in A and each variable in B. We say that A is independent of B given Z $(A \perp\!\!\!\perp B \mid Z)$ is all paths between each variable in A and B are closed when we condition on (or in other words observe) Z. If any path is open, we cannot claim independence but also cannot claim dependence. We would have to examine the conditional probability tables to verify the independence claims if there is no d-separation.

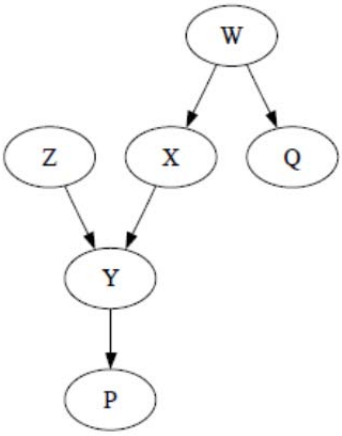


Fig. 4.5 directed acyclic graph

Using above graph, we can express the following statements:

- $(Q \perp\!\!\!\perp X, Y, Z, P \mid W)$: $Q \rightarrow W \rightarrow X$ is a divergent path that is closed since we set condition on W.
- $(Z \perp\!\!\!\perp X, W, Q \mid \emptyset)$: $Z \rightarrow Y \leftarrow X$ is a closed convergent path since we do not condition on Y or it's descendent P
- $(Z, Y, P \perp\!\!\!\perp W, Q \mid X)$: $W \rightarrow X \rightarrow Y$ is closed sequential path since we condition on X

4.6 Algorithm for Bayesian Network Queries

There are mainly four different types of querying in Bayesian network.

- Diagnostic inference
- Causal inference
- Inter-causal inference
- Mixed inference

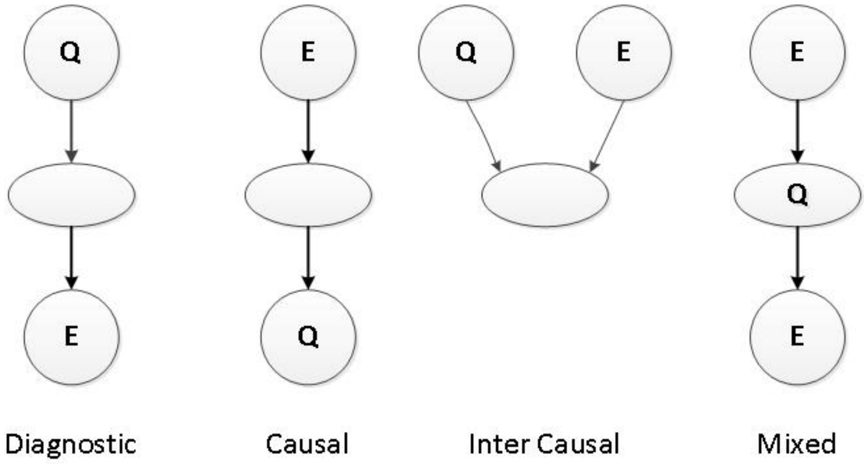


Fig. 4.6 Types of reasoning (Diagnostic, Causal, Inter-Causal, Mixed)

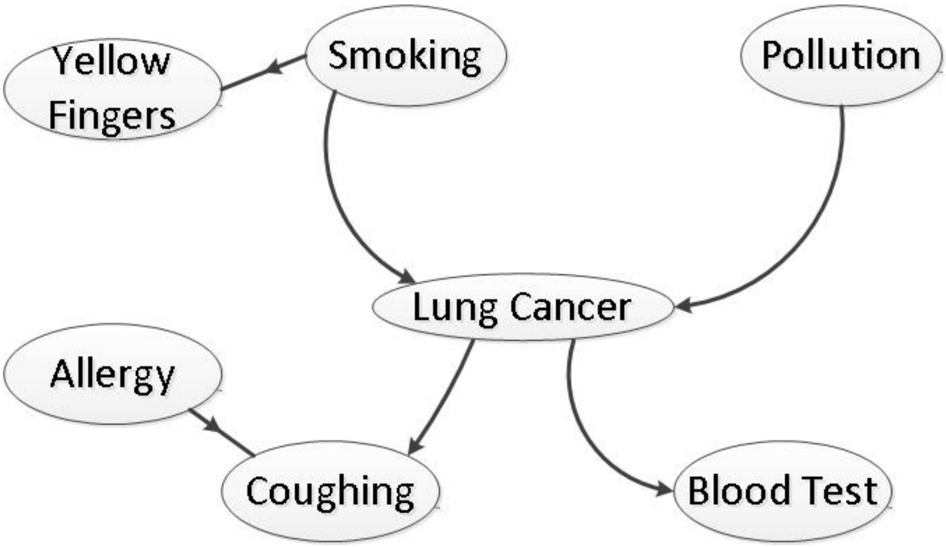


Fig. 4.7 Simple Bayesian network for Lung Cancer diagnosis

Using above Bayesian network (Fig 4.7), we can explain above four inferences for singly connected Bayesian network.

Diagnostic inference: Reasoning from symptoms to cause, such as when a doctor observes blood test report and then updates doctor belief about lung cancer and whether the patient is a Smoker. In above diagram, it goes opposite direction to arrows.

Causal inference: The patient may tell the doctor that he is a smoker; even before any symptoms have been assessed, the physician knows this will increase the chances of the patient having lung cancer. It will increase possibility of having symptom such as coughing.

Inter-Causal inference: It can be represented by a v-structure in the BN. In above simple Bayesian network, smoking and pollution has common effect and that is lung cancer. Suppose, the patient has lung cancer, then it will increase probability of he is smoker or he lives in a polluted area. If we found that he is a smoker, then it decrease the probability of he lives in a polluted area.

Mixed inference: It is combination of one or more above inferences. For example, the probability of a patient having lung cancer, while he lives in a polluted area and coughing. In this chapter, we will create generic algorithm for this case.

The 6.2.2 chapter describes to how to calculate above inferences using DFX failures diagnosis Bayesian network.

Mixed inference pattern is more general inference and we are going to create an algorithm for it in this chapter. All other inferences are special cases of Mixed inference.

Let’s consider the following belief network which is poly-tree. In the poly-tree, there is at least one undirected path between any two nodes. In this poly-tree, X is query variable and evidence are given by E.

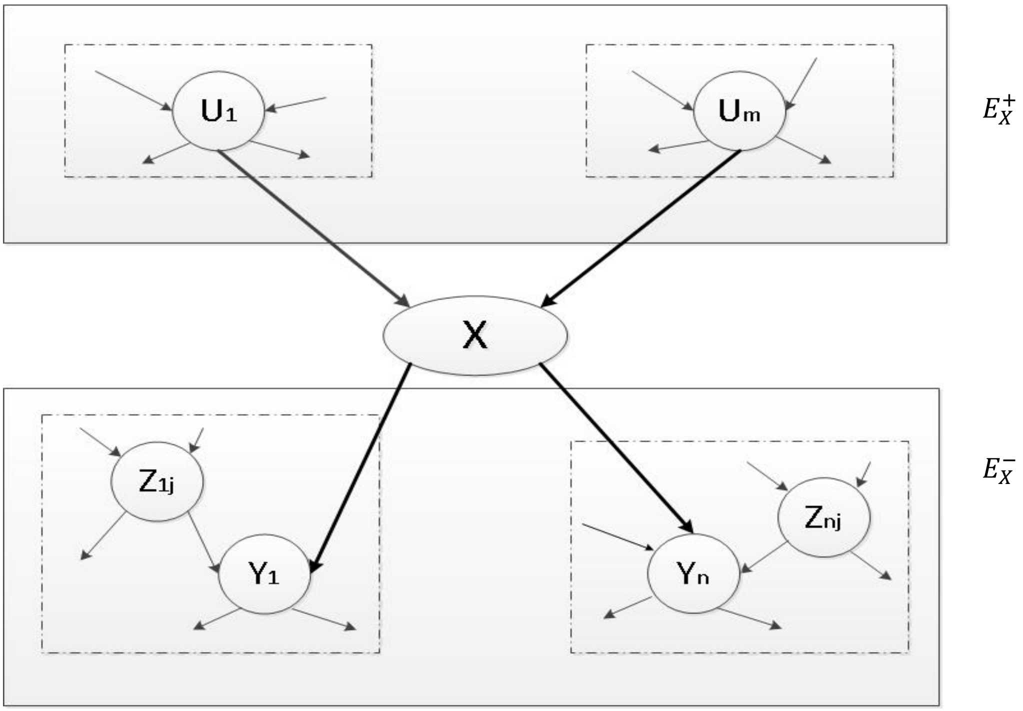


Fig. 4.8 Poly-tree Belief network X is query variable and Ex+ and Ex- evidence nodes

U_1,…., U_m are parent nodes of X

Y_1,…., Y_m are children nodes of X

E_X^+ are evidence nodes which are causal support for X node

E_X^- are evidence nodes which are evidential support for X node

Computation of P(X | E)

$$P(X \mid E) = P(X \mid E_X^-, E_X^+) \quad ------(1)$$

$$= \frac{P(E_X^- \mid X, E_X^+)\, P(X \mid E_X^+)}{P(E_X^- \mid E_X^+)}$$

Since X is d-separated E_X^+ from E_X^- ,we can use conditional independence to simply the first term in the numerator.

$$= \frac{P(E_X^- \mid X)\, P(X \mid E_X^+)}{P(E_X^- \mid E_X^+)}$$

Let’s take denominator as constant (α)

$$= \alpha\ P(E_X^- \mid X)\, P(X \mid E_X^+) \quad ------- (2)$$

Both of them $P(E_X^- \mid X)$ & $P(X \mid E_X^+)$ are similar to causal terms

Computation of $P(X \mid E_X^+)$ from equation (2):

Let U be the vector parents U_1, U_2, ……, Um and all of them are parents of X as shown in Fig 4.8.

$$P(X \mid E_X^+) = \sum_U P(X \mid U, E_X^+)\, P(U \mid E_X^+)$$

In the above poly-tree, U_i nodes are d-separates from X node. As result of that, we can simplifies to $(X|U)$:

$$P(X|\, E_X^+) \;=\; \sum_U P(X|U)\; P(U|\, E_X^+)$$

As seen in the above ploy tree, each U_i connects other only thru X. So that; E_X^+ d-separates each U_i from the others. The probability of a conjunction of independent variables is equal to the product of their individual probabilities.

$$P(X|\, E_X^+) \;=\; \sum_U P(X|U) \prod_i P(U_i|\, E_X^+)$$

The last term can be simplified by partitioning E_X^+ into $E_{U_1\backslash X}, \dots\dots, E_{U_m\backslash X}$ and noting that $E_{U_i\backslash X}$ d-separates U_i from all the other evidence in E_X^+

$E_{U_i\backslash X}$ are evidence nodes which are connected to node U_i except thru the path from X

$$P(X|\, E_X^+) \;=\; \sum_U P(X|U) \prod_i P(U_i|\, E_{U_i\backslash X})$$

- $P(X|U)$ is a lookup in the conditional probability table of X
- $P(U_i|\, E_{U_i\backslash X})$ is a recursive (smaller) sub-problem

Now, we have simplified second term of the equation (2) to above equation.

Computation of $P(E_X^- \mid X)$ from equation (2):

Let Z_i be the parents of Y_i other than X_i, and let Z_i be an assignment of values to parents. The evidence in each Y_i box is conditionally independent of the others given X

$$P(E_X^-|\, X) \;=\; \prod_i P(\, E_{Y_i\backslash X} \mid X\,)$$

Averaging over Y_i and Z_i yields:

$$P(E_X^- \mid X) = \prod_i \sum_{y_i} \sum_{z_i} P(\, E_{Y_i \backslash X} \mid X, y_i, z_i\,)\; P(\; y_i, z_i \mid X\,)$$

$E_{Y_i \backslash X}^+$ are evidence nodes which are connected to node Y_i except thru the parents of X

Breaking $E_{Y_i \backslash X}$ into the two independent components $E_{Y_i}^-$ and $E_{Y_i \backslash X}^+$

$$= \prod_i \sum_{y_i} \sum_{z_i} P(\, E_{Y_i}^- \mid X, y_i, z_i\,)\, P(\, E_{Y_i \backslash X}^+ \mid X, y_i, z_i\,)\; P(\; y_i, z_i \mid X\,)$$

$E_{Y_i}^-$ is independent of X and z_i given y_i and $E_{Y_i \backslash X}^+$ is independent of X and y_i

$$= \prod_i \sum_{y_i} P(\, E_{Y_i}^- \mid y_i\,) \sum_{z_i} P(\, E_{Y_i \backslash X}^+ \mid z_i\,)\, P(\; y_i, z_i \mid X\,)$$

Apply Bayes' rule to $P(\, E_{Y_i \backslash X}^+ \mid z_i\,)$

$$= \prod_i \sum_{y_i} P(\, E_{Y_i}^- \mid y_i\,) \sum_{z_i} \frac{P(\,z_i \mid E_{Y_i \backslash X}^+\,)P(E_{Y_i \backslash X}^+)}{P(z_i)}\; P(\, y_i, z_i \mid X\,)$$

Rewriting the conjunction of y_i and z_i

$$= \prod_i \sum_{y_i} P(E_{Y_i}^- \mid y_i\,) \sum_{z_i} \frac{P(\,z_i \mid E_{Y_i \backslash X}^+\,)P(E_{Y_i \backslash X}^+)}{P(z_i)} P(y_i \mid X, z_i\,)P(z_i \mid X)$$

$P(z_i \mid X) = P(z_i)$ because Z and X are d-separated. Also $P\left(E^{+}_{Y_i\setminus X}\right)$ is a constant

$$= \prod_{i} \sum_{y_i} P(E^{-}_{Y_i} \mid y_i) \sum_{z_i} \beta_i P(z_i \mid E^{+}_{Y_i\setminus X})\, P(y_i \mid X, z_i)$$

The parent of y_i (Z_{ij}) are independent of each other

We also combine the β_i into β

$$\mathrm{P}(\mathrm{E}^{-}_{\mathrm{X}} \mid \mathrm{X}) = \beta \prod_{\mathrm{i}} \sum_{\mathrm{y_i}} \mathrm{P}(\mathrm{E}^{-}_{\mathrm{Y_i}} \mid \mathrm{y_i}) \sum_{\mathrm{z_i}} \mathrm{P}(\mathrm{y_i} \mid \mathrm{X}, \mathrm{z_i}) \prod_{j} P(Z_{ij} \mid E_{Z_{ij}\setminus Y_i})$$

- $\mathrm{P}(\mathrm{E}^{-}_{\mathrm{Y_i}} \mid \mathrm{y_i})$ is a recursive instance of $P(E^{-}_{X} \mid X)$
- $\mathrm{P}(\mathrm{y_i} \mid \mathrm{X}, \mathrm{z_i})$ is a conditional probability table entry of Y_i
- $P(Z_{ij} \mid E_{Z_{ij}\setminus Y_i})$ is recursive sub-instance of the P(X | E) calculation

(Prof. P. Dasgupta (2008))

Now, we have simplified first term of the equation (2).If we use only Bayesian analysis, it will be complex for larger Bayesian network.

5 IDC DATA PROCESSING SYSTEM

IDC software acquires time-series data from stations of International Monitoring System. The data are passed through a number of automatic and interactive analysis stages, which culminate in the estimation of location and in the origin time of events such as earthquakes and volcanic eruptions in the earth, including its oceans and atmosphere. The automatic processing pipeline processes data through the following computer software components such as Stations Processing, Network Processing, Post-location Processing, Event Screening, Time-series Tools, Time-series Libraries.

In this master thesis, only the following two software components were considered in the examples.

- Station Processing

  This software scans data from individual time-series stations for characteristic changes in the wave forms (detections of onsets) and characterizes such onsets (feature extraction). The software then classifies the detections as arrivals in terms of phase type.

- Network Processing

  This software combines arrivals from several stations originating from one event and infers the location and time of its origin.

Station Processing consists of two main configurable software components, which are Detection and Feature Extraction (DFX) and Station Processing (StaPro) software items

5.1 Seismology

Every day there are about fifty earthquakes worldwide that are strong enough to be felt locally, and every few days an earthquake occurs that is capable of damaging structures. Each event radiates seismic waves that travel throughout Earth, and several earthquakes per day produces distant ground motion that, although too weak to be felt, are readily detected with modern instruments anywhere on the global. Seismology is the science that studies these waves and what they tell us about the structure of and the physics of earthquakes. (Shearer, 2009)

5.1.1 Body Waves

In 1830 Poisson used the equations of motion and elastic constitutive laws to show that two fundamental types of waves propagate through the interior of homogeneous solids: P waves (compressional waves involving volumetric disturbances, and directly analogous to sound waves in fluids) and S waves (shear waves with only shearing deformation and no volume change, which can therefore not propagate in fluids). The sense of particle motions relative to the direction of propagation for P- and S- waves disturbances is shown in Figure 5.1. These two types of motion are called body waves, because they traverse the interior of the medium. P (primary) waves travel faster than S (secondary) waves and are thus the first motion to be detected from any source in an elastic solid. (Thorne Lay & Terry C. Wallace(1995))

5.1.2 Surface Waves

In 1887 Lord Rayleigh demonstrated the existence of additional solutions of elastic equations of motion for bodies with free surfaces. These are Rayleigh waves, involving wave motions confined to and propagating along surface of the body. By 1911 a second type of surface-wave motion, produced in a bounded body with layered material properties, was characterized by Love and is hence called Love wave. Rayleigh and Love waves are surface waves result from the interaction of P and S waves with the boundary conditions on the body. The sense of particle motions for these surface waves is indicated on Figure 5.1. Body and surface waves are influenced by changes in material properties with depth, such as the existence of internal boundaries in the Earth that can reflect energy. These interactions can be quantitatively analysed by solving boundary-value problems, and they are expressed in terms of reflection and transmission coefficients. (Thorne Lay & Terry C. Wallace(1995))

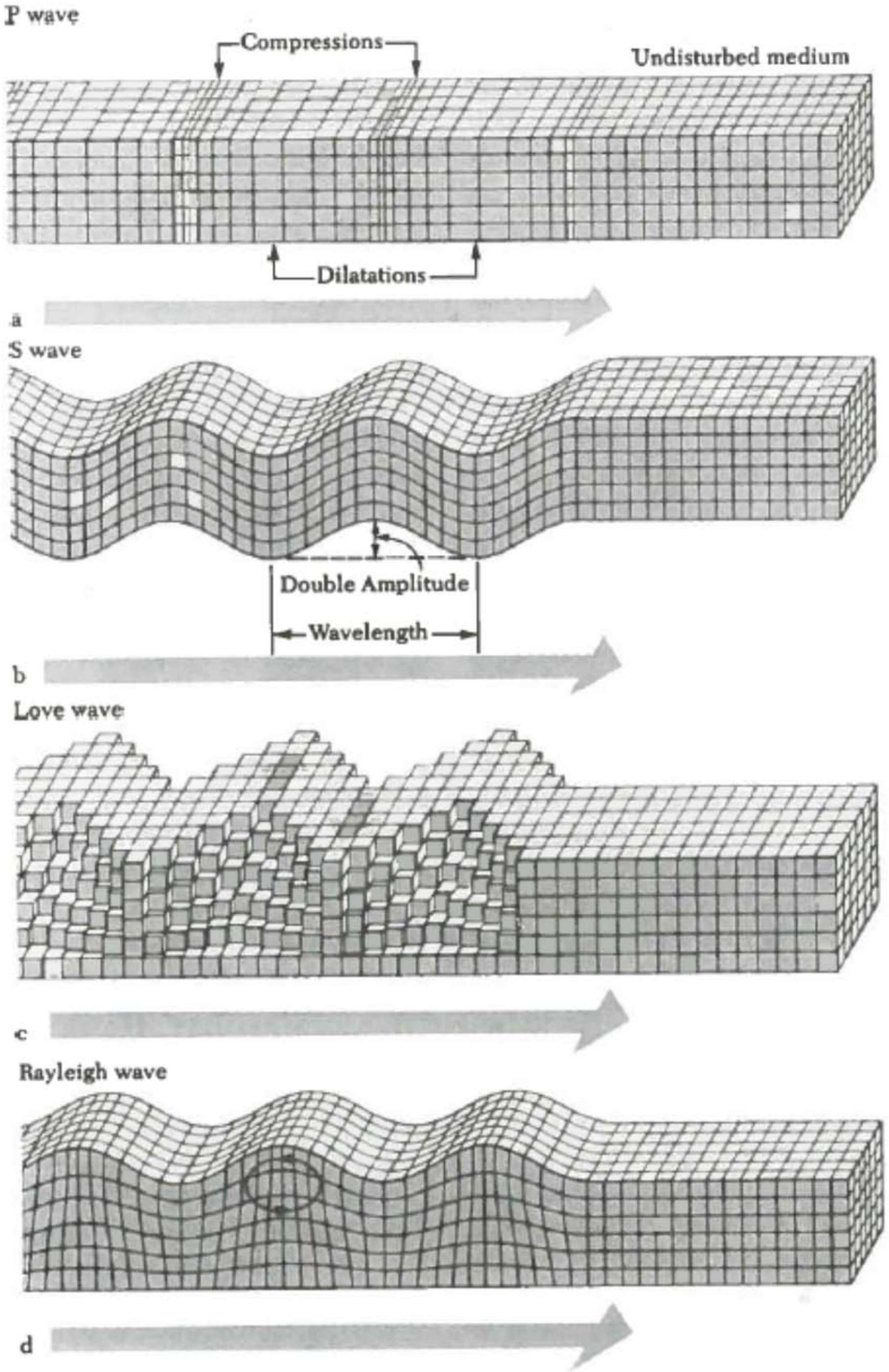


Fig. 5.1 Schematic of the sense of particle motions during passage of the two fundamental elastic body waves. (a) P and (b) S waves, as well as the two surface waves in the Earth. (c) Love and (d) Rayleigh waves. The waves are all propagating from left to right, with the surface of initial particle motion corresponding to the wave front. The relative velocity of each wave type decreases from top to bottom.

(Thorne Lay & Terry C. Wallace(1995)

5.1.3 Whole Earth phases

Here the main layers are the mantle, the fluid outer core, and the solid inner core. P- and S-wave legs in the mantle and core are labelled as follows:

P – P wave in the mantle

K – P wave in the outer core

I – P wave in the inner core

S – S wave in the mantle

J – S wave in the mantle

c – reflection off the core-mantle boundary (CMB)

i – reflection off the inner-core boundary (ICB)

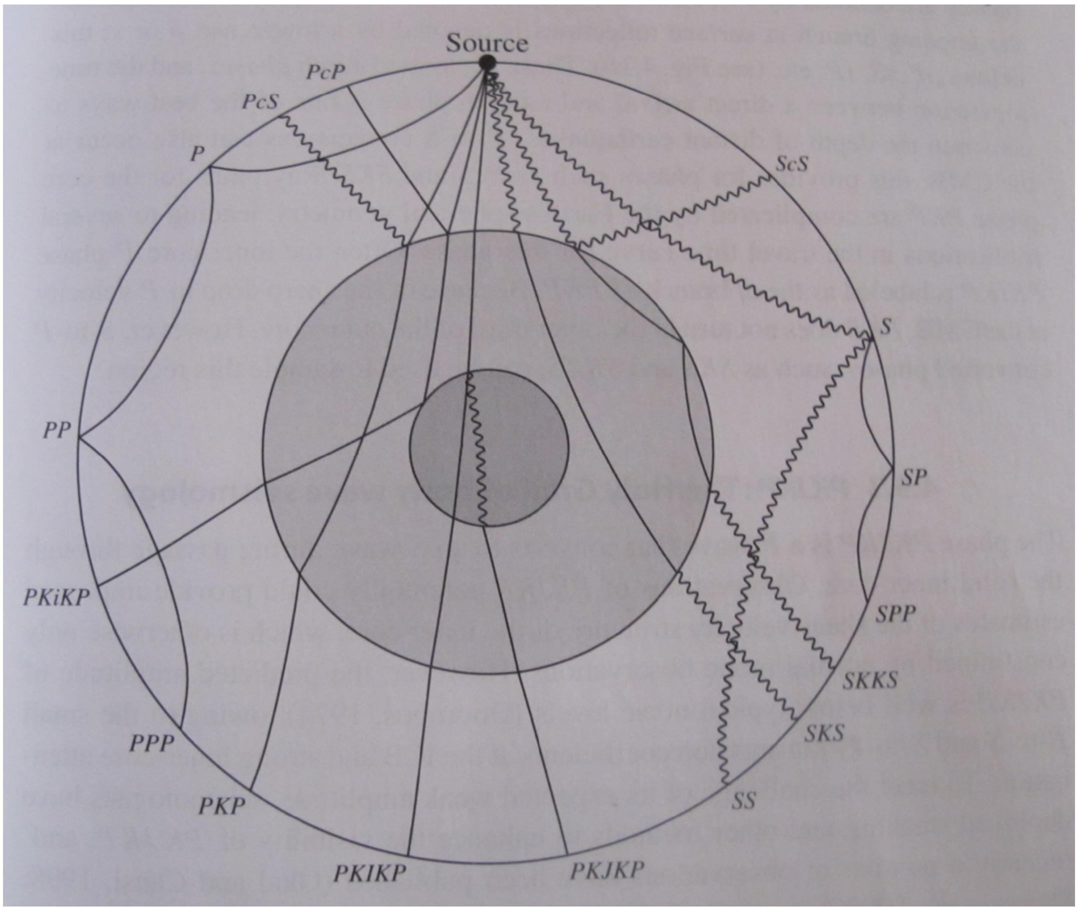


Fig. 5.2 Global seismic ray paths and phase names, computed for the PREM velocity model. P waves are shown as solid lines, S waves as wiggly lines. The different shades indicate the inner core, the outer core, and the mantle (Shearer, 2009).

For P and S waves in the whole earth, the above abbreviations apply and stand for successive segments of the ray path from source to receiver. Some examples of these rays, paths and their names are shown in Figure

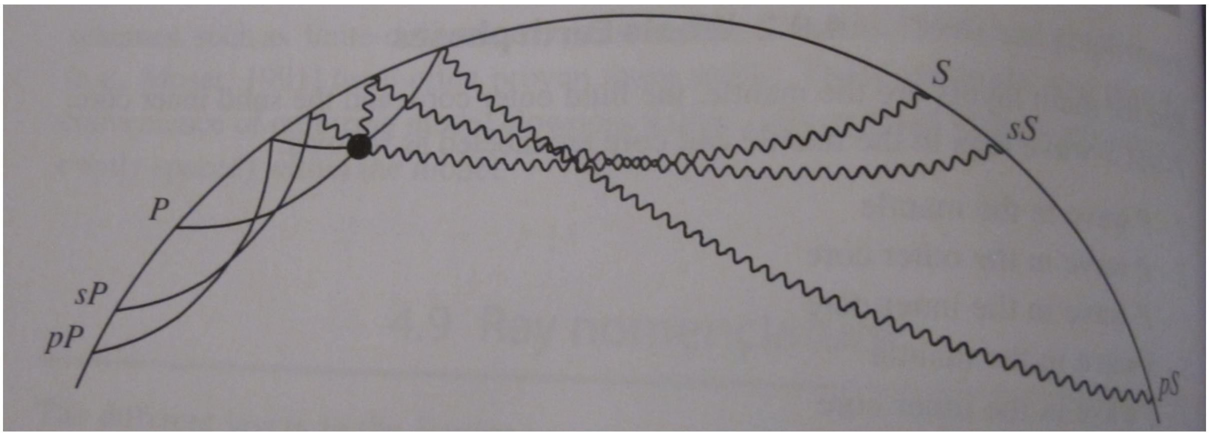


Fig. 5.3 Deep earthquakes generate surface-reflected arrivals, termed depth phases, with the up going leg from the source labelled with a lower-case p or s. Ray paths plotted here are for an earthquake at 650 km depth, using the PERM velocity model.

(Shearer, 2009)

5.1.4 PREM Model

For many years the most widely used 1-D model of Earth's seismic velocities has been the preliminary Reference Earth Model (PREM) of Dziewonksi and Anderson (1981). This model was designed to fit a variety of different data set, including free oscillation centre frequency measurements, surface wave dispersion observations, travel time data for a number of body-wave phases, and basic astronomical data (Earth's radius, mass and movement of inertia). In addition to profiling the *P* and *S* velocities, PERM specifies density and attenuation as functions of depth. Although these parameters are known less precisely than the seismic velocities, including them is important because it makes the model complete and suitable for use as a reference to compute synthetic seismograms without requiring additional assumptions. In order to simultaneously fit Love and Rayleigh wave observations, PRERM is transversely isotropic between 80 to 220 km depth in the upper mantle. This is a spherically symmetric form of anisotropy in which *SH* and *SV* waves travel a different speeds.

(Shearer, 2009)

5.2 CTBTO IMS Network

The International Monitoring System (IMS) of CTBTO currently consist of 337 facilities (Fig 5.4) worldwide to monitor the planet for signs of any natural and man-made events. These facilities contain seismometers, which is an instrument that converts ground motion into electric voltage. Different types of seismometers are used at seismic stations. The IMS uses the following three types of technologies to receive continuous data.

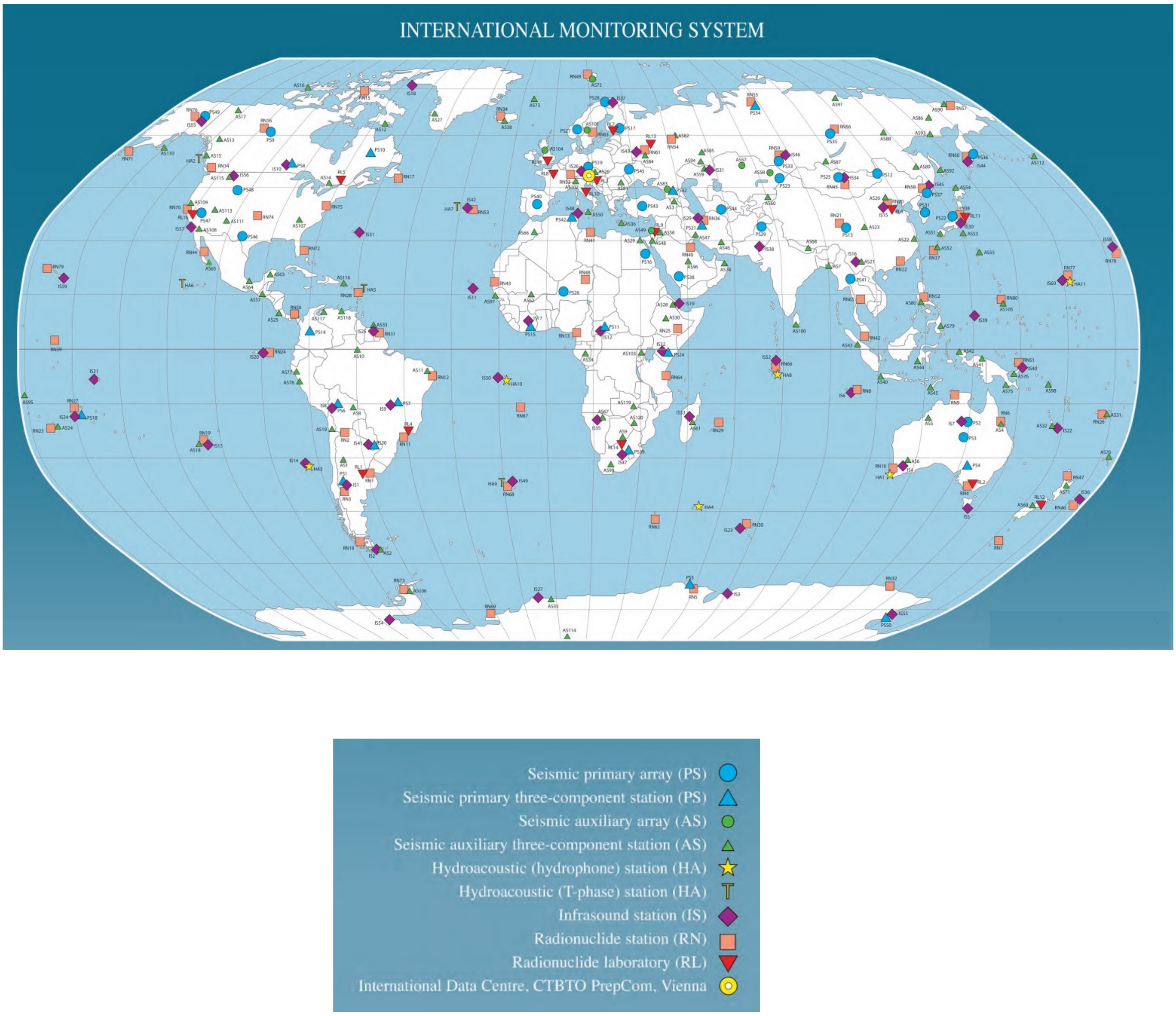

INTERNATIONAL MONITORING SYSTEM
Seismic primary array (PS) ●
Seismic primary three-component station (PS) ▲
Seismic auxiliary array (AS) ●
Seismic auxiliary three-component station (AS) ▲
Hydroacoustic (hydrophone) station (HA) ☆
Hydroacoustic (T-phase) station (HA) T
Infrasound station (IS) ◆
Radionuclide station (RN) ■
Radionuclide laboratory (RL) ▼
International Data Centre, CTBTO PrepCom, Vienna ◎

Fig. 5.4 IMS monitoring stations and laboratories will operate in 89 countries around the world. (CTBTO Public Information)

5.2.1 Seismic network

CTBTO’s seismic network consists of 50 primary stations and 120 auxiliary stations. The task of the network is to detect natural events such as an earthquake and other events such as an explosion.

Seismic arrays employ several seismic sensors arranged in a certain geometric pattern across an area that can range from a few to several hundred square kilometres. A seismic array employs two types of seismic sensors, which measure both types of seismic waves: body waves and surface waves. Seismic arrays help identify the location of an event based on information about the direction of a signal and its speed.

5.2.2 Hydroacoustic Network

An International Monitoring System network consists of eleven stations used to detect any events. Hydroacoustic station monitors the big oceans for hydroacoustic waves to detect underwater events. There are two types of stations – hydrophone stations and T-phase stations. Hydrophone stations use hydrophones, essentially underwater microphones, to detect hydroacoustic waves. T-phase stations measure seismic waves that converted from hydroacoustic waves when hitting land. These stations are usually located on oceanic islands.

5.2.3 Infrasound Network

A network consists of 60 infrasound stations, which use microbarometers to detect low-frequency sound waves in the atmosphere. A monitoring station consisting of four to eight infrasound array elements, arranged in different geometric patterns. Stations located in windy locations on isolated islands require more array elements to improve their detection capacity. At each array element, microbarometers measure the pressure changes in the air produced by infrasonic waves

5.3 International Data Centre (IDC)

International Data Centre (IDC) located at the headquarters of the CTBTO in Vienna, Austria Primary stations from above mentioned networks deliver data continuously to IDC, whereas auxiliary stations provide data upon request.

Each station needs to be equipped with communication devices to send the data for analysis to the IDC. The Global Communication Infrastructure (GCI) is developed to provide a functioning communication system for the timely, reliable and accurate transmission of data. Very Small Aperture Terminal (VSAT) is a set-up on the ground called earth station that allows for communication via a satellite.

Then, these data are stored in the file system and data are passed through a number of automatic analysis stages. Detection and Feature Extraction (DFX) and Station Processing (StaPro) are first two steps out of them.

5.4 Detection and Feature Extraction (DFX)

(IDC doc (DFX), 2001)

DFX applications perform a variety of tasks. Their primary functions are to make detections and to measure features from waveforms. DFX processes data from all three waveform-based technologies (seismic, hydroacoustic, and infrasonic). In the current system, DFX is used in automatic station processing, interactive analysis and automatic post-analysis processing. In automatic station processing, DFX detects transient signal and estimates features in the vicinity of these detections. In interactive processing, DFX estimates or updates features for detections that have been modified or added by the analysts. In automatic post-analysis processing, DFX makes a final update of the detection features, and it make measurements, based on event hypotheses, which can be used to characterize the event. DFX is also used to beam form array data.

Automatic processing begins with continuous waveforms arriving from the IMS primary stations through the Continuous Data Subsystem. The detections and features are measured in the DFX detection applications, which include: Automatic Seismic Detection, Automatic Hydroacoustic Detection and Automatic Infrasonic Detection. In addition to feature extraction, the seismic and infrasonic DFX applications produce and save detection beams.

The next process in the pipeline, StaPro (6.5), classifies detections into phase types based on the extracted waveform features. StaPro can also make single station locations at seismic stations.

5.5 Station Processing (StaPro)

(IDC doc (StaPro), 2000)

The first function in StaPro initializes station specific processing and reads detection features from the database. The core station processing consists of three main functions: Determining Signal Type, Grouping Signals and Identifying phases. Feature estimates from all signals found during DFX detection processing are written to the station processing database. StaPro uses this information to determine most likely signal types. Signals are then assembled into groups representing possible events. Phase names for regional seismic arrivals are determined by using a Bayesian analysis method and phase prediction.

Initialization

StaPro initialization includes reading user parameters specified for the station being processed, opening a database connection, loading station-specific CLIPS rules and neural network weights into memory.

Determining signal type

StaPro determines signal types (for example, P or S) of each detected signal. Signal types differ depending on the data technology, so this functional area has separate modules for S/H/I technologies. Each module has the same basic concept of evaluating feature characteristics to determine each signal type.

Grouping Signals

The purpose of this function is to place signals into groups in which each member has similar characteristics suggesting they were generated by the same event. Groups follow rules based on geophysical principles, such as P phases proceeding P coda phases or S phases.

Identifying Phases

The main purpose of this function is to identify phase names for regional seismic data. Phase names for teleseismic, hydroacoustic, and infrasonic data are limited to a more simplified naming convention by StaPro. For these data, signals are labelled similar to their signal type.

Estimating Location and Magnitude

The purpose of this function is to estimate single-station location and local magnitude for seismic groups that have been formed. An International Association of Seismology and Physics of the Earth’s Interior (IASPEI) velocity model is currently employed. However, other models could be used.

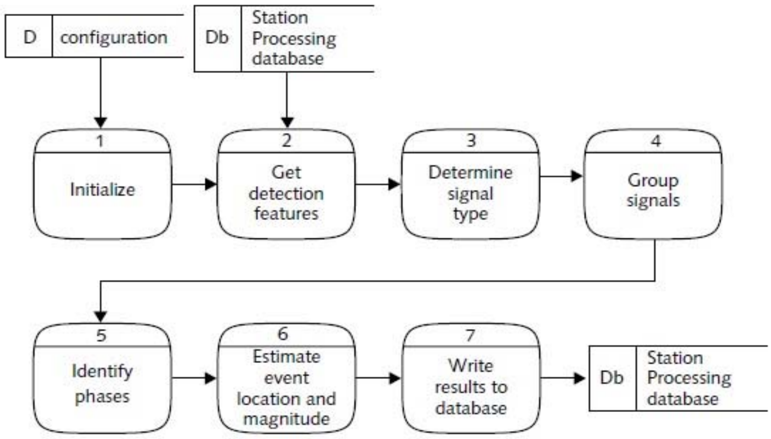

D
configuration
Db
Station Processing database
1
Initialize
2
Get detection features
3
Determine signal type
4
Group signals
5
Identify phases
6
Estimate event location and magnitude
7
Write results to database
Db
Station Processing database

Fig. 5.6 Process Flow of Station Processing (IDC doc (StaPro), 2000)

5.6 Global Association(GA) Sub System

(IDC doc (GA), 2001)

GA is the process in the automatic pipeline that forms event hypothesis. GA reads arrival and amplitude data for a time interval and forms set of associations using an exhaustive search algorithm. These association set define the events, which then are located and have their magnitude estimated.

GA's components (Five program and one library) are identified as follows :

- GAassoc
- GAconflict
- GA_DBI
- GAcons
- GAgrid
- libGA

GAassoc constructs initial event hypotheses by associating arrivals from different stations using a grid search algorithm.

The GAconflict program resolves conflicts between sectors and between time intervals. In addition, the program predicts and associates defining and non-defining phases after relocating initial event hypotheses, and it applies a number of geo-physical checks on the associations and events. It modifies and removes associations that do not pass these checks.

GA_DBI performs a few auxiliary functions and is specific to particular configuration of GA.

The GAcons process is a stand-alone program that builds the propagation knowledge base grid file and the static grid file to be used by the pipeline-activated programs GAassoc and GAConflict.

GAgrid is a GUI program that allows visualization of one of the two grid files (the propagation knowledge base grid file) built by the GAcons program. The information in the grid file is used by GAassoc to form trial event hypotheses in the initial phase of automatic association process.

6 VALIDATION AND DIAGNOSIS OF DATA PROCESSING SYSTEM

CTBTO IMS stations are subjected to various upgrades such as station equipment changes, stations parameter updates, relocation of station's sites and calibration exercises for seismometers at stations. After these changes, stations' results need to be validated. This chapter will describe how to apply statistical methods for validation process of new station parameters.

6.1 Validation of station and processing parameter changes

IMS stations send continuous data to IDC and these data are segmented to 10 min time intervals. Station data processing software (DFX and StaPro) processes these intervals and identifies detections.

It is necessary to introduce new versions of DFX and StaPro applications to improve performance of the processing pipeline. At the same time, we should install software patches to solve the software bugs in the application. These changes are implemented in a test environment at IDC. Current version of the software is being used in the production environment during this period. At the same time, both test and production environments receive same data stream from stations.

To statistically verify these software changes, 'Two-sample t-Test' hypotheses testing can be carried out for detections' mean.

After implementing these changes in the test environment at IDC, DFX application run on 10 min data intervals and it will generate detections. In this example, number of detections will be counted for same hours in both test and production environments.

During the testing period, random samples are collected from test environment. It is necessary to collect samples for same hours from production environment, where old software is installed.

Data set for test and production environments can be represented with the notation of 'D_{LHH}'

L : Represents test or production environment

HH : Represents hour of the day

E.g D_{T05} means the data sample, which is collected from test environment between 04 and 05 time interval.

Let $D_{T01}, D_{T02}, D_{T03}, \ldots, D_{TNN} \sim N(\mu_T, \sigma_T^2)$ and $D_{P01}, D_{P02}, D_{P03}, \ldots, D_{PNN} \sim N(\mu_P, \sigma_P^2)$

Assuming that mean of detections has t-distribution for both test and production environment;

μ_T and μ_P denotes the true detection mean of test and production environments respectively. Now, we consider the following hypotheses:

$$H_o : \mu_T = \mu_P \text{ versus } H_A : \mu_T \neq \mu_P$$

$$t = \frac{(\bar{x}_T - \bar{x}_P) - (\mu_T - \mu_P)}{\sqrt{S_T^2/n_T + S_P^2/n_P}}$$

$\bar{x}_T$: Mean of detections in test environment sample ($\sum D_{THH} / n_T$)

$\bar{x}_P$: Mean of detections in production environment sample ($\sum D_{OHH} / n_O$)

S_T: Std. deviation of detections in test environment sample

S_P: Std. deviation of detections in production environment sample

n_T: Number samples of test environment

n_P: Number samples of production environment

If the null hypothesis is true (H_o) : $\mu_T - \mu_P = 0$;

$$t = \frac{(\bar{x}_T - \bar{x}_P\ )}{\sqrt{S_T^2/n_T + S_P^2/n_P}}$$

If the null hypothesis is true, degrees of freedom can be calculated from the following equation:

$$v = \frac{\left(S_T^2/n_T + S_P^2/n_P\right)^2}{\left(S_T^2/n_T\right)^2/(n_T - 1) + \left(S_P^2/n_P\right)^2/(n_P - 1)}$$

P $(t \geq t_{exp})$ can be calculated with values of t and v using t-distribution table. This p value will be used to compare with value set in the Table 3.1 and corresponding descriptive language can be used to final decision.

In addition to software changes, there are various other changes are taken place at processing pipeline such as software parameter updates, database upgrades and operating system upgrades. After implementing these changes, it is possible to apply same hypotheses testing to verify results of DFX and StrPro applications using same above calculation.

6.2 Diagnostics Bayesian Network for station processing application (DFX)

The Bayesian network has been used by many systems for the development of diagnostic systems. Bayesian networks are successfully applied to a variety of applications such as machine diagnosis, robotics, data mining and natural language interpretation and planning. This chapter describes practical aspects for creating a Bayesian network model as a diagnostic support tool for DFX processing.

6.2.1 Bayesian model for DFX processing

The Bayesian network consists of two parts, qualitative part and a quantitative part. The qualitative part represents the graphical part of the network and quantitative part consists of the conditional probability tables. This diagram shows only qualitative part and next chapter will describe both these parts in a simple diagram.

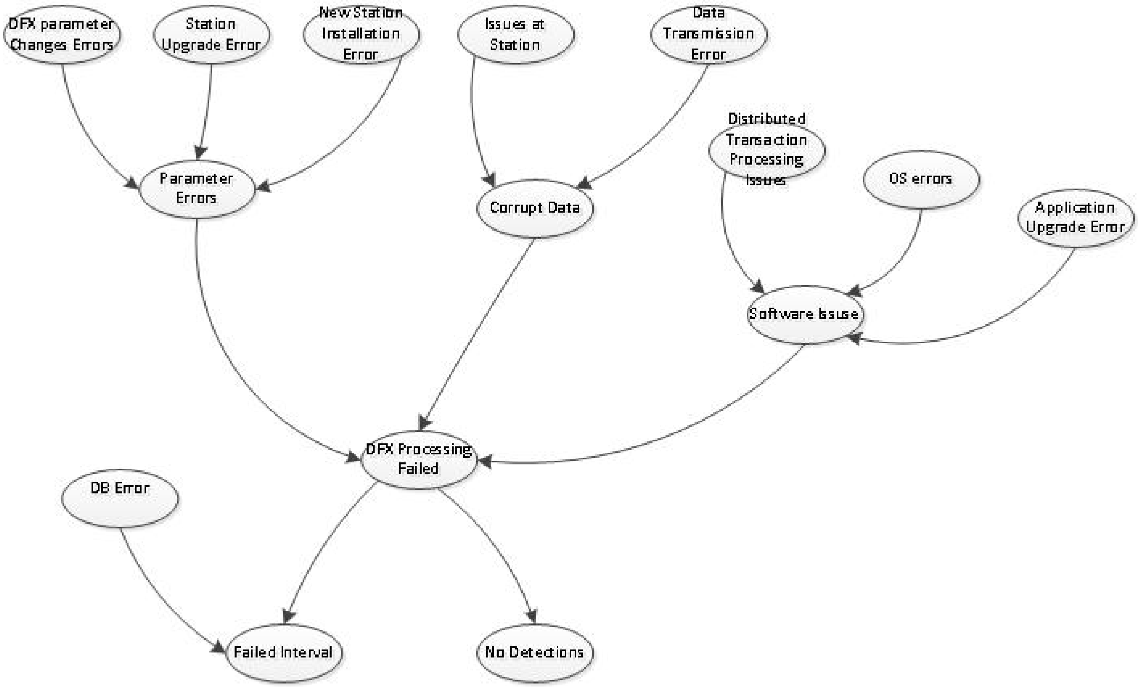


Fig. 6.1 Bayesian network for diagnosis DFX processing failures

In this Bayesian network, there are three intermediate cause nodes:

Parameter Changes Errors: DFX application uses application parameter files and station related parameters. These parameter values are modified due to software changes and SHI station changes.

IMS stations undergo various upgrades and it requires updating station specific values in station parameter files. There may be errors in these updates (*Station Upgrade Error* node)

When installing new SHI stations, it is required to generate new station specific files and update the existing shared parameter files. *New station installation* node represents probability of errors in these changes.

Corrupt Data: SHI stations send corrupt data to IDC due to failures at stations or data transmission problems. DFX processing may fail while processing these corrupt data. *Issues at Stations* and *Data Transmission Errors* are causes for corrupted data.

Software Issues: Activities such as installing DFX application patches, upgrading OS system and upgrading middleware (Distributed Transaction Processing) are taken place in processing pipeline. These activities may effect to processing of DFX application. (*Distributed Transaction Processing Issues*, *OS errors* and *Application Upgrade Errors* nodes in the figure)

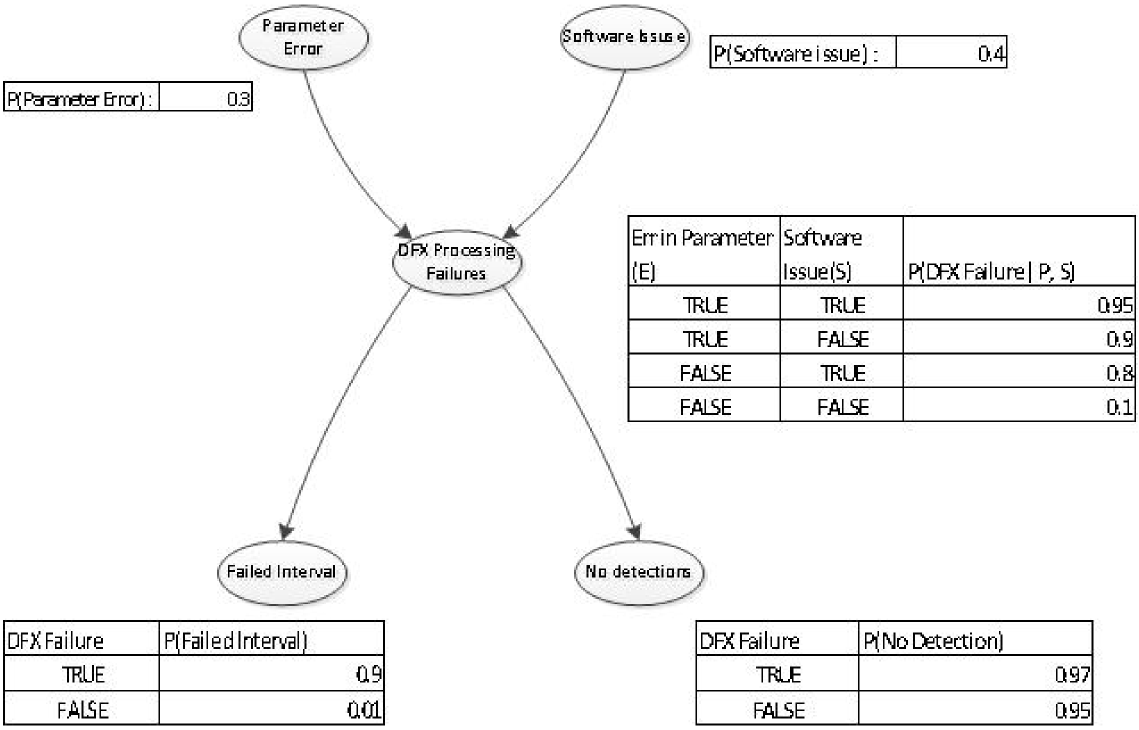
There are two evidence nodes:

Failed Interval: DFX application run on segmented station data intervals (10 min). When there is an error in DFX processing, the processing interval will be changed to "Failed" status. These failed processing intervals can be observed in a special workflow.

No Detections: The primary functions of DFX processing are to make detections and to measure features from waveforms. If there are any processing failures in DFX application, it will possible to have intervals with no station detections. Looking at database entries, it is possible to check number of detections for each station or each interval.

6.2.2 Simple Bayesian model for DFX processing

In this sub-chapter, simple version of Bayesian network is used to perform calculations for main four types of queries. All the probability values in Bayesian model are estimated values and not real.



P(Parameter Error) :	0.3

P(Software issue) :	0.4

Err in Parameter (E)	Software Issue(S)	P(DFX Failure | P, S)
TRUE	TRUE	0.95
TRUE	FALSE	0.9
FALSE	TRUE	0.8
FALSE	FALSE	0.1

DFX Failure	P(Failed Interval)
TRUE	0.9
FALSE	0.01

DFX Failure	P(No Detection)
TRUE	0.97
FALSE	0.95

Fig 6.2 Simple version of Bayesian network for diagnosis DFX processing failures

6.2.2.1 Diagnostic inference Calculation

Using above simplified version of Bayesian Network, diagnostic interference can be calculated as follows :

P (Parameter Error (E) | No detections (D))

E = P (Parameter Error)

S = P (Software Issue)

D = P (No Detections)

F = P (DFX processing Failure)

I = P (Failed Interval)

$$P(E \mid D) = \frac{P(D\,E)}{P(D)} \quad --- \quad (1)$$

$$P(D\,E) = P(D\,E\,F) + P(D\,E\,F')$$

$$= P(D \mid FE).P(FE) + P(D \mid F'E).P(F'E)$$

$DFX\ processing\ failure\ node\ seperates\ No\ Detection\ \ node\ from\ Parameter\ Error\ node$

$P(D \mid F\,E) = P(D \mid F);$

$Applying\ same\ condition;\ \ P(D \mid F'E) = P(D \mid F')$

$$= P(D \mid F).P(FE) + P(D \mid F').P(F'E) \quad --- \quad (2)$$

$Let's\ calculate\ \ P(FE);$

$$P(F\,E) = P(E\,F\,S) + P(E\,F\,S') \quad --- \quad (3)$$

$$= P(F \mid ES).P(ES) + P(F \mid ES').P(ES')$$

$Parameter\ Error\ (E)\ and\ Software\ issue\ (S)\ are\ independent\ events\ each\ other$

$$= P(F \mid ES).P(E).P(S) + P(F \mid ES').P(E).P(S')$$

$$= 0.95 * 0.3 * 0.4 + 0.9 * 0.3 * 0.6$$

$$= 0.276$$

$Calculating\ \ P(F'E);$

$$P(F'\,E) = P(F'\,E\,S) + P(F'E\,S') \quad --- \quad (4)$$

$$= P(F' \mid ES).P(ES) + P(F' \mid ES').P(ES')$$

$$= 0.05 * 0.3 * 0.4 + 0.1 * 0.3 * 0.6$$

$$= 0.024$$

$Applying\ to\ (2)\ equation\ ;$

$$P(D\ E) = P\ (D \mid F\ ) \,.\, P\ (FE\ ) + P\ (D \mid F'\ ) \,.\, P\ (F'E\ )$$

$$= 0.97 * 0.276 + 0.95 * 0.024$$

$$= 0.290$$

$Calculating\ \ P(\ D\ );$

$$P(D\ ) = P\ (D\ F\ ) + P\ (D\ F'\ )$$

$$= P\ (D \mid F\ ) \,.\, P(\ F\ ) + P\ (D \mid F') \,.\, P\ (F'\ ) \quad --- \quad (5)$$

$Calculating\ \ P(\ F\ );$

$$P(F\ ) = P(\ F\ E\ S\ ) + P(\ F\ E\ S') + P(\ F\ E'\ S\ ) + P(\ F\ E'\ S') \quad --- \quad (6)$$

$$= P(\ F \mid E\ S\ ).\ P(\ ES\ ) + P(\ F \mid E\ S').P(\ ES'\ ) + P(\ F \mid E'S\ ).\ P\ (\ E'S\ ) + P(\ F \mid E'\ S').P(\ E'S'\ )$$

$Parameter\ Error\ (E)\ and\ Software\ issue\ (S)\ are\ independent\ events\ each\ other$

$$= P(\ F \mid E\ S\ ).P(\ E\ ).P(\ S) + P(\ F \mid E\ S').P(\ E\ ).\ P(\ S'\ ) \ldots\ldots..$$

$$= 0.95 * 0.3 * 0.4 + 0.9 * 0.3 * 0.6 + 0.8 * 0.7 * 0.4 + 0.1 * 0.7 * 0.6$$

$$= 0.114 + 0.162 + 0.224 + 0.042$$

$$= 0.542$$

$Calculating\ \ P(\ F'\ );$

$$P(\ F'\ ) = 1 - P(\ F\ )$$

$$= 1 - 0.542 = 0.458$$

$Applying\ to\ (5)\ equation\ ;$

$$P(\ D) = P\ (D \mid F\ ) \,.\, P(\ F\ ) + P\ (D \mid F') \,.\, P\ (F'\ )$$

$$= 0.97 * 0.542 + 0.95 * 0.458$$

$$= 0.96084$$

$Applying\ to\ (1)\ equation$;

$$P(E \mid D) = \frac{P(D\,E)}{P\,(D)}$$

$$= \frac{0.290}{0.96084}$$

$$= 0.302$$

6.2.2.2 Causal inference Calculation

Causal inferences reason top-down from causes to effects. To illustrate this using above Bayesian network, the following example can be used:

P (Failed Interval (I) | Software Issue (S))

$$P(I \mid S) = \frac{P(\mathrm{I}\,S)}{P(\mathrm{S})} \quad --- \quad (7)$$

$$P(I\,S) \;=\; P(I\,S\,F) \;+\; P(I\,S\,F')$$

$$= \; P(\;I \mid FS\,).P(\,FS\,) + P(I \mid F'S\,).P(\,F'S\,)$$

$DFX\ processing\ failure\ node\ seperates\ Failed\ interval\ \ node\ from\ Software\ issue\ node$

$P(I \mid FS) = P\,(I \mid F)$;

$Applying\ same\ condition$; $P(I \mid F'S) = P(I \;\mid F')$

$$= \; P(I \mid F\,).P(\,FS\,) + P(I \mid F'\,).P(\,F'S\,)$$

$Perform\ the\ calculation\ for\ P(\,FS\,)\ and\ P(\,F'S\,)\ as\ explained\ in\ the\ previous\ example\ (3\,)$;

$P(FS) \;= 0.338$

$P(F'S) = \; 0.062$

$$P(IS) = P(I \mid F).P(FS) + P(I \mid F').P(F'S)$$

$$= 0.9 * 0.338 + 0.01 * 0.062$$

$$= 0.30482$$

$Applying\ to\ (7)\ equation$

$$P(I \mid S) = \frac{P(\mathrm{I}\,S)}{P(\mathrm{S})}$$

$$= \frac{0.30482}{0.4}$$

$$= 0.76205$$

6.2.2.3 Inter-causal inference calculation

Inter-causal reasoning is a common inference pattern involving probabilistic dependence of causes of an observed common effect. In the following example, probability of software issue for given DFX failure can be calculated as below:

P (Software Issue (S) | DFX failure (F))

$$P(S \mid F) = \frac{P(S\,F)}{P(\mathrm{F})} \quad --- \quad (8)$$

$Calculating\ \ P(S\,F);$

$$P(F\,S) = P(F\,S\,E) + P(F\,S\,E')$$

$$= P(F \mid S\,E).P(S\,E) + P(F \mid S\,E').P(S\,E')$$

$Parameter\ Error\ (E)\ and\ Software\ issue\ (S)\ are\ independent\ events\ each\ other$

$$= P(F \mid S\,E).P(S).P(E) + P(F \mid S\,E').P(S).P(E')$$

$$= 0.95 * 0.4 * 0.3 + 0.8 * 0.4 * 0.7$$

$$= 0.338$$

$P(\ F\ )\ is\ calculated\ in\ the\ previous\ example\ (6);$

$P(\ F\ ) = 0.542$

$Applying\ to\ (8)\ equation$

$$P(\ S\ |\ F\ ) = \frac{P(\ S\ F\ )}{P(\ \mathrm{F}\ )}$$

$$= \frac{0.338}{0.542}$$

$$= 0.623$$

Assuming there is more evidence, let's calculate same probability as follows:

P (Software Issue (S) | DFX failure (F) ^ Parameter Error (E))

$$P(\ S\ |\ F\ E\ ) = \frac{P(\ S\ F\ E\ )}{P(\ F\ E\ )} \quad --- \quad (9)$$

$Calculating\ \ P(\ S\ F\ E\ );$

$$P(\ F\ S\ E\ )\ =\ P(\ F\ |\ S\ E\ ).P(\ S\ E\ )$$

$Parameter\ Error\ (E)\ and\ Software\ issue\ (S)\ are\ independent\ events\ each\ other$

$$=\ P(\ F\ |\ S\ E\ ).P(\ S\ ).P(\ E\ )$$

$$=\ 0.95 * 0.4 * 0.3$$

$$=\ 0.114$$

$Calculating\ \ P(\ F\ E\ );$

$$P(\ F\ E\ )\ =\ P(\ F\ E\ S\ ) + P(\ F\ E\ S'\ )\ ---\ \ (10)$$

$$=\ P(\ F\ |\ E\ S\ ).P(\ S\ ).P(\ E\ ) +\ P(\ F\ |\ E\ S'\ ).P(\ S'\ ).P(\ E\ )$$

$$=\ 0.95 *\ 0.4 * 0.3\ +\ 0.9 * 0.6 * 0.3$$

$$=\ 0.276$$

$Applying\ to\ (9)\ equation$

$$P(\ S \mid F\ E\ ) = \frac{P(\ S\ F\ E\ )}{P(\ F\ E\ )}$$

$$= \frac{0.114}{0.276}$$

$$= 0.41$$

With additional evidence, the probability goes down to 0.41

6.2.2.4 Mixed inferences calculation

It combines above inferences in Bayesian network.

P (DFX failure (F) | Failed Interval(I) ∧ Parameter Error (E))

$$P(\ F \mid I\ E\ ) = \frac{P(\ F\ I\ E\ )}{P(\ I\ E\ )} \quad --- \quad (11)$$

$Calculating\ \ P(\ F\ I\ E\ );$

$$P(\ F\ I\ E\ )\ =\ P(\ I \mid F\ E\ ).P(\ F\ E\ )$$

$DFX\ processing\ failure\ node\ seperates\ Failed\ interval\ \ node\ from\ parameter\ error\ node$

$$=\ P(\ I \mid F\ ).P(\ F\ E\ )$$

$P(\ FE\ )\ is\ calculated\ in\ the\ previous\ example\ (10);$

$$P(\ F\ E\ )\ =\ 0.276$$

$$P(\ F\ I\ E\ )\ =\ P(\ I \mid F\ ).P(\ F\ E\ )$$

$$=\ 0.9 *\ 0.276$$

$$=\ 0.248$$

$Calculating \;\; P(\,I\;E\,);$

$$P(\,I\;E\,) \;=\; P(\,I\;E\;F\,) \;+ P(\,I\;E\;F'\,)$$

$$\;=\; P(\,I\;|\;E\;F\,).P(\,E\;F\,) + P(\,I\;|\;E\;F'\,).P(\,E\;F'\,)$$

$DFX\;processing\;failure\;node\;seperates\;Failed\;interval\;\;node\;from\;parameter\;error\;node$

$$\;=\; P(\,I\;|\;F\,).P(\,F\;E\,) + P(\,I\;|\;F'\,).P(\,E\;F'\,)$$

$P(\,FE\,)\;is\;calculated\;in\;the\;previous\;example\;(10);$

$$P(\,F\;E\,) \;=\; 0.276$$

$P(\,F'E\,)\;is\;calculated\;in\;the\;previous\;example\;(10);$

$$P(\,F'\;E\,) \;=\; 0.024$$

$$\begin{aligned}
P(\,I\;E\,) \;&=\; P(\,I\;|\;F\,).P(\,F\;E\,) + P(\,I\;|\;F'\,).P(\,E\;F'\,) \\
&=\; 0.9 * 0.276 \;+\; 0.01 * \;0.024 \\
&=\; 0.9 * 0.276 \;+\; 0.01 * \;0.024 \\
&=\; 0.2486
\end{aligned}$$

$Applying\;to\;(11)\;equation$

$$\begin{aligned}
P(\,F\;|\;I\;E\,) &= \frac{P(\,F\;I\;E\,)}{P(\,I\;E\,)} \\
&=\; \frac{0.248}{0.2486} \\
&=\; 0.9974
\end{aligned}$$

6.3 Diagnostics Bayesian Network for GA application

The Diagnostic Bayesian network for GA application consists of two parts, qualitative part and quantitative part. The qualitative part represents the graphical part of the network and quantitative part consists of the conditional probability tables. The figure 6.3 shows only qualitative part.

In this Bayesian network, there are four intermediate cause nodes:

Parameter Changes Errors: GA application uses application parameter files and these parameter values are subjected to modify due to software changes. GA application also reads earth-model configuration data. Any errors occur during modifications to these parameter files will effect on GA processing.

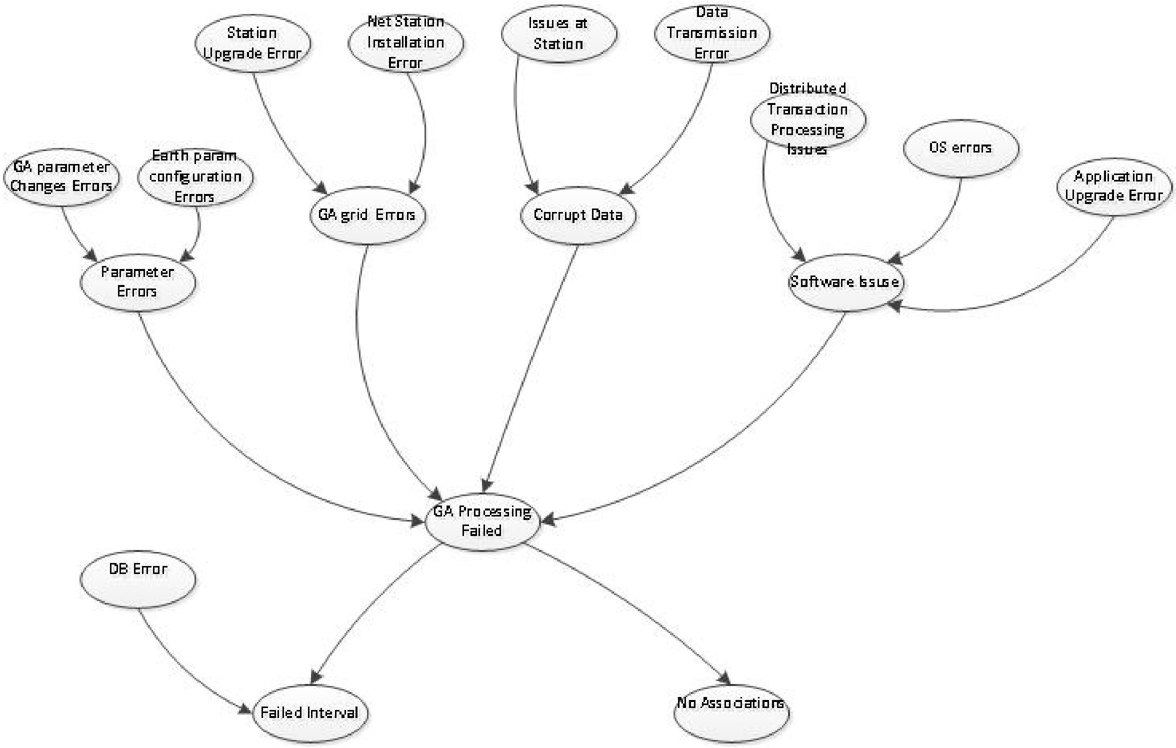


Fig 6.3 Bayesian network for diagnosis GA processing failures

GA grid errors: IMS stations undergo various upgrades and it results updating station specific parameter values in parameter files. When installing new SHI stations, it is required to generate new station specific files and update existing shared parameter files.

After installing the new primary stations, it must be required to generate new GA grid. It may be required to generate new GA grid file for station parameter upgrades depending on the changes. Station upgrade and new station installation errors will contribute to GA grid errors.

Corrupt Data: SHI stations send corrupt data to IDC occasionally due to failures at stations or data transmission problems. GA application may fail while processing these corrupt data. *Issues at Station* and *Data Transmission Errors* nodes represent causes for corrupted data.

Software Issues: Installing GA application patches, upgrading OS system and upgrading middleware (Distributed Transaction Processing) are major activities, which carried out on processing pipeline. These activities can cause to GA processing. *Distributed Transaction Processing Issues*, *OS errors* and *Application Upgrade Errors* nodes are cause nodes for software issues.

There are two evidence nodes:

Failed Interval: GA processing executes on 10min station data segments. When there is an error in GA processing, the status of processing interval will be "Failed". These failed intervals can be observed in a special workflow. At the same time, this failed interval also can be caused by database errors because interval data are read and written to a database table.

No associations: The primary function of GA is to read arrival and amplitude data for a time interval and forms set of associations using an exhaustive search algorithm. If there are processing failures in GA, there will no associations. It is possible to count number of associations for each station at each interval by querying at database entries.

6.4 Populating Conditional Probability Table

It is easy to understand failures in ‘diagnostic’ form than ‘causal’ form to diagnostic experts. This is due to the fact that they are primarily interested in determining component failure given test results. For example, if an electrical system indicator light is illuminated on an automobile dashboard, and automotive diagnosis expert will have little difficulty determining the probability that the car has, say, an alternator malfunction. However, to determine the likelihood that a particular dashboard light is on or off given alternator failure may be hard to answer, because it is equivalent to asking the expert to pass judgment on the effectiveness of the light to capture various forms of alternate failures. This is a question on test design relative to functional modes of the observed component, which may fall outside of the expert’s domain knowledge.

(K. Wojtek Przytula & Don Thompson (2000))

To fill out Conditional Probability Tables (CPT), it required knowledge of domain (diagnostics) experts and processing engineers, who has lot of experiences about processing failures in DFX and GA applications.

Diagnostic inference in the Bayesian network context refer to conditional probabilities of the form P (Parameter Error | No detections) from figure 6.1. This indicates probability of parameter error for given failure condition of “No detections”. To obtain this information, it is required to look at previous DFX and GA processing failures with help of processing engineers.

Determining the prior probability of any errors such P(OS error) will not be easy task but system operators can obtain this information from past log files or log tickets, which contains failures on mean time.

In above Bayesian networks, the prior component probability and conditional diagnostic probabilities may be not sufficient to calculate conditional causal probabilities of the network.

An example from Fig 6.3:

P (software issue | distribution transaction error, ~OS error, ~application upgrade error)

It is required to elicit additional diagnostic or prior probability information to network. The following theorem can be used to build a complete diagnostic Bayesian model, capable of forward and backward reasoning with reduced burden for the domain expert.

Theorem 1:

(K. Wojtek Przytula & Don Thompson (2000))

Suppose we have the Bayesian network depicted in figure 6.4. Given the complete diagnostic conditional joint distribution of “component defectiveness given T”. (i.e the complete set of probabilities of the form {P($C_{1,}$ $C_{2, ...,}C_n$ | T)}, over all complemented and un-complemented value combinations of the C_i), the single probability P($C_{1,}$ $C_{2, ...,}C_n$) and the single probability P($C_{1,}$ $C_{2, ...,}C_n$ | T’), it is possible to calculate the complete joint probability distribution of Ci and T, and thus, all probabilities pertaining to these variables. In particular, it is possible to calculate all causal probabilities.

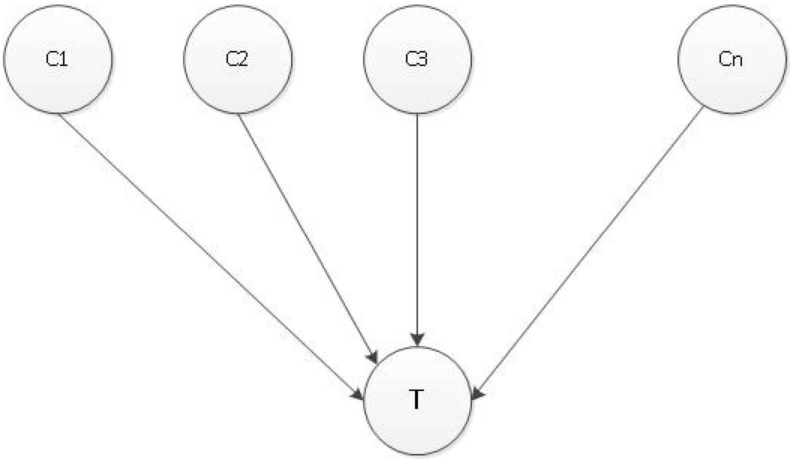


Fig 6.4 Bayesian network node with C and T primary events

Proof of Theorem

The proof follows inductively on the number of components. The root case of one component follows.

First of all, it is clear that from our given information we may calculate P(C’), P(C’|T) and P(C’|T’)

$$P(C') = 1 - P(C)$$

$$P(C'|T) = 1 - P(C|T)$$

$P(C'|T') = 1 - P(C|T')$

Next, using the laws of probability we have:

$$\begin{aligned} P(C|T) &= P(C,T)/P(T) \\ &= P(C,T)/(P(C,T) + P(C',T)) \end{aligned}$$

$$\begin{aligned} P(C|T') &= P(C,T')/P(T') \\ &= (P(C) - P(C,T))/(1 - P(C,T) - P(C',T)) \end{aligned}$$

Solving for $P(C|T)$ and $P(C|T')$ we see that are led to the matrix system:

$$\begin{bmatrix} P(C'|T) & P(C|T) \\ P(C'|T') & P(C|T) \end{bmatrix} \begin{bmatrix} P(C,T) \\ P(C',T) \end{bmatrix} = \begin{bmatrix} 0 \\ P(C) - P(C|T) \end{bmatrix}$$

The determinants of the coefficient matrix of this system reduce to:

$$P(C|T')P(C'|T') - P(C'|T)P(C|T') = P(C,T)P(C',T') - P(C',T)P(C,T')/(P(T)P(T'))$$

Which has a vanishing numerator only if $P(C,T)/P(C',T) = P(C,T')P(C',T')$, which is equivalent to C and T being independent events. We assume that this is not the case, else our conditional probabilities all collapse to prior probabilities, an uninteresting case.

Upon solving the above matrix system, we get $P(C,T)$ and $P(C',T)$; hence also

$P(C|T') = P(C) - P(C|T)$

$P(C'|T') = P(C) - P(C'|T)$

Thus, we can complete determine the joint distribution of C and T. This will uniquely determine all pertinent probabilities in this two-node network, including the causal probabilities.

7 SUMMARY AND FUTURE WORK

Hypothesis testing is one of the most widely used methodologies in statistics. Testing hypothesis is a powerful statistical method in validating changes and modification, which have been carried out in data processing pipeline at IDC.

There were no actual data collected during this master thesis. It is required to collect data samples and analyse their distribution for future work. These data samples can be collected through system operators at IDC.

At the same time, only statistics validation will not be enough to accept or reject validation changes. Most of these changes will be required additional testing with system experts, data analysis and scientists. Testing hypothesis methodology will be an additional auxiliary tool for validating changes in processing pipeline.

You might to perform several experiments to find out correct significance level, minimum number of samples required for hypothesis testing and length of data segment e.g. when we are computing number of detections, it is necessary to define a time period (1hrs, 2hrs or more lengthy hour) for data collection. Once these details are finalized, some script can be developed to collect necessary samples from real data. Currently, there are many numbers of commercial statistical applications available in the market. These collected samples can be fed to the statistical applications to get the result faster.

Bayesian network section of Artificial Intelligence is a popular method in the modelling under the uncertain knowledge. This topic has become a research field recently and there are many books which take a broad look at the literature on Bayesian networks. This master thesis covered most discussed topics in this Bayesian networks field. However, specific or research level topics have not been discussed.

Chapter 4 explained about Bayesian Network and its theories. In the same chapter, we discussed how to create generic algorithm for Bayesian queries. In Chapter 6, these theories were applied to create diagnostic support tool for DFX and GA applications. we have discussed about four inferences calculation using simple diagnostic Bayesian Network for DFX application in that chapter.

Finding out prior probabilities of errors such as station upgrade errors, OS errors, application upgrade errors and any other error is a difficult task. Calculating conditional probabilities need more domain experts’ knowledge. We discussed about populating conditional probability tables (CPT) in chapter 6.4. But probability elicitation for Bayesian network discussion is beyond this thesis. Probability elicitation will be an interesting area for the further research.

DFX and GA are only two sub systems of Automatic processing pipeline at IDC but there are more complex subsystems are included in the pipeline. After gaining knowledge of building Bayesian network of those two sub systems, Bayesian network knowledge engineers can investigate about other sub systems as well.

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
BN	Bayesian Network
CTBTO	Comprehensive Nuclear-Test-Ban Treaty Organization
DFX	Detection and Feature Extraction
GA	Global Association
GCI	Global Communication Infrastructure
IASPEI	International Association of Seismology and Physics of the Earth's Interior
IDC	International Data Centre
IMS	International Monitoring System
MB	Markov Blanket
PREM	preliminary Reference Earth Model
StaPro	Station Processing
S/H/I	Seismic/Hydroacoustic/Infrasonic
VSAT	Very Small Aperture Terminal

LIST OF FIGURES

LIST OF TABLES

REFERENCES

Peter M.Shearer (2009): Introduction to SEISMOLOGY (Second Edition)
pages 1, 87, 88, 349

Laura Chihara and Tim Hesterbery (2011): Mathematical Statistics
pages 1,215-217, 221

Morris H. DeGroot (1986): Probability and Statistics (Second Edition)

ELA KUMAR (2008): Artificial Intelligence *pages 240-245, pages 256-269*

Tomasz Ku laga, 2006: The Markov Blanket Concept in Bayesian Networks (Master Thesis)
pages 18-20

Prof. P. Dasgupta (2008): Lecture notes on Artificial Intelligence

Thorne Lay & Terry C. Wallace(1995): Modern Global Seismology, *pages 5-6*

K. Wojtek Przytula & Don Thompson (2000) Construction of Bayesian Networks for Diagnostics *pages 6*

IDC-Documentation (2001) : Detection and Feature Extraction (DFX) Scheme Files
IDC-Documentation (2001) : Global Association (GA) Subsystem
IDC-Documentation (2000) : Station Processing (StaPro)