Performance of Expert Judgement Methods with Expert Modelling

Marieke M. J. W. van Rooij

August 16, 2005

Preface

This thesis describes the result of my graduation project performed partly at the Harvard Center for Risk Analysis (HCRA) in Boston, partly at Delft University of Technology (TU Delft). From September 2004 to March 2005 I have worked with Jim Hammit at the HCRA and from March until August 2005 I have finished the project and this thesis. The whole project was performed under the supervision of my board of examiners:

Prof. dr. Roger M. Cooke
Prof. dr. Tom Mazzuchi
Dr. Eric Cator

First I would like to thank Jim and Roger for giving me the opportunity to live and work in Boston for 7 months. Apart from this thesis I feel that living in Boston is a most valuable experience of my years at the TU Delft. Furthermore I would like to thank Tom who has given me many useful comments as well as the motivation to finish this project. Also Tom and his family offered me their home when I was in Washington DC for the summer of 2003. Another very valuable and fun experience.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In mathematical modelling data, is important. Unfortunately real data is very often incomplete or even unavailable. One alternative method to obtain or generate data can be Expert Judgement (EJ). Expert Judgement is the common name for all methods that use experts. Not all Expert Judgement methods produce valuable scientific data. In order to be scientifically adequate one must be able to reproduce the results and have insight in the process. The basic idea behind Expert Judgment methods is to ask a group of experts in the field of interest to quantify their degree of belief regarding the value and uncertainty of some unknown variables. These opinions can then be combined into one opinion. In general an Expert Judgement method involves the way of eliciting experts, scoring and weighting them and combining the various opinions into one result.

One of the components of each expert judgement method is thus to combine the expert opinions. This can be done in several ways. In this thesis three Expert Judgement methods are considered. The Classical Model (Cooke, 1991), the Bayesian Belief Net Method (Small et al, 2005) and the Copula Method (Jouini and Clemen, 1998). The latter one is not studied here but forms a basis of some comparisons.

In 2001, a project was started at the Harvard Center for Risk Analysis to evaluate and compare the performance of the Classical Model and the Copula Method. (Hammit and Cohen, 2001) Cohen and Hammit proposed a comparison of the two combining Expert Judgements methods: the classical and copula methods by using synthetic data. Their goal was to provide useful guidances to decision makers and practioners. [1]

Another way to compare the two studies is by using real data. TU Delft has performed many EJ studies over the years and evaluation of the two methods using data from these studies can be found in: [2, 3].

The results of comparing the two methods on real data showed that the Classical Model performed better than the Copula Method. Based on these early results from real data and the fact that the copula method has not been profiled much the past couple of years, The Harvard group focused on the Classical Model and use of synthetic data to gain more insight in the statistical behaviour of the model.

In the past years, many new Expert Judgement methods have been published that describe the use of Bayesian techniques. Mitchell Small et al [5] proposed the use of Bayesian Belief Nets in weighting and aggregating multiple expert judgements. Although the method seems very attractive and easy to implement, there are a few theoretical comments.

1.1 Objectives

1.1.1 Evaluation of the Classical Model

This thesis describes a model to produce synthetic expert opinions that gives insight in the statistical behaviour, convergence and performance of the Classical Model. The advantage of using synthetic data it that one knows exactly what processes produce the experts' judgements and hence the true value of their calibration, dependence and other characteristics that are not known with real expert judgement data.

Statistical Insight

It is important to investigate the way Expert Judgement methods respond to changes in expert characteristics. Early results evaluating the Classical Model using synthetic data suggest that it responds in a sensible fashion to differences in the expert distributions. The way the Classical model responds to changes in expert biases, variance, correlation as well as various numerical variations can be found in Chapter 3, Modelling Experts.

Convergence

When performing an Expert Judgement project in real life it is useful to have some ideas or guidelines on the method's convergence properties. For instance, if the Decision Maker consults very few experts, an extra expert could make a significant difference in the final result. On the other hand consulting a large number of experts to ensure convergence results in practical problems, time and money constraints. This results in:

1. Under what conditions does the Classical Model converge?
2. If it converges, at what rate?

1.1.2 Evaluation of the Bayesian Belief Net Method

Many methods have been proposed on weighting and combining experts using Bayesian techniques. It would be interesting to know to what extend these methods are useful in the everyday practice of Expert Judgement studies and what they add to already existing methods like the Classical Model. In evaluating the theoretical justification and by applying the method on real data from Expert Judgement studies performed at the TU Delft I have tried to answer these questions.

Applying the method to existing data from Expert Judgement studies performed at the TU Delft shows that performance-based weighting as used on the Classical Model outperforms the Bayesian Belief Net method. This thesis presents the results on performance of the Bayesian Belief Net method and our theoretical evaluation of the method as well as an overview of Bayesian Belief Nets in general.

1.2 Outline of this Report

II Chapter 2 will provide the reader with some background on Expert Judgement in general. Concepts that are used throughout this thesis are explained and defined.

III Chapter 3 presents model I. This is a model to generate synthetic expert distributions. By applying the Classical Model to this synthetic data, we gain more insight in the Classical Model. The Bin-model in Chapter 3 uses the same way to model experts and without generating any expert distributions, leads to the probability density function of the performance-based weight in the Classical Model.

IV The results from Chapter 3 lead to Chapter 4. In this chapter the convergence of the Classical Model is illustrated and comparing with other results lead to practical suggestions regarding the use of groups of experts.

V and VI Chapter 6 gives an overview on Bayesian Belief Nets and Chapter 5 will give the results from evaluating the Bayesian Belief Net Method. By comparing this method with other weighting methods on real data from Expert Judgement studies from the TU Delft, we can give some conclusions regarding the performance of the Bayesian Belief Net method.

VII Finally this thesis will end with conclusions and recommendations in Chapter 7.

Chapter 2

Basic concepts of Expert Judgement

2.1 Introduction

Expert Judgement deals with data obtained from experts that can be used in various decision-making problems involving uncertainty. Expert Judgement is typically used in cases where data is sparse of difficult to obtain.

2.1.1 Experts and Uncertainty

Before giving more information about some Expert Judgement models and methods, let us look more closely at the concepts experts and uncertainty. In general experts are people that work in the field of interest. They are for instance prominent researchers, educators or practioners in the field. At the TU Delft, many Expert Judgement studies have been performed. Typically the group of expert consists of $10-15$ persons. Throughout this thesis E will denote the total number of experts in one Expert Judgement study or analysis and the expert are represented by $e_1, e_2, \ldots, e_E$.

Uncertainty There are many ways to elicit experts as can be found in [6]. In the case of continuous variables, experts can either estimate the exact value or express their uncertainty about a variable by giving a distribution. Although experts may in general possess valuable knowledge about the problems or field of interest, their knowledge is not certain. Therefore eliciting a distribution for each variable is preferred to eliciting point estimates. Eliciting distributions can be realised in non-parametric and parametric forms. Parametric elicitation in the case when experts are able to assess certain parameters of distributions of a known type (e.g. Normal, Weibull, Exponential.) In the case of non-parametric elicitation, the expert can express his belief by specifying so-called quantiles.

Quantiles One way to elicit a distribution is by asking experts to assess a number of quantiles. By quantifying various quantiles, the expert expresses his belief about his uncertainty and the value of an unknown variable. Quantifying the $a-$ and $b-$quantile where $a < b$, the expert expresses his belief that the probability that the true value of the unknown variable lays in between the two quantiles is $(b - a)$. In most Expert Judgement studies at the TU Delft the 0.05, 0.50 and $0.95-$ quantiles where elicited. In this thesis we will denote the quantiles with $q_1, \dots, q_K$ where K is the number of quantiles.

Example An expert assesses values $x_{0.05}$, $x_{0.50}$ and $x_{0.95}$ to the 0.05, 0.50, 0.95 quantiles respectively. This means that the expert's subjective belief about the true value of this variable is that it is smaller than $x_{0.05}$ with probability 0.05, between $x_{0.05}$ and $x_{0.50}$ and between $x_{0.50}$ and $x_{0.95}$ with probability 0.45 and larger than $x_{0.95}$ with probability 0.05.

2.1.2 Scoring and Weighting

When the experts have given their subjective probability distribution for each variable, the next step is to combine these distributions into one distribution for each variables, called the Decision Maker's distribution. There are many ways to combine experts. One way is to weight each expert and combining them as a weighted average Decision Maker. Weighting can be based on scoring rules. A scoring rule defines a measure to quantify the experts' performance or significance. In order to assure reliable results, one requirement of scoring rules is that they should be (strictly) proper. The informal definition of a (strictly) proper scoring rule is that it should be defnied such that an expert receives maximal score if (and only if) his stated assessment corresponds to his true opinion.

This requirement ensures that an expert can only optimise his weight by telling his true belief [13].

Using a ((strictly) proper) scoring rule to quantify each expert's performance should result into weights for all experts that sum up to one. Thus, based on their performance (quantified in some way,) each experts receives a weight $w_1, w_2, \dots, w_E$, and the Decision Maker's distribution for each variable is the weighted average of the expert's assessments:

$$\mathrm{DM_i(x)} = \sum_{\mathrm{j-1}}^{\mathrm{E}} \mathrm{w_i f_{j,i}(x)} \tag{2.1}$$

where $f_{j,i}(x)$ denotes expert j's distribution for variable i.

Seed Variables In order to quantify the performance of the experts, in Expert Judgement studies one can ask expert to assess variables that are already known to the analyst or will be known during the analysis. These variables are called seed variables and in this thesis are also referred to as seeds. N will denote the number of seed variables. The values of the seed variables are called their realisations.

2.2 The Classical Model

This section briefly describes the Classical Model as proposed by Cooke. For an extended explanation of the model, we would like to refer to [6]. The Classical Model has been thoroughly reviewed and used in over 30 real-life Decision-making problems. Appendix C gives an overview of all these studies.

The Classical Model describes a way to quantify the performance of experts who give their subjective probability functions in the form of a number of specified quantiles. Two measures are used and the product of those two measures for each expert gives his unnormalised weight.

2.2.1 Performance-based Weighting

The Classical Model uses a performance-based weighting scheme to weight the experts. The basic idea behind this is that experts who are performing ”well” on seed variables might also perform well predicting the unknown variables of interest. To express the performance of experts in a quantitative manner, two quantitative measures are used in the Classical Model: the Calibration and Information.

2.2.2 Calibration

In the Classical Model the Calibration measures the statistical likelihood of the hypothesis that the realizations are sampled independently from distributions agreeing with the expert's assessments.

An expert states a number of n fixed quantiles (say three) of his subjective probability distribution for each uncertain variable. Then for each distribution there are $n + 1$ bins between the n quantiles in which the actual value of the variable may fall. Now we define the probability vector P as the vector containing the probabilities corresponding to the size of the bins. Suppose an expert gives his quantile assessments for N seed variables. Then of all N realisations of the seed variables, a number falls below the smallest quantile, a number falls in between the quantiles and a number of seed variables falls strictly above the upper quantile. These numbers can be computed as relative frequencies and the empirical probability vector containing these relative frequencies is denoted by S.

Calibration Now, we can calculate the Calibration by testing the hypothesis that the realisations are an independent sample from the multinomial distribution with probability vector P. In case of using the $5-, 25-, 50-, 75-$ and 95-quantiles, the above hypothesis means that exactly 5% of the realisations of the seed variables fall under the 5-quantile of the corresponding distribution given by the expert, 20% falls between the 5- and 25-quantile of the corresponding expert distribution and so on. A known results in statistics is that when sampling S independently from a multinomial distribution with probability vector P, the quantity

$$2NI(S,P) = 2N\sum_{i=1}^{N} S_i \ln(\frac{S_i}{P_i}) \tag{2.2}$$

is asymptotically Chi-square distributed with n degrees of freedom (n is the number of assessed quantiles.)

Let χ_n^2 denote the cumulative probability function for a Chi-square variable with n degrees of freedom. Then the Calibration is defined as the probability that a Chi-square distributed random variable with n degrees of freedom is larger than $2NI(S,P)$ under the hypothesis. Thus:

$$C = 1 - \chi_n^2(2NI(S,P)) \tag{2.3}$$

The larger this probability, the better the calibration. Recall that P is the same for all experts as all experts assess the same quantiles, and S depends on the expert assessments.

2.2.3 Relative Information

Calibration tells us something about how far away expert assessments are from the realisations of the seed variables. Another way to quantify the expert performance is by means of his Relative Information. Information tells us something about the spread of the expert assessed distributions.

Information In the Classical Model this Information is computed as the average Relative Information of the expert distributions with respect to a user-defined background. This is either a uniform or log-uniform distribution on the intrinsic range. The intrinsic range is the smallest interval containing all assessments for a given item plus the realization if available, overshot by 10% above and below. The expert's information scores are affected by the choice of the overshoot. A very large overshoot moves the information scores of the experts relatively close together. In general the Relative Information between two densities f and g on an intrinsic range I_j is :

$$\begin{array}{ll} I\left(f,g\right) = \int_{u \in I_j} f(u) \ln\left(\frac{f(u)}{g(u)}\right) du & \text{in the continuous case} \\ I(f,g) = \sum_{u \in I_j} f(u) \ln\left(\frac{f(u)}{g(u)}\right) & \text{in the discrete case} \end{array} \tag{2.4}$$

The Relative Information of an expert's distribution with respect to the uniform background is:

$$c\sum_{e\in I_j} f(u)\ln(\frac{f(u)}{\frac{1}{I_j}}) \tag{2.5}$$

$$\tag{2.6}$$

$$=\sum_{e\in I_j} f(u)\ln(f(u)I_j) \tag{2.7}$$

$$=\sum_{e\in I_j} f(u)\ln(f(u))+f(u)\ln(I_j) \tag{2.8}$$

2.2.4 Global Weights

The performance of each expert can thus be quantified with both the Information and the Calibration. The performance-based weights or global weights are now proportional to the product of the Calibration and average Information over seed variables. This is the strictly proper scoring on which the Classical Model is based.

2.3 Other Methods

Of many other Expert Judgement methods I would like to mention Bayesian models and the Social Network method.

2.3.1 Bayesian Models

Bayesian models are Expert Judgement models in which the expert assessments are treated as observations and Bayes' theorem is used to update the prior distribution of an unknown variable based on the observations. [6] A comprehensive review of Bayesian methods is given in [14]. In Chapter 6 an Expert Judgement method using Bayesian Belief Nets is reviewed and compared with performance-based global weights and equal weights.

2.3.2 Social Networks

Recently Huang and Cooke review the use of Social Networks as a tool for Expert Judgement. [13] Social Network (SN) theory views social relationships in terms of nodes and ties and focuses on relationships among social entities. In comparison with performance-based global weighting and combining of experts SN does not perform well. It does however outperform equal weighting and combining. More on this subject can be found in [13] and Chapter 6.

Chapter 3

Modelling Experts

This chapter gives the result of modelling expert opinions with a stochastic process. Instead of real expert opinions these stochastic processes result in synthetic data. This data will be interpreted as data produced by experts and applying the Classical Model to this data gives results on how the method responds to changes in experts characteristics.

Model I produces synthetic expert distributions and the Classical Model is used to aggregate the expert distributions. This leads to some results illustrating the statistical behaviour of the Classical Model. The experts' characteristics are represented as the parameters of their underlying distribution. In the next section we will look further into the expert characteristic distributions. Using bins we can construct the probability density function of the unnormalised weight. In this bin-model there are no expert distributions generated. Together the results from the two different models lead to some interesting conclusions about the way the Classical Model deals with experts' distributions in terms of the experts' characteristics.

3.1 Model I - Modelling Experts with Synthetic Data

By modelling experts in order to produce synthetic data, we will know exactly what processes produce the expert judgements. Applying an Expert Judgement method to this synthetic data will result in more insight in the statistical behaviour of the method. Model I produces expert judgements in such a way. In [1, 17, 18] the authors propose a model which we will use. The model incorporates the fact that an expert does not know his own characteristics but the analyst does. The latter is not true in real life. The expert specific parameters are an input to the model and enable the analyst to "play around" with and investigate the effect of changes in the parameters of an expert on the final aggregated distribution for the target variable or on the (un)normalised weight an expert receives.

3.1.1 Model Assumptions

This model has as main goal to produce synthetic expert data. First we define E as the number of experts and $N+1$ as the number of variables. Of these variables N are used as seed variables for computing the performance-based weights in the Classical Model as explained in Chapter 2. For each of these variables model I will generate the assessments of all E experts. Furthermore let the subscript e denote an arbitrary expert, $e = 1, 2, \ldots, E$ and the subscript j denote an arbitrary variable, $j = 1, 2, \ldots, N+1$. Finally. let the true value of all variables be zero.

Expert Distribution Instead of reporting a point estimate for each variable the expert expresses his uncertainty about a variable by reporting a distribution. Another model assumption is that all expert distributions are normal distributions. Before looking more carefully at these normal distributions and their parameters we will first give the notation.

Expert e's distribution for variable j is represented by:

$$\mathcal{N}(x_{e,j}, \sigma_e^2) \tag{3.1}$$

Medians Finally, we assume that the medians of the expert distributions $x_{e,j}$ are also randomly distributed. The medians of all expert distributions are modelled as realisations of a random variable characteristic of expert e. For this, we choose the random variables $X_e \sim \mathcal{N}(\mu_e, \sigma_e^2)$. X_e is thus normal and its parameters are characteristic for each expert. This is a logical choice. It is both mathematically attractive and a realistic representation of errors in general. [4]

3.1.2 Expert Characteristics

Each expert has his own characteristics that influence the errors he makes in assessing each unknown variable. An expert can for example tend to overestimate. In the case of overestimating, the average median of the expert distributions will be higher than zero, the true value of all unknown variables. The other way around, the higher the medians, the more an expert tends to overestimate. We will represent the expert characteristics by an expert-specific underlying distribution. This underlying distribution is normal and determines the medians of the expert distributions. The underlying distribution of expert e: $\mathcal{N}(\mu_e, \sigma_e^2)$ where the mean μ_e is the expert e's underlying bias and the underlying variance σ_e^2 represents the underlying spread of the expert judgements. In the case of the overestimating expert, μ_e will be a positive number.

Now the medians of the expert distributions are drawn from his underlying distribution. Thus $x_{e,1},\dots,x_{e,N+1}$ are realisations from the random variable $X_e \sim \mathcal{N}(\mu_e, \sigma_e^2)$ and for variables $j = 1,\dots,N+1$ expert e reports

$$\mathcal{N}(x_{e,1}, \sigma_e^2), \dots, \mathcal{N}(x_{e,N+1}, \sigma_e^2) \tag{3.2}$$

In practice experts are asked for a few points of their distribution, the quantiles. In this analysis each expert reports five quantiles of his normal distribution. These five quantiles are: $\begin{bmatrix} 0.05 & 0.25 & 0.5 & 0.75 & 0.95 \end{bmatrix}$. Recall from Chapter 2 that by quantifying the 0.25- and 0.5-quantile an expert expresses that his belief is that the true value of the variable he assesses falls between the two quantiles with probability $0.5 - 0.25 = 0.25$.

3.1.3 Correlation

Another characteristic for each expert is the dependence between his assessments. A measure of this dependence is the correlation between these assessments. In practice it makes sense to assume that the medians of an expert's distributions $x_{e,j}$'s for the various variables are correlated with each other. Correlation between the various experts is another realistic assumption, but is not yet incorporated in the model. To draw a sample for each median of the expert distribution and to assure that these medians are correlated, we will sample from a multivariate normal distribution.

Multivariate Normal Distribution

Each expert has his own underlying mean μ_e and variance σ_e^2 which represents the expert's characteristics and are input for the model. Besides these characteristics each expert has an underlying correlation that represents the dependence between the medians of his reported distributions. By imposing a correlation between the medians of the distributions (the variances are identical) we model this dependence.

To determine the medians of the expert distributions $x_{e,1},\cdots,x_{e,N+1}$ for expert e with a non-zero correlation ρ_e between them and underlying distribution $\mathcal{N}(\mu_e, \sigma_e^2)$, we will sample from a multivariate normal distribution. The parameters of the multivariate normal distribution differ among the experts and are a vector of means **Mu** and a covariance matrix **S**. The multivariate normal distribution must have a dependence structure that involves a correlation ρ_e between the variables for each expert.

Using the following general relations:

$$corr(X_i, X_j) = \tfrac{cov(X_i,X_j)}{\sigma_i \sigma_j} \quad corr(X_i, X_i) = \tfrac{cov(X_i,X_i)}{\sigma_i^2} = 1 \tag{3.3}$$

the covariance matrix that is one of the parameters of the multivariate normal distribution for expert e is given by:

$$\mathbf{S} = \begin{pmatrix} \sigma_e^2 & \rho_e * \sigma_e^2 & \cdots \\ \rho_e * \sigma_e^2 & \sigma_e^2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \tag{3.4}$$

The other input parameter, the vector of means **Mu** contains the underlying bias μ_e and is given by

$$\mathbf{Mu} = \begin{pmatrix} \mu_e \\ \vdots \\ \mu_e \end{pmatrix} \tag{3.5}$$

Drawing samples from a multivariate normal distribution with covariance matrix **S** and mean **Mu** results in $N+1$ medians for expert e. The medians are correlated with correlation ρ_e and together with the expert underlying variance σ_e^2 the expert distributions are generated.

Example Let us look at three experts, all with underlying mean zero and variance one. Let the correlation between the medians of expert 1's distributions be 0.1, for expert 2's be 0.5 and expert 3's distributions be 1 (completely dependent observation errors.) Say, we have 3 seed variables and one target variable, the the covariance matrix S_1 for expert 1 is:

$$\mathbf{S_1} = \begin{pmatrix} 1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 1 \end{pmatrix} \tag{3.6}$$

Then the medians of the distributions the expert reports for each variable are sampled from the multivariate normal distribution with covariance matrix S_1 and zero mean.

Positive and Negative Correlation

A correlation close to zero indicates that the two variables are unrelated. A positive correlation indicates that the two variables move together, and the relationship is stronger the closer the correlation gets to one. A negative correlation indicates the two variables move in opposite directions, and that relationship also gets stronger the closer the correlation gets to minus 1. In the simulations we have only considered positive correlation between the medians of an expert distribution. Negative correlation is also something to investigate but could in our simulations lead to a non positive (semi-)definite covariance matrix from which we cannot sample. In section 3.1.7 the results are shown of some negatively correlated experts.

3.1.4 Summary Model I

- The true values of all variables are assumed to be zero.
- For each variable j expert e reports five quantiles of a normal distribution $\mathcal{N}(x_{e,j}, \sigma_e)$.
- The medians of the distributions $x_{e,1}, \ldots, x_{e,N+1}$ are sampled from the expert underlying characteristic distribution $\mathcal{N}(\mu_e, \sigma_e)$.
- When incorporating some correlation between the medians of the expert distribution, we will have to draw realisations $x_{e,1}, \ldots, x_{e,N+1}$ from a multivariate normal distribution.

3.1.5 Implementation of the Classical Model with Synthetic Data

Expert e's assessment for variable j are the five quantiles of a normal distribution that has the parameters $x_{e,j}$ and σ_e^2. For each variable and each expert quantile points, $f(j, e)$ is the minimal information density function fitted to expert e's quantiles for variable j. With this minimal information density the Calibration is computed. The Information is the average Information score with respect to the background over all seed N variables. The product of these two scores is the unnormalised weight per expert and normalising the weights lead to the final performance based weight for each expert.

3.1.6 Background

The Classical model assumes that there is no prior information on the expert distributions. Therefore the background measure is chosen either uniform or loguniform on the intrinsic range. Recall that the intrinsic range is the smallest interval containing all expert quantiles and the realisations of the seed variables plus a ten% overshoot on both sides. In our model for producing synthetic expert distributions however we do make a prior assumption about the expert distributions. Namely that all expert distributions are normal. This means that a normal background would be a better choice. However, since we are interested in the performance of the Classical Model we have used the uniform background. This is a slight disadvantage of the Classical Model. Figure 3.1 shows $N + 1$ expert distributions generated as described above and respectively the uniform and normal background on the intrinsic range. In **??** the authors propose to find the parameters of the normal background density function as a solution of an optimisation problem. In order to find a normal background density that minimises the Relative Information of the experts together with respect to the background we must maximise the following expression:

$$\prod_{j=1}^{E} \prod_{i=1}^{4} \left(\int_{x_{q_i}^j}^{x_{q_{i+1}}^j} f(x)dx \right)^{\Delta q_i} \tag{3.7}$$

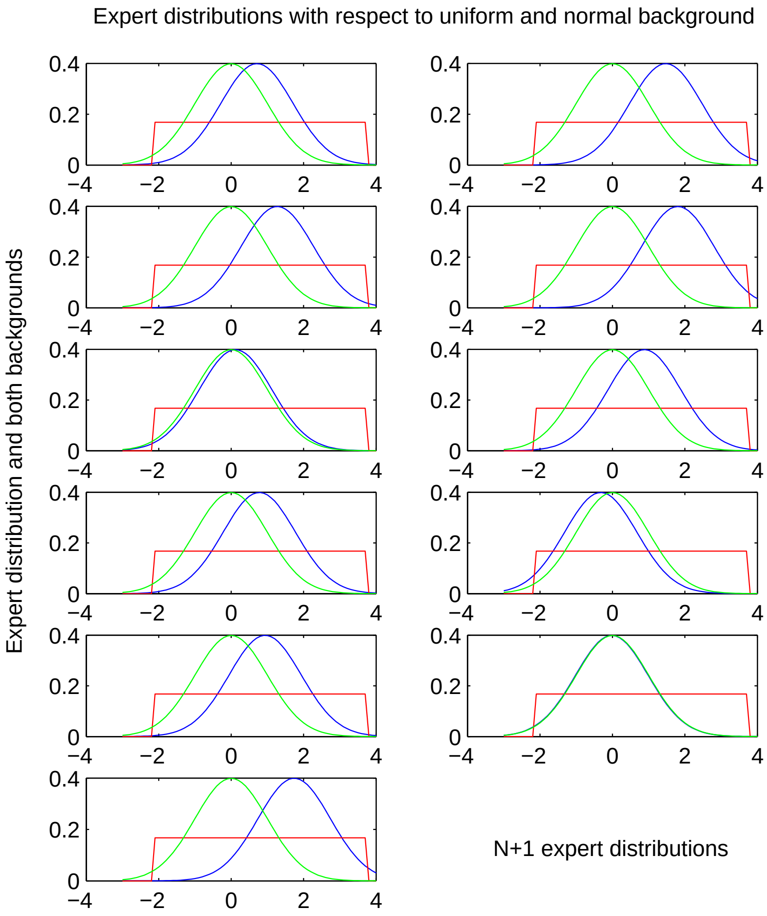

Figure 3.1: Expert distributions with respect to both uniform and normal background.

where $f(x)$ is the normal density function, q_i's are the quantile points for seed variable i and $x^j_{q_i}$ is expert j's assessment for quantiles q_i.

The values of the relative information between the expert distributions and both backgrounds corresponding to the expert distributions shown in figure 3.1 are listed in table 3.1 As can be seen if table 3.1 the Relative Information of the same expert distributions with respect normal background is on average higher than the Relative Information with respect to the uniform background. This means that we disadvantage the Classical Model by using the uniform background whilst assuming normal expert distributions. There is a theoretical explanation given in AppendixC.

Table 3.1: Relative Information for both uniform and normal backgrounds

Background	
Uniform	Normal
0.0536	0.0356
0.0559	0.1491
0.0547	0.1125
0.0585	0.2188
0.0559	0.0011
0.0536	0.0545
0.0535	0.0423
0.0595	0.0068
0.0536	0.0616
0.0570	0.0000
0.0580	0.2043
Average	
0.0558	0.0806

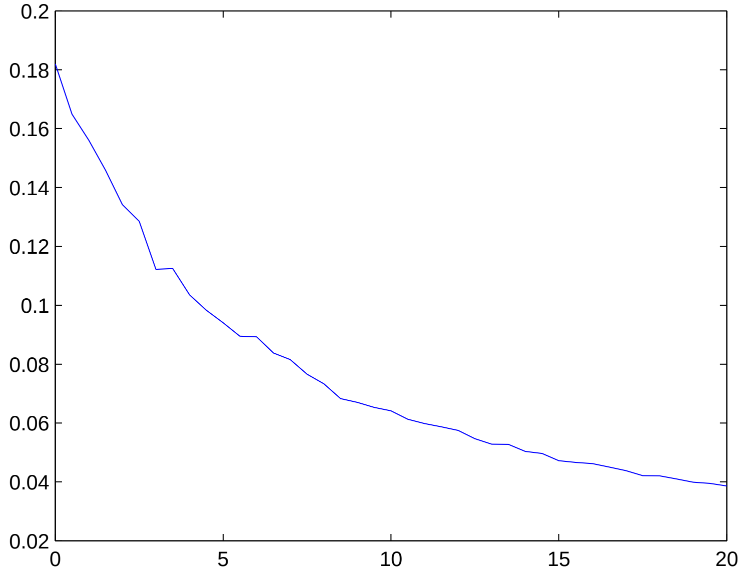

Figure 3.2: Expert's weight as function of his underlying mean.

3.1.7 Results Modelling Experts with Synthetic Data

The above describes a model to generate synthetic data. To this data we can apply the Classical Model. This mainly gives more insight in the statistical behaviour of the Classical Model. Here are some of the results where unless noted otherwise, the number of draws and seed variables is ten and there is one target variable.

Mean and Variance Changes in the experts underlying mean and variance reflect in changes in his rewarded weight. A high underlying mean means that the expected value of the medians of the expert distribution is also greater than zero. It is in fact μ_e. Having an expert distribution with positive (negative) median can still lead to a good relative information score. On calibration however the expert will perform worse with higher (lower) medians than with medians close to zero.

In figure 3.2 the experts weight decreases when the mean of his underlying normal distribution increases. The mean of his underlying distribution characterises the expert's bias. An increasing underlying mean μ_e means that the medians of the expert's distributions are likely to be bigger than zero. Having reported a distribution for the seed variables with positive medians leads to lower weights.

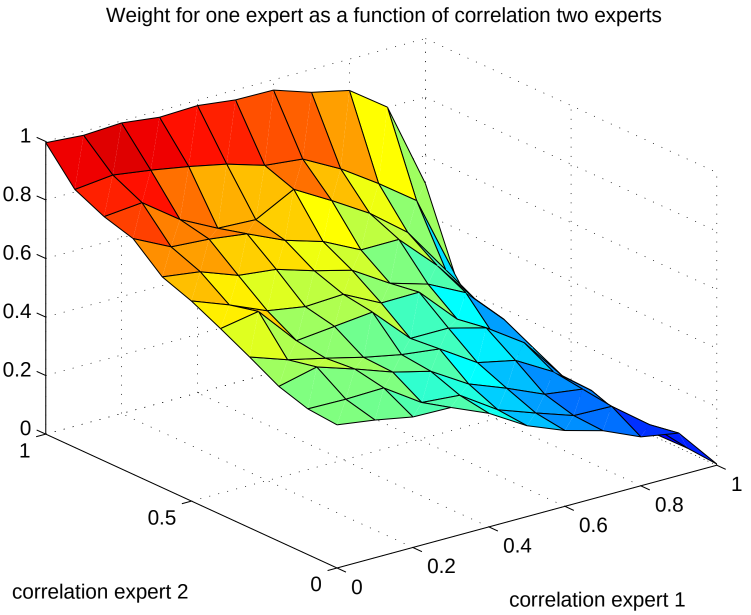

Figure 3.3: Expert 1's weight as function of two experts' correlations.

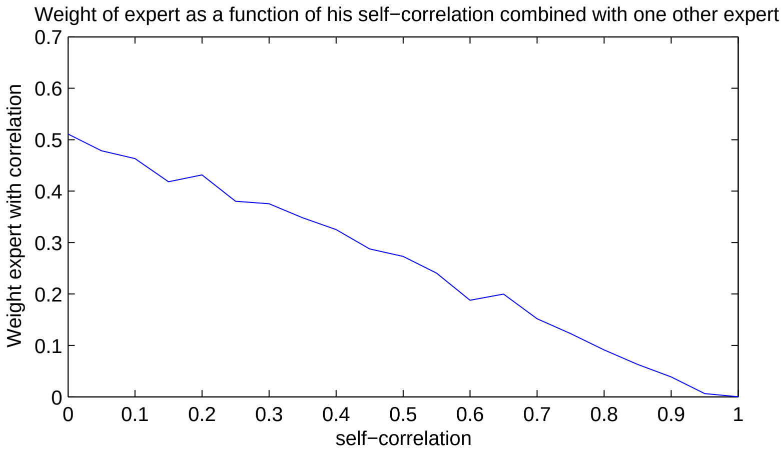

Figure 3.4: Expert weight as a function of his self-correlation - 100 runs

Correlation In section 3.1.3 we introduced the so-called self-correlation ρ_e. This is the correlation between the medians of the expert distributions. Each expert has his own specific underlying correlation, his self-correlation. One would expect that when an expert's self-correlation is closer to one this will result in a lower weight. The self-correlation mainly influences an expert's Calibration score. Recall from Chapter 2 that the Calibration score is computed by testing the hypothesis that the expert distributions are such that the realisations of the N seed variables are an independent sample from the multinomial distribution with parameter P. A high positive correlation between the medians of the expert distributions will in fact reduce the effective number of samples. This is equivalent with scoring the expert on too many seed variables. Instead of a χ^2-test of $2NI(S,P)$ one should reduce M. Not doing this leads to a calibration score that is too low.

Figure 3.3 shows the weights of each expert when considering two experts as a function of the correlation between their expert distribution medians. The two experts have identical underlying characteristic distributions. In the upper left corner of the figure, expert 1 has correlation zero and expert 2 has correlation one. It shows that expert 1 who is the better one in terms of uncorrelated medians of his distributions and equally good in terms of underlying bias receives weight one. In the lower left corner both experts have correlation zero and thus receive weight 0.5. Finally, the lower right corner of figure 3.3 shows that in the situation of a completely correlated expert 1 and a uncorrelated expert 2, expert 1 receives zero weight. For completeness, figure 3.4 confirms that an experts performs worse and consequently receives a lower weight as the correlation between the medians of his distributions increases.

To illustrate the influence of experts' underlying correlation we will look at how the Calibration changes as a function of this self-correlation. Since increasing the underlying correlation decreases the number of effectively independent samples within the χ^2-test, the test might need a correction. Higher self-correlation means that we are no longer testing whether the expert distributions are such that the N realisations of the seed variables are N independent samples from the multinomial distribution.

In figure 3.5 and 3.6 an expert's calibration using ten seed variables is displayed as a function of the self-correlation. The correlation is averaged over the number of simulation runs. It seems that the relation between the self-correlation and the Calibration is linear. Does this mean that we could correct N somehow in order for the χ^2-test to remain reasonable?

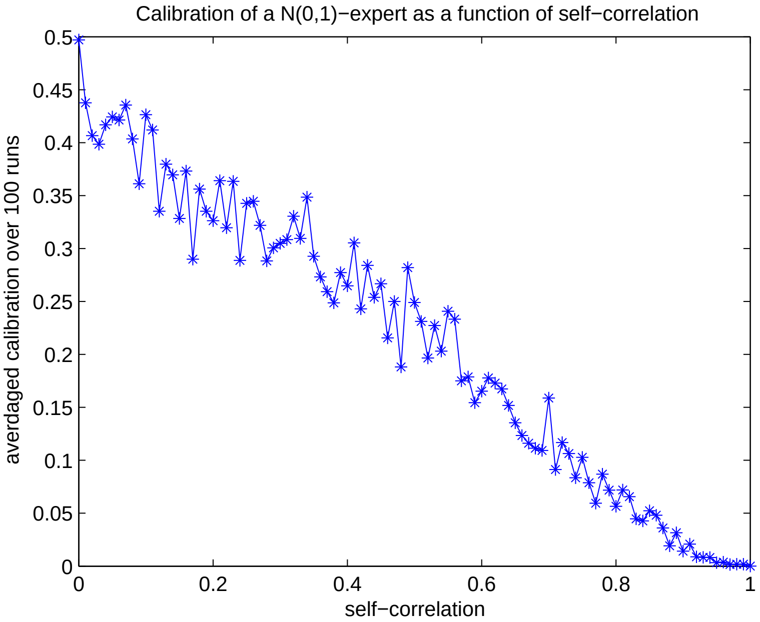

Figure 3.5: Calibration as a function of self-correlation.

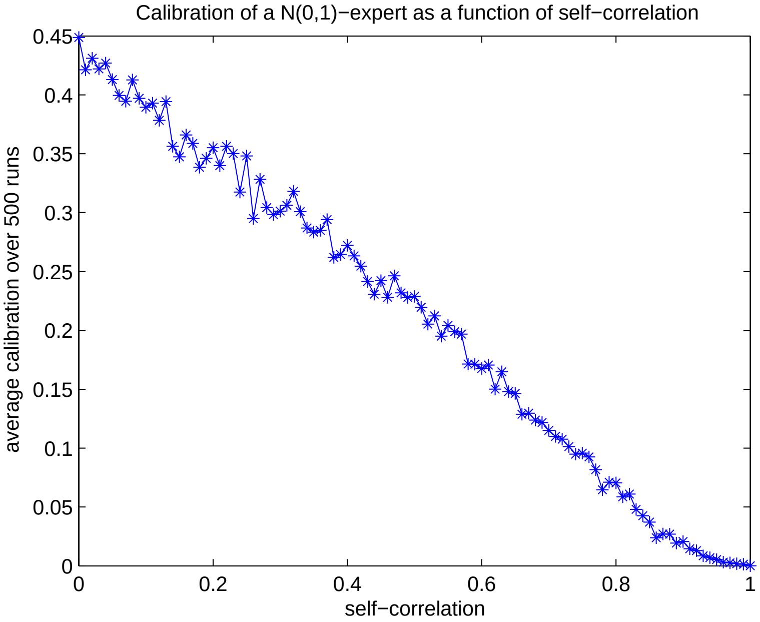

Figure 3.6: Calibration as a function of self-correlation - more runs.

3.2 Probability Density of the Unnormalised Weight

Typically Expert Judgement studies at the TU Delft use around ten experts and seed variables and all use the Classical Model to weigh and combine the experts. It is unknown whether this number is sufficient to obtain a stable combined judgement. It is possible that adding one single expert to ten experts changes the final judgement for some unknown variable significantly. An interesting question to be answered is the matter of convergence of the Classical Model.

One approach can be to consider the following: How sensitive is the combined distribution e.g. the distribution for a target variable to adding another expert as a function of the characteristics of that expert?
In other words: How much influence has a new expert on the final combined judgement and how does this influence depend on his characteristics and that of the other experts?

As a measure for the sensitivity of the combined distribution to adding new experts we have taken the relation between the unnormalised weights of the new expert and that of the previous experts. Let us consider for example two experts who both have an equal unnormalised weight. Adding a new, third expert who has a significant higher unnormalised weight can result in a completely different combined distribution. In this case the third expert will add new information to the combined distribution. If however the new expert has a much lower unnormalised weight he has less influence on the combined distribution.

Eventually, we are interested in the influence of adding new experts onto the Decision Maker's distribution and how this depend on the experts characteristics. Measuring this using the unnormalised weight, we would like to know the dependence of the unnormalised weight on expert characteristics. We will therefore construct the probability density of the unnormalised weight depending on the expert characteristics as input variables. This will enable us to compare the influence of different experts or adding expert depending only on their characteristics.

3.2.1 Quantifying Expert Performance with his Underlying Distribution and Bins

In the previous section, a model was proposed to generate synthetic expert distributions. In this section we will continue with this model. Recall that an expert-specific distribution represents the expert characteristics. Each expert is characterised by a normal distribution which is parameterised by his underlying mean and variance, μ_e and σ_e^2:

$$\mathcal{N}(\mu_e, \sigma_e^2) \tag{3.8}$$

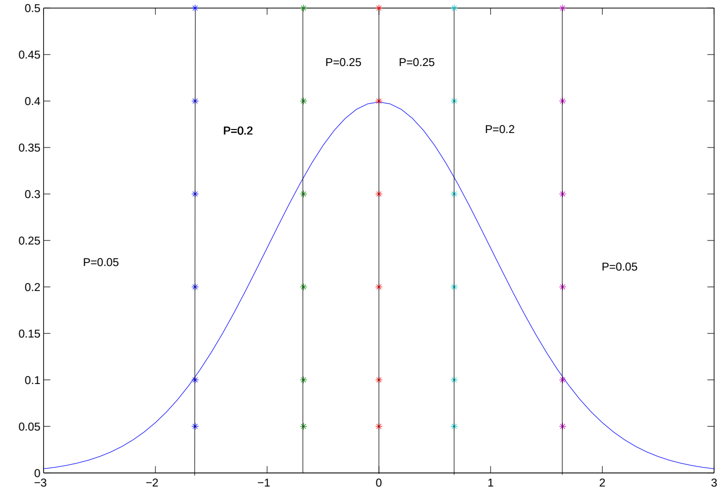

Figure 3.7: Bins in which the Seed Variables can fall

The realisations of the seed variables are still zero and the expert distributions are assumed normal distributions.

Bins

To construct the probability distribution for the unnormalised weight the six intervals between the five quantiles of each distribution are represented as six bins. The bins of each expert distribution are the intervals in which the realisation of the seed variables can fall. Figure 3.7 shows the six bins for an expert distribution with median zero and variance one. Now, before generating the expert distributions for each seed variable, his underlying distribution already determine the stochastic process of determining the positions of the bins. For example, figure 3.8 shows that the value zero of a seed variable falls into another bin if the median of the expert distribution would shift from zero to two. Thus, the bins in which the seed variables fall are completely determines by the expert's underlying distribution. Let us illustrate this with an example:

Example We consider an expert with standard normal underlying distribution. $\mathcal{N}(\mu_e, \sigma_e^2) = \mathcal{N}(0, 1)$. The medians of this expert, $x_{e,i}$'s are all realisations of his underlying distribution. Now, we consider one seed variable and median x_e. The probability that the median of the expert distribution for this seed variable is such that its realisation falls into the first bin is also completely

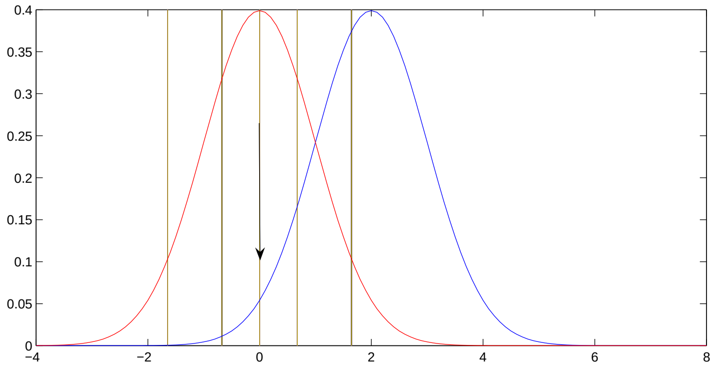

Figure 3.8: Bins as a function of shifting underlying bias.

determined by the expert's underlying distribution:

$$\begin{aligned}
&P(\text{seed variable} \in \text{bin}_1) \\
&\\
&= P(x_{q_1} > 0 | X_e \sim \mathcal{N}(0,1)) \\
&= P(x_e > x_{q_5} | X_e \sim \mathcal{N}(0,1)) \\
&= 1 - P(x_e \leq x_{q_5} | X_e \sim \mathcal{N}(0,1)) \\
&= 1 - q_5 \\
&= q_1
\end{aligned} \tag{3.9}$$

where q_i denotes the ith quantile (e.g. $0.05,\ 0.25,\ \ldots,\ 0.95$) and x_{q_i}denotes the value of the expert distribution at that quantile.

Now, we have illustrated how the underlying distribution determines the positions of the bins and therefore determines the probabilities with which the realisations of the seed variables fall into the bins. Knowing the probability with which the seed variables fall into the bins leads to quantifying the expert's performance as described in the next sections.

Using bins to construct the desired PDF

In order to quantify the expert distributions, we will first describe the construction of frequency vectors. Frequency vectors represent the bin in which the seed variable falls. The stochastic process for a seed variable falling into one of the six bins is the multinomial distribution and will be explained in section 3.2.3. The multinomial distribution gives each possible frequency vector a probability. In 3.2.5 will be explained how each frequency vector leads to a value for the calibration and relative information. The calibration and relative information determine the unnormalised weight and together with the corresponding probabilities this leads to a probability density for the unnormalised weight.

To determine the probability density of the unnormalised weight we consider a single expert to begin with. The distribution however depends on other experts since the relative information is calculated with respect to the background measure. The background measure depends on the distributions of all experts and is explained in 3.2.5

3.2.2 Frequency Vector

The intervals between the quantiles of the unknown expert distribution are represented as six bins. Let us consider the six bins represented by S_1, S_2, S_3, S_4, S_5 and S_6 and N seed variables. When a seed variable would fall into bin i S_i is increased by one, otherwise S_i remains the same. Then we introduce the frequency vector which contains all $S_i's$: $S = [S_1 \quad S_2 \quad S_3 \quad S_4 \quad S_5 \quad S_6]$. Taking into account N seed variables increases the number of frequency vectors very quickly. Each of the possible values of the frequency vector occurs with a probability depending on the parameters of the expert's underlying distribution: $\mathcal{N}(\mu_e, \sigma_e^2)$.

3.2.3 Probabilities

Recall that the coordinates of the six bins are random and the underlying stochastic process is determined by the expert characteristic distribution.

The bins in which the seed variable falls depends on the underlying distribution. Assuming that each seed variable has true value zero there are many different possible frequency vectors. Each frequency vector has a probability of occurring and leads to a different value of the unnormalised weight. The stochastic process describing seed variables falling into one of multiple bins is the multinomial distribution.

Multinomial Distribution

The multinomial distribution is a discrete distribution and an extension of the binomial distribution involving joint probabilities. It involves a similar statistical experiment, but this time there are more than two possible outcomes. Specifically, each trial can result in any of the k events $E_1, E_2, \ldots, E_k$, with respective probabilities $P_1, P_2, \ldots, P_k$. In this case, the multinomial distribution is the joint probability distribution of the set of random variables $X_1, X_2, \ldots, X_k$, where X_i is the number of occurrences of $E_i, i = 1, 2, \ldots, k$ in n independent trials. It has a probability mass function of the following form:

$$P(x_1, x_2, \ldots, x_k | P_1, P_2, \ldots, P_k, n) = \binom{n}{x_1 x_2 \cdots x_k} P_1^{x_1} P_2^{x_2} \cdots P_k^{x_k} \qquad (3.10)$$

where

$$\sum_{i=1}^{k} x_i = n, \qquad \sum_{i=1}^{k} P_i = 1$$

In other words, the multinomial distribution gives the probability of choosing a given collection of m items from a set of k items with repetitions and the probabilities of each choice given by $P_1, \dots, P_k$. These probabilities are the parameters of the multinomial distribution. Conjugate prior of the parameters of the multinomial distribution is the Dirichlet distribution. **??**

Parameter of the Multinomial Distribution

The intervals between the five quantiles of the possible expert distribution are represented as six bins in which the true value of the seed variables could fall. The probability model for this process is thus a multinomial distribution with parameter P. The probability vector P depends on the characteristics of the expert and is recalculated using the expert underlying distribution. Each parameter of the multinomial distribution P is calculated in the following way:

Recall from the example in section 3.2.1 that if an expert underlying bias is zero then the probabilities with which the seed variables fall into the various bins are the standard sizes of the bins. The probability vector is $P_{standard} = [\ 0.05 \quad 0.2 \quad 0.25 \quad 0.25 \quad 0.2 \quad 0.05]$. If however the expert underlying bias is unequal to zero, the probability vector $P_{standard}$ changes into P_{biased}. The x-coordinates of the quantiles and thus the bins shift to the right with the underlying bias μ_e. The new x-coordinates are therefore:

$$\underline{x}_{new} = \ [x_{0.05,old} + \mu_e \quad x_{0.25,old} + \mu_e \quad x_{0.5,old} + \mu_e \quad x_{0.75,old} + \mu_e \quad x_{0.95,old} + \mu_e]$$

and the new probability vector P_{biased} are the increments of the cumulative density function of x_{new} with parameters μ_e and σ_e^2. This means that when an expert has an underlying bias larger than zero, the probability that the median of the expert distribution is such that the first bin contains value zero increases and the probability that the last bin contains value zero decreases.

3.2.4 Calculating the Density of the Frequency Vector

There is an underlying stochastic process for the bins in which the seed variables fall. This means that the the parameter P for the multinomial distribution is equal for each seed variable. Since the true values of the seed variables is always zero, the frequency vector has the following form: zeros on all but one entry and one on the remaining entry.

The probability distribution for the number of seeds variables that fall into the various bins thus follows from the multinomial distribution:

$$\begin{aligned} P(S_1 = s_1, S_2 = s_2, \dots, S_{n+1} = s_{n+1}) = \\ \frac{N!}{s_1! s_2! \cdots s_{n+1}!} P_1^{s_1} P_2^{s_2} \cdots P_{n+1}^{s_{n+1}}, \end{aligned} \qquad (3.11)$$

where $\mathbf{S} = (S_1, S_2, \ldots, S_{n+1})$ is the vector of numbers of seed variable realisations that fall into each of the bins and P is the probability vector. All possible combinations of seed variables in the various bins, S and the multinomial distribution together result in the (discrete) probability density for the frequency vectors of the seed variables.

The outcome of the multinomial distribution gives a probability for each of the possible frequency vectors. The frequency vectors all lead to a different value of the unnormalised weight and together with the corresponding probabilities this leads to the desired PDF of the unnormalised weight.

3.2.5 Probability Density of the Unormalised Weight

The previous section described how we have come to all possible frequency vectors and their corresponding probabilities of occurrence. Each frequency vector leads to a value for the relative information and the calibration in a way that will be explained here. The product of these values is the unnormalised weight. Each frequency vector has a specific probability and therefore the corresponding values of the calibration and relative information leading to the unnormalised weight also have this probability. Together, this leads to a probability distribution of the unnormalised weight. First we will show how each frequency vector S_i leads to a value of the unnormalised weight:

Calibration

Using a recursive algorithm we will obtain all possible frequency vectors as explained in 3.2.2 The frequency vector represents the number of seed variables that fall into one of the six bins. The calibration for each frequency vector S_i is now calculated as:

$$C_i = 1 - \chi^2(2N \sum_{i=1}^{6} \frac{S_i}{N} \ln \left(\frac{S_i/N}{P_{i,standard}} \right), 5) \tag{3.12}$$

Here $P_{i,standard}$ is the standard probability vector for the multinomial distribution as explained in 3.2.3.

Relative Information and Background Measure

The relative information is calculated with respect to a background measure. Within the Classical Model this is either the uniform or loguniform density on an intrinsic range. Recall from Chapter 2 that the intrinsic range is the interval of minimal length that contains all expert assessments and all realisations of the seed and target variables plus a ten percent overshoot. It is not possible to specify the intrinsic range up front. It depends on the expert distributions. In our model however we do not specify the expert distributions but we still would like to be able to determine the intrinsic range. this is necessary for computing the unnormalised weights. In the simulations, the parameters of the expert's underlying distribution are bounded by μ_{max} and σ_{max}. In practice it appears to works well if we specify the intrinsic range as follows:

$$I = [1.1(CDF^{-1}(0.95|\mu_{max}, (\sigma_{max}+1)^2) - CDF^{-1}(0.05|\mu_{max}, (\sigma_{max}+1)^2))]$$

Since the parameters of the expert underlying distributions are bounded by μ_{max} and σ_{max} and their values are $\ll 10$, a suitable background measure will always be the uniform density as opposed to the log uniform density over the fixed intrinsic range I as specified above.

The relative information for expert e and the variable i in the Classical Model is:

$$I(f_{i,e}, g_i) = \sum_{k=1}^{n} f_{i,e}(x_k) ln \left(\frac{f_{i,e}(x_k)}{g_i(x_k)} \right) \tag{3.13}$$

where $f_{i,e}$ is the minimal information density function fitted to the expert's quantiles and g_i is either the uniform or the loguniform density function depending on the scale of the variable. In our case this g_i is thus the uniform density function.

The unnormalised weight is the product of the expert's relative information and the calibration. An important fact to keep in mind is that the relative information of each expert depends on the other experts. This is a problem in our simple model of one expert. Instead of modelling the sample distribution for the unnormalised weight as a function of the characteristics of one expert, we should take into account the assessments of all other experts. The question is now whether it is possible to make an assumption about the influence of the tails of the background measure?

Let us look at the continuous equivalent: Let S be the fitted density to expert quantiles and assume it is the standard normal density function. Let P be the background measure uniform on $[-b, b]$ where b is unknown. Then the relative information is given by:

$$\begin{array}{l} R.I. = \int_{-b}^{b} S(x) \ln\left(\frac{S(x)}{P(x)}\right) dx = \\ \int_{-b}^{b} S(x) \ln\left(\frac{S(x)}{\frac{1}{2b}}\right) dx = \\ \int_{-b}^{b} S(x) \ln(S(x)) + S(x) \ln(2b) dx = \\ \int_{-b}^{b} S(x) \ln(S(x)) dx + \ln(2b) \int_{-b}^{b} S(x) dx \end{array} \tag{3.14}$$

Since $S(x)$ is a density function, the second integral of the last equation will go to $\ln(2b)$ as b becomes large and the first integral will go to zero. This means that there remains a factor $\ln(2b)$. This shows that the influence of b remains even if the tails of the uniform distribution get very large. The simple model however only takes into account one expert. Conclusion from figures: the value of b does affect the relative information, but will do so in a constant way. Therefore the model will use one expert and a fixed background.

Unnormalised Weight When the Calibration and information score have been computed for the set of frequency vectors, the unnormalised weight for each frequency vector is simply the product of the two scores. Finally, the set of unnormalised weights and the set of probabilities, both corresponding with the set of frequency vectors determine the discrete probability density function of the unnormalised weight of one expert as a function of his underlying distribution.

3.2.6 Results and Conclusions

In the previous section we have explained the construction of the probability density of the unnormalised weight depending on expert characteristics. Recall from the introduction that the main reason to construct the probability density of the unnormalised weight was our interest in the influence of adding new experts on the Decision Maker's distribution and how this depends on expert characteristics.

Here are some results about the relation between the unnormalised weight and expert characteristics. The results are subdivided into prototype experts. We look at the standard expert, the extreme experts and some variations in between. The standard expert has an underlying $\mathcal{N}(0,1)$ distribution and the extreme expert has an underlying $\mathcal{N}(3,4)$ distribution. These numbers come from the bounds μ_{max} and σ_{max}. Varying the underlying mean from the standard to the extreme mean and the same for the variance gives the variations in between the two extremes.

Standard Expert

Figure 3.9 and 3.10 show the discrete probability function of the unnormalised weight for the standard expert. The first figure shows the possible values for the unnormalised weight for only one seed variable and the second one for ten seed variables. Figure 3.9 displays three different values for the unnormalised weight. To explain this we need to realise that the calibration score is symmetric: If an expert assesses a distribution such that the seed variable falls into the first bin of his distribution he is just as poorly calibrated as when he assesses a distribution such that the seed variable falls into the last bin. Therefore six bins lead to three different values of the unnormalised weight.

Figure 3.10 shows many different values for the unnormalised weight, but shows the same spread and structure as the first graph: The first bar in figure 3.9 counts for the many small bars for low values in figure 3.10 and so on. The expected unnormalised weight in case of ten seed variables however is lower than when using one seed. The reason that the expected weight is lower than in case of one seed variable is the expected calibration depends on the χ^2-test that is performed.

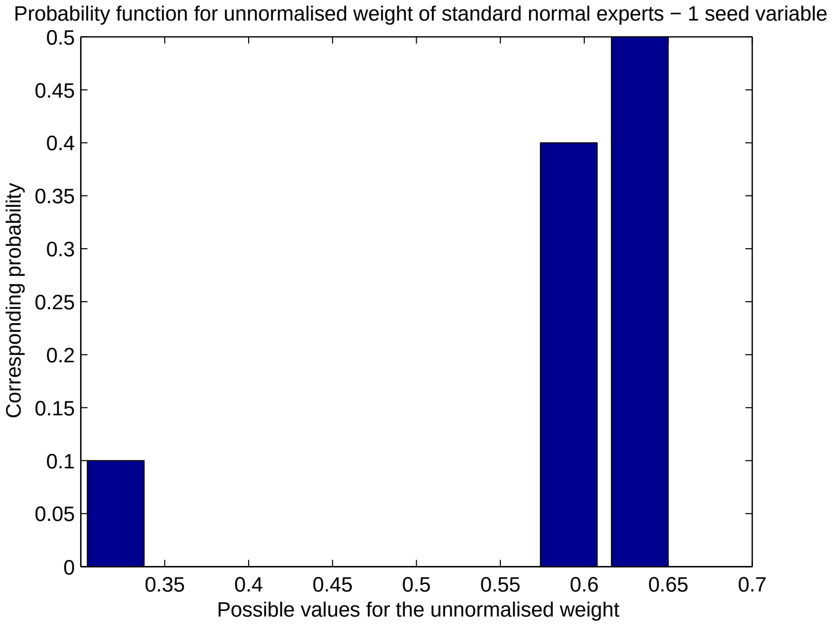

Figure 3.9: Discrete Density for Weight for Standard Normal Expert

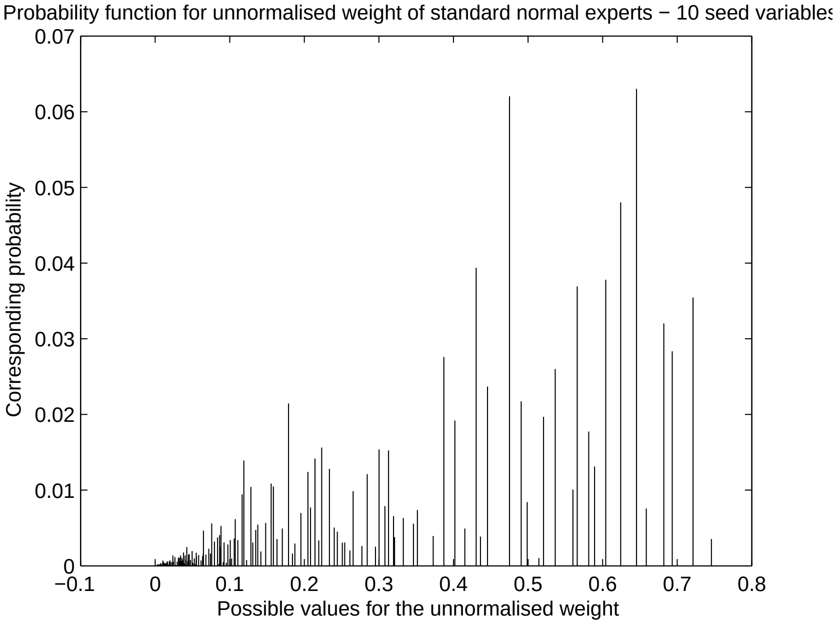

Figure 3.10: Discrete Density for Weight for Standard Normal Expert

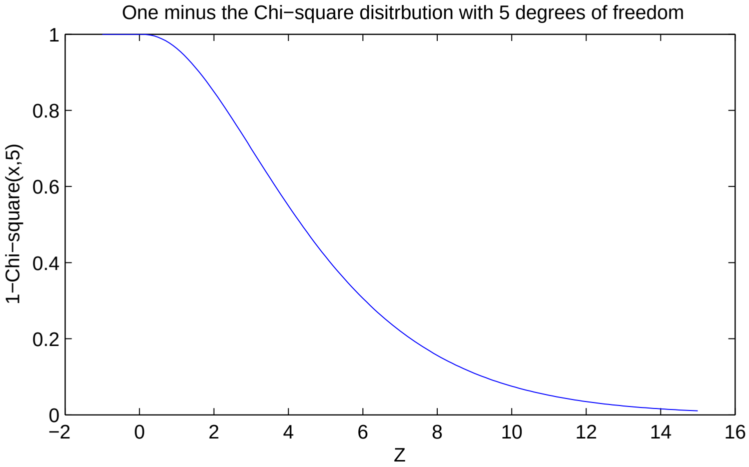

Figure 3.11: 1-Chi-square distribution with 5 degrees of freedom

Let Z be:

$$Z = 2N \sum_{i=1}^{6} \frac{S_i}{N} \ln\left(\frac{S_i/N}{P_{i,standard}}\right) \tag{3.15}$$

If Z is a χ^2_K-distributed variable, the expectation of Z is:

$$Z \sim \chi^2_K : \quad \mathrm{EZ} = \mathrm{K} \tag{3.16}$$

When the number of seed variables increases, the number

$$2N \sum_{i=1}^{6} \frac{S_i}{N} \ln\left(\frac{S_i/N}{P_{i,standard}}\right) \tag{3.17}$$

increases as well. If Z is χ^2_K, the expected value of Z remains K no matter how many seed variables there are. The calibration score is:

$$1 - \chi^2(Z, 5) \tag{3.18}$$

Expression 3.2.6 decreases when Z increases as is illustrated by figure 3.11. This explains why the expected weight decreases with increasing number of seed variables.

Extreme Expert

We continue with considering only one seed variable. This results in clear graphs. Figure 3.12 shows the discrete probability function of the unnormalised weight for the most extreme expert we consider. This expert has an underlying mean of three and variance of four. The expected weight for the most extreme expert is very low compared with the standard expert. Adding the extreme expert to standard normal experts will thus change approximately nothing about the DM's distribution.

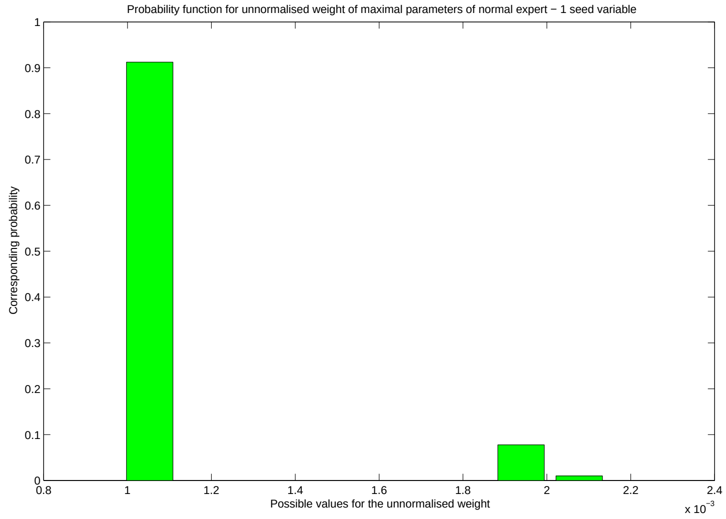

Figure 3.12: Probabilities for unnormalised weight for maximal parameters of expert distribution.

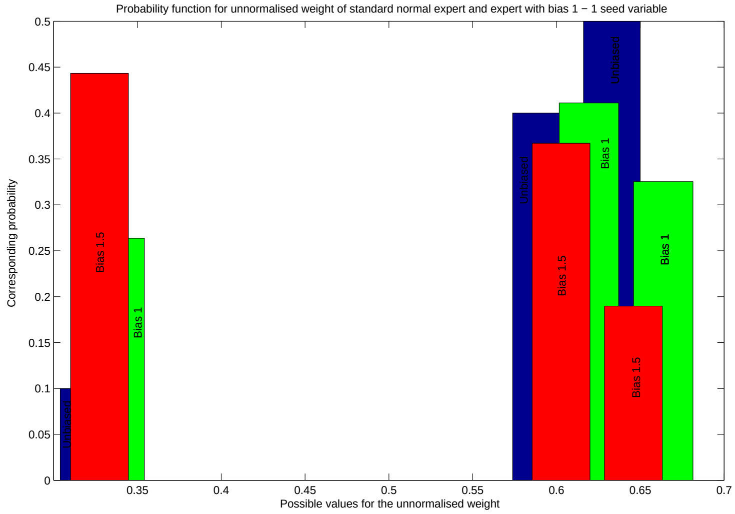

Figure 3.13: Probabilities for unnormalised weight when bias is increasing from 0 to 1 to 1.5

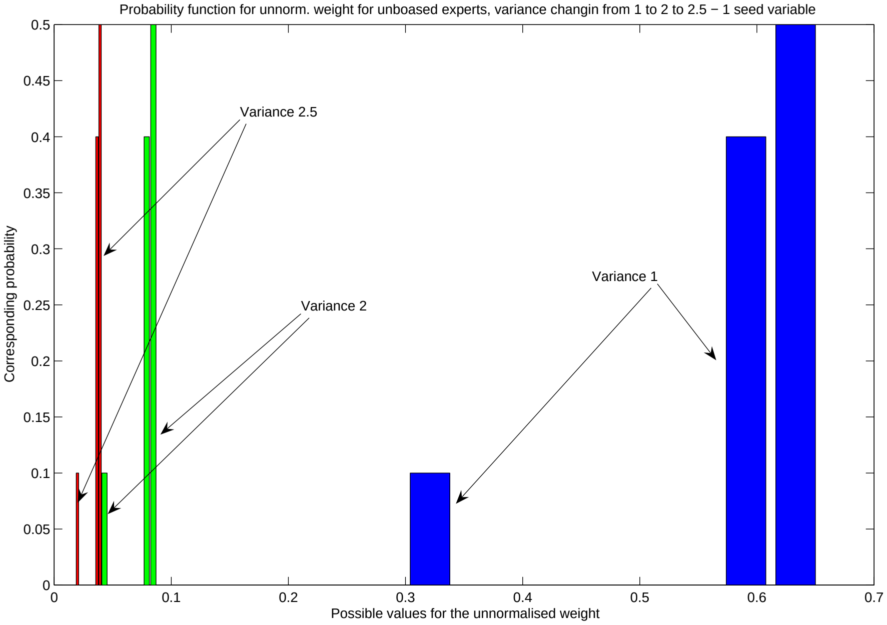

Figure 3.14: Probabilities for unnormalised weight when variance is increasing from 1 to 2 to 2.5

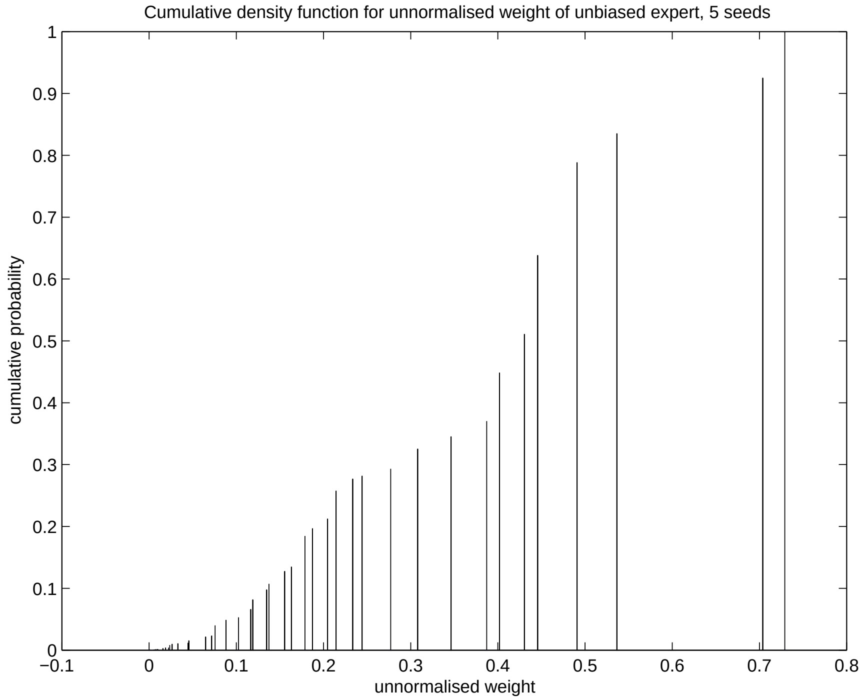

Figure 3.15: Cumulative distribution function for unnorm. weight standard normal expert - 5 Seeds

Variations between Standard and Extreme Experts

Now we can ask ourselves what will be the difference in the influence of adding an expert with a higher underlying mean μ_e and adding an expert with a higher underlying variance σ_e^2? Figure 3.13 shows the probability function of the weight for three different experts. All three experts have the same underlying variance. Their underlying mean however changes. The probability functions show a shift from a high probability of a value around 0.6 to a high probability of a value around 0.3 twice as low.

To compare the influence of the expert underlying bias and variance, figure 3.14 again shows the probability function for three different experts. The three experts all have underlying zero mean. Their underlying variances change with the same increments as the underlying mean in figure 3.13. The probability functions show a much quicker shift to low unnormalised weights. The corresponding probabilities however are the same. This means that a larger spread in the expert distributions has more influence than his bias.

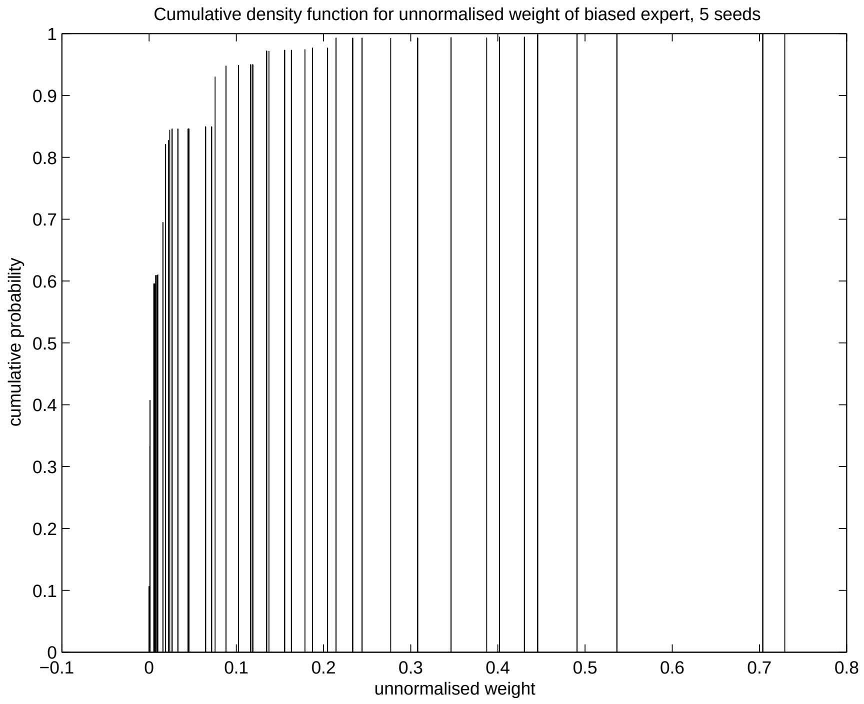

Figure 3.16: Cumulative distribution function for unnorm. weight most extreme expert - 5 Seeds

CDF of the Unormalised Weight

Until now we have given results of the probability density function of the unnormalised weight. Figure 3.15 and 3.16 however show the cumulative density of the unnormalised weight for five seed variables. The first figure shows the cumulative probability function for a standard normal expert. It can be seen that the CDF is approximately linear. The second figure corresponds to the most extreme expert, the one with underlying maximal mean three and variance four. His CDF is not linear.

Thus, the cumulative density function of an expert with underlying zero mean seems to approach a linear function. Calculating the CDF with more seeds show that the asymptotic CDF indeed is linear. This result corresponds with the underlying theory: Let $S^{(n)}$ be the minimal information density fitted to the expert distribution. We know that if $S^{(n)} \sim \prod^n P$ the asymptotic distribution of $2nI(S^{(n)}, P)$ is χ^2. Therefore the cumulative density function of the unnormalised weight of an zero-mean expert should simply be the χ^2 distribution of a χ^2 parameter which is asymptotically linear.

Influence on the DM

The questions posed in this section was: How much influence has a new expert on the final combined judgement and how does this influence depend on his characteristics and that of the other experts? When taking the unnormalised weight as a measure of the influence the results have shown that the weight decreases with increasing bias and even more with increasing underlying variance. This means that an expert that gives assessment with a small spread and intermediate bias still may influence the distribution of the Decision Maker in a fair amount.

Chapter 4

Convergence of the Classical Model

Using the same approach the bin-model this chapter shows the results of keep adding experts with the same or different characteristics and aggregate them using the Classical Model. The main purpose is to obtain some interesting results about the convergence of the method.

4.1 Introduction

When performing an Expert Judgement analysis it is important to have some ideas or guidelines on the way the method converges. Is there something we can say about the number of experts from which it does not make a significant difference to add a new one? In other words: Does the Classical Model converges and with how many experts? In this chapter we will address the question of convergence. Both the Decision Maker's characteristics as the relative information of consecutive DM's are used as a measure of convergence.

Outline First section 4.2 gives some results on the influence of more and less biased experts on the aggregated distribution of the Decision Maker. Then instead of considering the experts and their influence on the Decision Maker we consider the Decision Maker himself. Does the distribution of the Decision Maker keep changing after adding a lot of experts? Finally the relative information is used as a measure of convergence of the Decision Maker's distribution when adding new experts followed by conclusions.

4.2 Influence on the DM

In the previous chapter we have constructed a probability density for the unnormalised weight of an expert. This unnormalised weight was taken as a

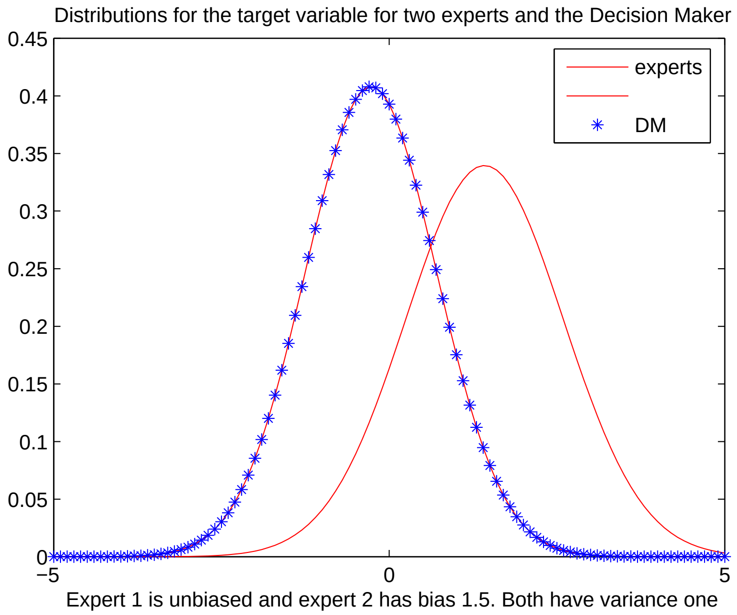

Figure 4.1: Aggregating two experts: A (0,1)-expert and a (1.5,1)-expert - 10 seed variables

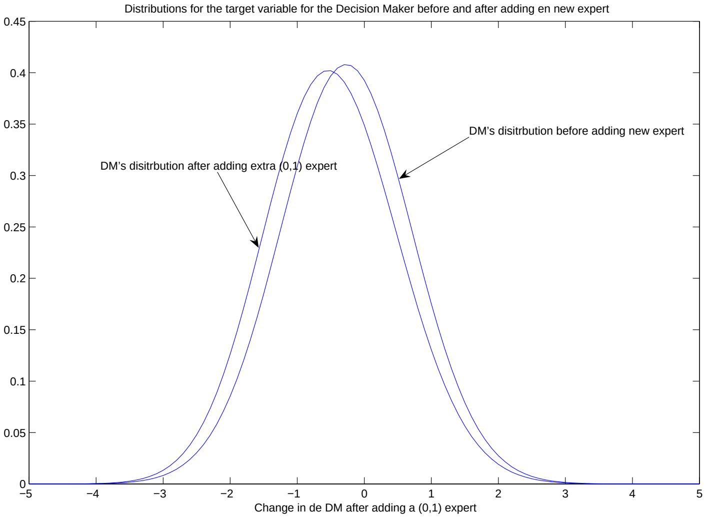

Figure 4.2: The DM's distribution changing with adding a third unbiased expert - 10 seed variables

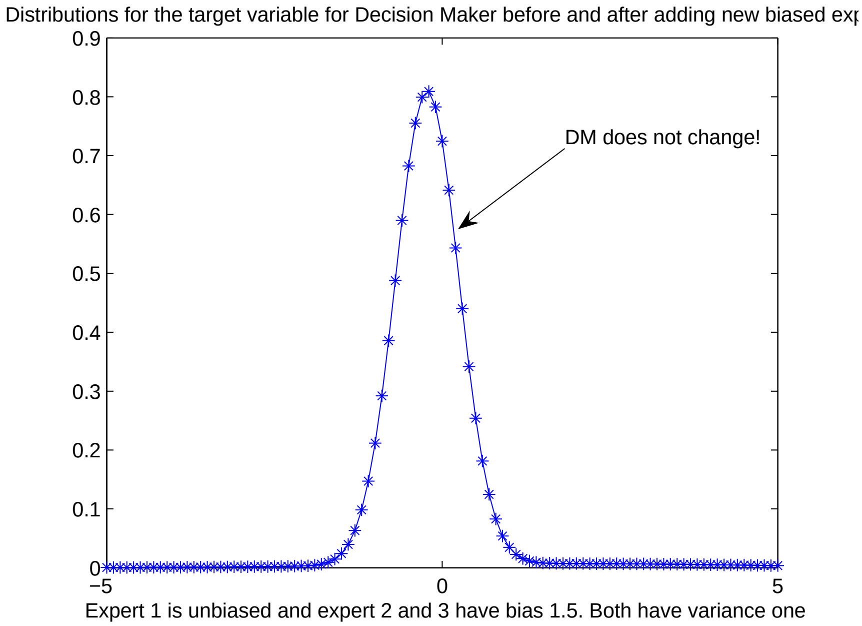

Figure 4.3: The DM's distribution changing with adding a third biased expert - 10 seed variables

measure of influence of an expert on the aggregated distribution of the Decision Maker. The following results will visualise this concept.

Figure 4.1 shows the two distributions of a $\mathcal{N}(0,1)$-expert and a $\mathcal{N}(1.5,1)$-expert and the distribution of their combination using ten seed variables. The figure illustrates that the unbiased expert has a much higher influence than the biased expert. The unbiased expert thus has a much bigger influence as can be seen in the figure. Adding a third unbiased expert, again with higher expected unnormalised weight should change the Decision Maker's distribution. Figure 4.2 shows this indeed. Here only the aggregated distributions of the two Decision Makers are shown. One for the Decision Maker constructed from an unbiased $\mathcal{N}(0,1)$-expert and a biased $\mathcal{N}(1.5,1)$-expert and one for those two experts plus an extra unbiased expert. The aggregated distribution changes, it shifts more to the left. So we have seen that adding a new expert with a high unnormalised weight with respect to the other experts, e.g. the unbiased new expert shows a change in the aggregated distribution of the Decision Maker.

Figure 4.3 on the other hand shows that adding a biased expert with a relative low unnormalised weight does not change the aggregated distribution. This agrees with our choice of taking the unnormalised weight as a measure of the

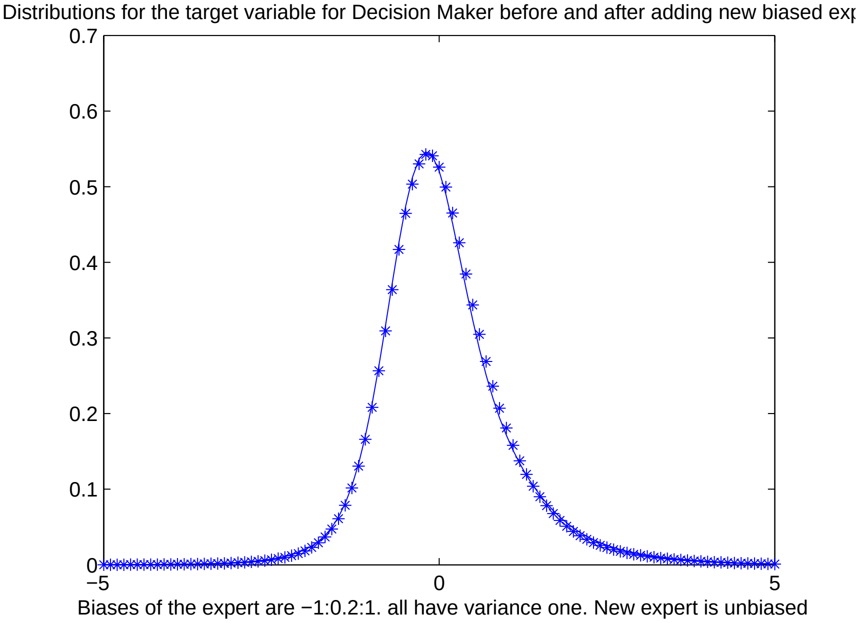

Figure 4.4: The DM's distribution changing with adding an unbiased expert to eleven other more or less biased experts - 10 seed variables

influence of an expert to the distribution of the Decision Maker: the biased expert has a very low expected unnormalised weight thus his expected influence is almost none. **N.B.** Figure 4.3 shows the results of a different simulation than figure 4.2. This explains the difference in values. It is still an illustration of the fact that a Decision Maker from two experts (an unbiased and a biased one) is approximately the same as from three experts (an extra biased one.)

Finally, figure 4.4 shows that adding a new unbiased expert to a Decision Maker who is already based on 11 unbiased experts does not show significant difference. Even thought the new expert is unbiased and has a high expected unnormalised weight. This results leads us at addressing the convergence of the Decision Maker himself.

4.3 Convergence of the Decision Maker

In the previous section we saw that adding an unbiased expert to two experts made more difference in the distribution of the Decision Maker than adding a new expert to eleven unbiased experts. However in order to say something more about the convergence of the Decision Maker himself instead of the influence experts have on the Decision Maker, let us look again at the Bin Model.

4.3.1 The Medians of the Expert Distribution

We thus follow the bin approach described in the previous chapter. Instead of leaving the expert distributions unknown we will draw a realisation for each median of the expert distribution. Recall that for each expert and each seed variable the medians of the expert distribution follows the expert underlying distribution: $\mathcal{N}(\mu_e, \sigma_e)$.

Expert e is characterised by the his underlying distribution $\mathcal{N}(\mu_e, \sigma_e)$. For each seed variable j he reports a normal distribution $\mathcal{N}(x_{e,j}, \sigma_{e,j})$ Each median $x_{e,j}$ is a realisation of the random variable X_e where $X_e \sim \mathcal{N}(\mu_e, \sigma_e)$.

4.3.2 Frequency Vectors

With six bins and a known expert distribution the seed variables with value zero will always fall into the same bin. Therefore once the observation error is drawn from the underlying normal distribution the frequency vector representing the bins in which the seed variables fall is also known. It always has the same form: five zeros and one entry with value one. Which entry has value one depends on the expert disitribution and in our case directly on the median of this distribution. As described in Chapter 3 this frequency vector leads to a value of the calibration, relative information and finally the unnormalised weight. This is a very simple approach, but makes it easy to look at what happens when adding a large number of experts.

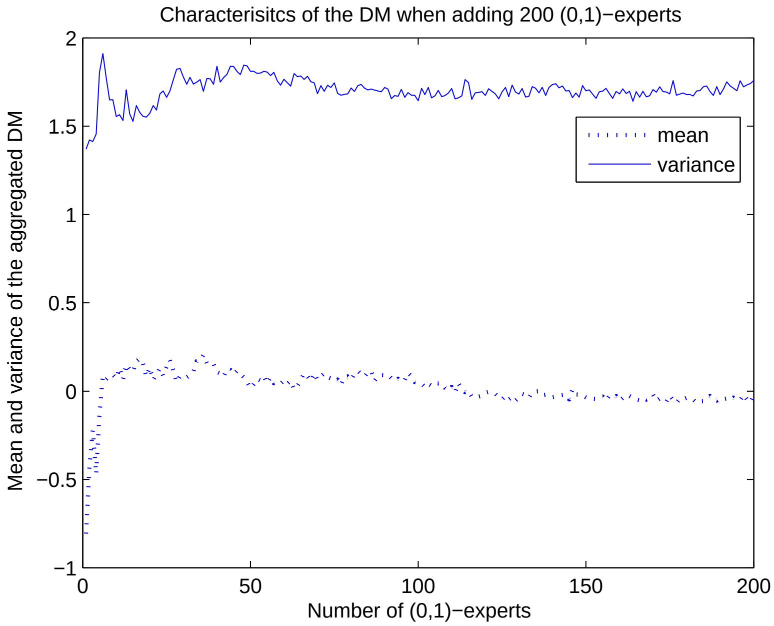

Figure 4.5: Mean and variance when adding 200 new (0,1)-experts.

Now we can use the extended bin model as described above to look at the consequences of adding a large number of more and less similar experts. This will give a suggestion on how the question from the introduction on the necessary number of experts needed for the distribution of the Decision Maker to convert should be answered.

4.4 Convergence Results

Because all the experts have un unbiased underlying distribution. Keeping adding $\mathcal{N}(0,1)$-experts one would expect that the aggregated distribution will go to some equilibrium. As a measure of the aggregated distribution the following figure 4.5 shows the mean and variance of the aggregated distribution as function of the number of added $\mathcal{N}(0,1)$-experts. Once the unnormalised weight of an expert is calculated it stays fixed. When adding a new expert only his unnormalised weight will be calculated and all normalised weights of course change. At first sight the mean and variance of the aggregated distribution stabilises after adding very few

Figure 4.6 shows the results of the same approach but instead of using experts with an underlying (0,1)-distribution, we now sample the mean of the underlying distribution again from a normal distribution. The more likely it is to produce a

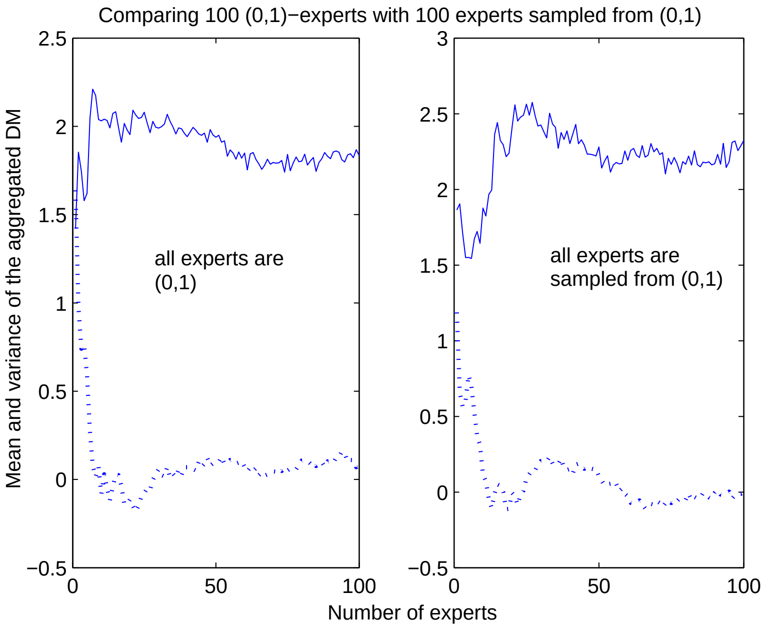

Figure 4.6: Comparison of the DM from (0,1)-experts and sampled (0,1)-experts

very bad expert, the slower is the convergence. Figure 4.7 shows the same results for using equal weights. It suggests that performance-based weights provide stronger convergence than equal weights.

This however is not a very satisfactory way of looking at the convergence of the Decision Maker. Due to statistical variation it is not so clear whether and from which point the curves of the mean and variance in figure 4.5 show some convergence.

4.4.1 Convergence in terms of Relative Information

The fact that the distribution of the Decision Maker is constructed from experts distribution using relative information and calibration a quantitative performance measures justifies out choice of the relative information between two successive Decision Makers as a measure of convergence. Each time after adding a new expert to the group of experts the distribution of the Decision Maker changes. If it does not change significantly we could say that the aggregated distribution of the Decision Maker converges. In that case the new expert does not add any new information to the Decision Maker's distribution.

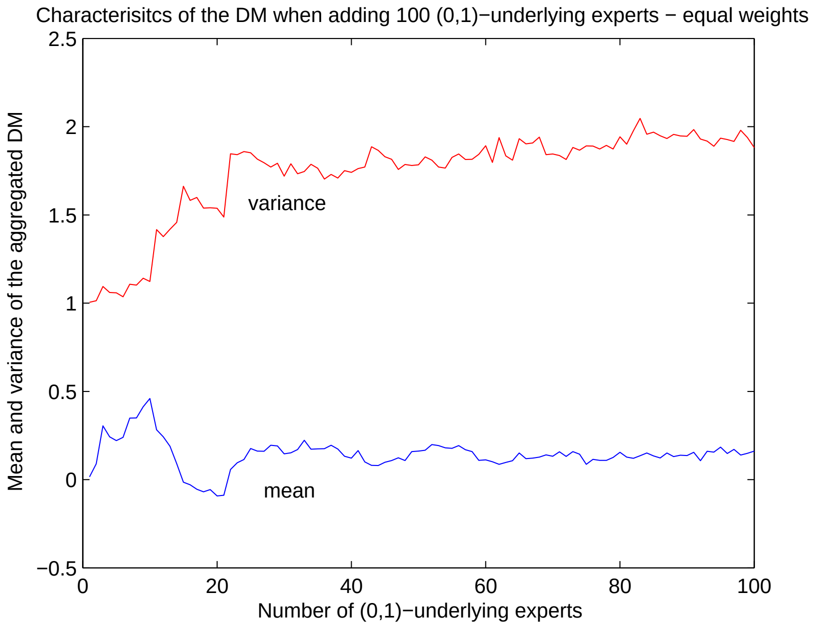

Figure 4.7: DM's characteristics as function of 100 (0,1)-underlying experts - equal weights

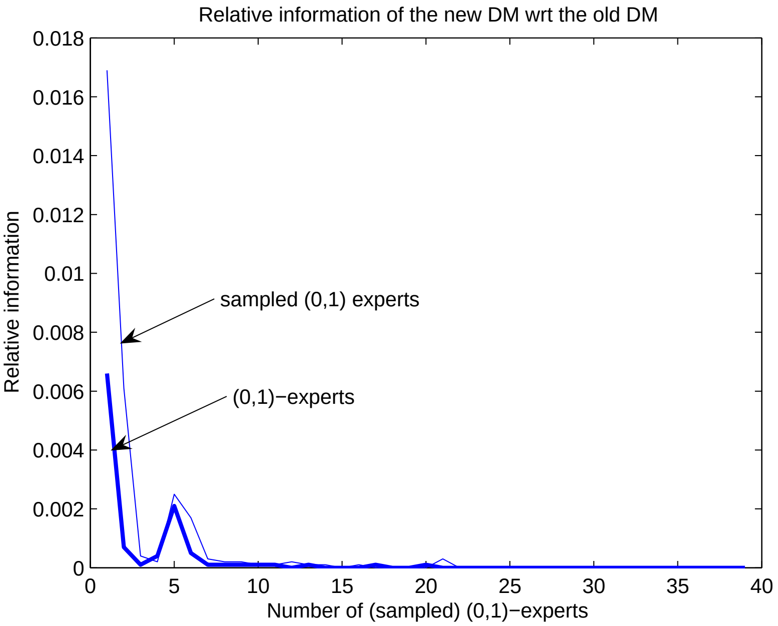

Figure 4.8: Convergence of the DM's distribution measured with the Relative Information

Twenty Experts Figure 4.8 shows again a comparison between $\mathcal{N}(0,1)$-experts and experts who's underlying mean is sampled from a $\mathcal{N}(0,1)$ distribution. This time however the measure of convergence of the Decision Maker's distribution is the relative information between each pair of two successive distributions of the Decision Maker. After adding approximately twenty experts, the relative information of the distribution of the new Decision Maker with respect to the distribution of the previous Decision Maker stays (approximately) zero. This suggests that in practice a group of twenty experts is sufficient to ensure convergence.

Note that we have added expert with similar characteristics; either the same or sampled from the same distribution. When using a broader range of characteristics it will be more likely that there is one or a few "best experts" that receive all the weight. In this case convergence should occur sooner that in the case a similar experts.

Equal Weights Again figure 4.9 shows the same results with equal weighting. Together with figure 4.7 it suggests that equal weighting has a much slower convergence rate than performance-based weights. This can easily be explained by the fact the with equal weighting even the really bad experts still receive

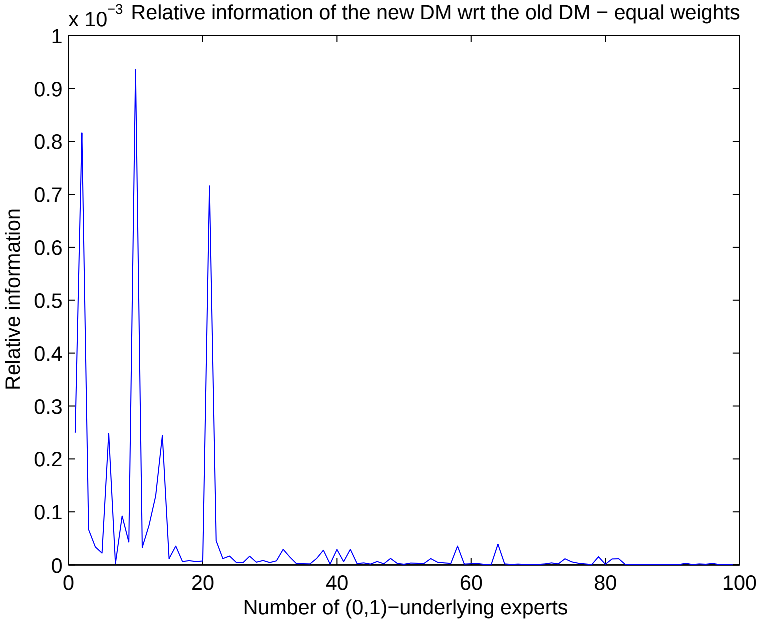

Figure 4.9: Relative information of new DM with respect to old DM when adding 100 (0,1)-underlying experts using equal weights

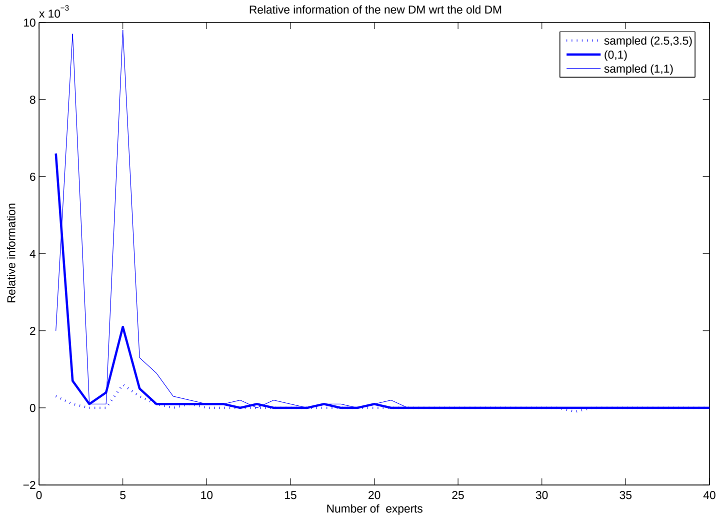

Figure 4.10: Convergence of the DM's distribution measured with the Relative Information

the same weight as the best expert. Adding new expert, regardless of their characteristics, will add new information to the Decision Maker and therefore the DM's distribution does not stabilise until the whole range of expert is added. With this we mean all experts from very good to very bad.

At last, figure 4.10 shows three completely different situations in one graph. Again convergence is measured by the relative information between two successive distributions of the Decision Maker. The dotted line represents a group of experts that have a highly biased underlying distribution. Their underlying mean is sampled from a $\mathcal{N}(2.5, 3.5)$ distribution and their underlying variance is one. It shows that the relative information between successive distributions of the Decision Maker is lower than in the other two cases where the experts are less biased. On the other hand adding the thirtieth-something expert still leads to a change in the distribution of the Decision Maker. Well at least in terms of relative information.

4.5 Conclusion

The results in this chapter suggest that performance-based weighting provides a quicker consensus on the Decision Maker's distribution than equal weighting. This is not surprising since performance-based weighting rises the situation where one or a few very good experts receive all the weight and the (many) bad ones receive zero weight. This is not possible when using equal weights.

Another result is that adding around twenty experts to the group is enough to ensure convergence. This results from both the DM's characteristics as the relative information between the pairs of consecutive DM distributions.

Chapter 5

Bayesian Belief Nets

5.1 Introduction

In many problems that deal with decision making and uncertainty, Bayesian Belief Nets (BBNs) are emerging. A Bayesian Belief Net is a graphical representation of a joint probability distribution and applications of BBNs range from medical, computer vision, financial to military domains. [7, 8] BBNs have proven to be a very satisfactory tool in all kinds of problems involving uncertainty. They bring back the number of quantities that have to be assessed and the conditional probabilities that specify the BBN in most cases have natural physical meaning to the Decision Maker. The definition of a Bayesian Belief Net [9, 10] is:

Definition 5.1 *A BBN is a pair (G, P) consisting of a directed acyclic graph $G = (V, E)$ and encodes a probability density P. The set V is the index of a set of variables $\{X_v\}_{v \in V}$. E is the set of directed arcs that connect the variables X_v and P is the probability density on $\{X_v\}$ specified by a set of conditional independence statements in the form of a acyclic directed graph and a set of probability functions. The joint probability function P is given by: $P(x_1, \ldots, x_v) = \prod_i P(x_i|\text{parents(i)}$*

The complete graph and conditional independence statements specify a joint probability over the variables $\{X_v\}_{v \in V}$ in the graph.

Each directed arc represents an influence from the node at the head of the arc which we will call the parent, to the child, the node at the tail of the arc. Also, each variable is fully specified by the variable itself and his parents in the graph. The conditional independence statements encoded in the graph allow us to calculate the joint probability as:

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|x_{parents(i)}) \tag{5.1}$$

where $f(x_i|x_{parent(i)}) = f(x_i)$ if $parents(i) = \emptyset$ [9, **?**]

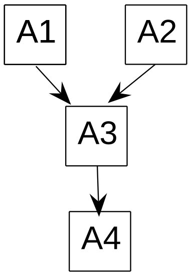

Figure 5.1: Example of Bayesian Belief Net.

We will distinguish between two kinds of BBNs, discrete and continuous. Discrete BBNs have discrete nodes and continuous BBNs have continuous nodes. The BBN with binary nodes in the following example is a special case of a discrete BBN. The difference in specifying the two different types of BBNs is in the joint distribution. A discrete BBN encodes conditional probability tables whilst a continuous BBN requires conditional probability functions and regression coefficients corresponding with the arcs in the graph.

5.2 Discrete BBNs

Table 5.1: Example of Conditional Probability Tables

$P(A_1)$		$P(A_2)$		$P(A_3\|A_1, A_2)$				$P(A_4\|A_3)$		
T	F	T	F	A_1	A_2	T	F	A_3	T	F
0.4	0.6	0.9	0.1	T	T	0.6	0.4	T	0.5	0.5
				T	F	0.2	0.8	F	0.5	0.5
				F	T	0.8	0.2			
				F	F	0.3	0.7			

Figure 5.1 shows a simple Bayesian Belief Net with corresponding conditional probability table 5.1. Together, figure 5.1 and table 5.1 specify a discrete BBN. The graph tells us that variables 1 and 2 are independent, variable 3 is conditioned on 1 and 2 and variable 4 is conditioned on 3. Table 5.1 shows the marginal distributions of variable 1 and 2 and the conditional distributions

of 3 and 4. Together these distributions specify the joint distribution over the variables. In case of our example the joint probability function is given by:

$$P(A_1, A_2, A_3, A_4) = P(A_1)P(A_2)P(A_3|A_1, A_1)P(A_4|A_3) \tag{5.2}$$

5.3 Continuous BBNs

When the nodes in the BBN represent continuous variables we speak of a continuous BBN. For these variables one needs to specify their conditional probability functions and a partial regression coefficient for each arc. A special case of continuous BBNs is the Gaussian Bayesian Belief Net. Here the joint distribution over the variables in the graph is normal.

5.3.1 Gaussian BBNs

We will call a Bayesian Belief Net Gaussian if the joint probability distribution of the variables in the BBN is multivariate normal. [19] Each variable has a conditional probability distribution, conditioned on its parents if the variable has parents. For joint normal variables constructing continuous BBNs is easier than discrete BBNs. Instead of specifying conditional probability tables we can interpret the influence among variables as partial regression coefficients when the child is regressed on the parents. [9, **?**, **?**]

In Gaussian BBNs, the conditional distribution for each variable is characterised by an unconditional mean, a conditional variance and partial regression coefficients. [9, 19] The definition of the conditional variance is given by:

Definition 5.2 *Conditional variance. If we are considering a conditional distribution $Y|X$, we define the conditional variance as*

$$var(Y|X) = E[Y - E(Y|X)]^2|X).$$

Say we have N variables $X_1, \dots, X_i, \dots, X_N$. Then the joint distribution for $X_1, \dots, X_N$ can be characterised by the unconditional means $\mu_i = EX_i$ and a covariance matrix:

$$\sum_{NN} = \mathrm{Var(X_N)} = \mathrm{EX_N X_N^T} - \mathrm{EX_N EX_N^T} \tag{5.3}$$

Also, in a Gaussian BBN each conditional variable $X_j|X_{parents(j)}$ is normally distributed with mean $\mu_j + \sum_{k \in parents(j)} b_{kj}(X_k - \mu_k)$ and variance ν_j where b_{kj} is the linear coefficient that represents the influence among variable k and j.

The Partial Regression coefficient

The linear coefficient b_{jk} can be written in terms of Yule's partial regression coefficient β in the following way: [19]

$$b_{kj} = \beta_{jk \cdot parents(j) \backslash (k)} \tag{5.4}$$

where β is defined by:

Definition 5.3 *Regression coefficient. Let us consider* X *and* Y *with covariance* $cov(X,Y)$ *and variances* σ_x^2 *and* σ_y^2*. Then the regression coefficients are given by*

$$\beta_{XY} = \frac{cov(X,Y)}{\sigma_x^2}, \qquad \beta_{YX} = \frac{cov(X,Y)}{\sigma_y^2}$$

In terms of linear predictors: Let X and Y be random variable with mean zero. Then b_{YX} minimises

$$E((X - b_{YX}Y)^2);$$

and b_{XY} minimises

$$E((X - b_{YX}Y)^2).$$

We are interested in the linear coefficients which are defined in terms of *partial* regression coefficients.

Definition 5.4 *Partial regression coefficient Let us consider variables* X_i *with mean zero,* $i = 1, \ldots, n$*. The numbers* $b_{12;3,\ldots,n}, \ldots, b_{1n;2,\ldots,n-1}$ *are the values that minimise*

$$E((X_1 - b_{12;3,\ldots,n}X_2 - \cdots - b_{1n;2,\ldots,n-1}X_n)^2).$$

Covariance Representation of Gaussian BBNs

The representation of a Gaussian BBN is closely related to the standard representation of the multivariate normal distribution in terms of the vector of means and correlation matrix. The key in this relation is the following theorem [9, 19]:

Theorem 5.5 *The covariance matrix* $\sum_{NN}$ *is positive (semi-) definite if and only if* $\nu_N > (\geq) 0$*. Further more, the rank of* $\sum_{NN}$ *is equal to the number of nonzero elements in* ν_N*.*

This theorem shows us how to verify whether a Gaussian BBN has a positive (semi-) definite covariance matrix and how to determine its rank. To fully specify a Gaussian BBN we need an unconditional mean and a conditional variance for each variable and a linear coefficient for each arc. The standard representation of multivariate normal distributions is in terms of unconditional statistics: a vector of means and a correlation matrix. Since the conditional variances in general have more meaning to expert than a correlation matrix, eliciting Gaussian BBNs is much more straightforward.

Once a Gaussian BBN has been assessed, the corresponding covariance matrix can be constructed. similarly, we can construct a Gaussian BBN from a covariance representation. In [19] the authors propose two algorithms to transform the BBN representation of a joint normal distribution into the covariance matrix representation and vice versa.

Chapter 6

BBN Expert Method

6.1 Introduction

In the past years, several methods are published to weight and combine expert judgements. Many of them describe the use of Bayesian techniques and seem very attractive and easy to work with. Mitchell Small et al [5] proposed the use of Bayesian Belief Nets in weighting and aggregating multiple expert judgements. Although, the method is mathematically attractive and easy to implement, there are a few theoretical comments. Applying their weighting method to existing data from Expert Judgement studies performed at the TU Delft shows that performance-based weighting of the Classical Model outperforms the underlying way of scoring of the Bayesian Belief Net method. This chapter presents the results of performance of the Bayesian Belief Net method and our theoretical evaluation of the method.

6.1.1 Outline

After explaining the method, there are three points on which the method is evaluated. First we show that the likelihood scoring rule is defined in such a way that experts can always maximise their score by telling the Decision Maker something other than their true belief. Then we look into the Decision Maker and show that although his distributions are the results of a Bayesian analysis, he himself is not Bayesian. Finally, we present a comparison of the likelihood score used on the BBN Method with a proper scoring rule on some real Expert Judgement data from studies performed at the TU Delft. It shows that the method's performance is not bad but can be improved.

6.2 Bayesian Belief Net Method [5]

The proposed BBN method builds upon available procedures for Bayesian model averaging and expert aggregation. Many researchers have developed Bayesian approaches for aggregating multiple expert models or expert opinions by using observed evidence to update the probability that the expert model or expert assessment is correct. These methods depend on the development of a likelihood function. This likelihood function is the probability that the observed evidence could have occurred given a particular expert model. In this section the method of weighting and combining experts using Bayesian Belief Nets will be explained. For a more detailed description we like the refer to [5].

6.2.1 Experts and BBNs

For each expert an individual Bayesian Belief Net is constructed by the Decision Maker. These BBNs differ among the experts only in the (conditional) probability tables or in case of continuous variables probability functions and partial regression coefficients. The child parents structure is the same for each expert and predetermined by the Decision Maker. This means that the Decision Maker has an important role as will be discussed further in section 6.4.

From now on the reader should interpret conditional probabilities as either continuous of discrete. The underlying directed acyclic graph $G = (V, E)$ is thus the same for each expert. Each expert provides the necessary prior and conditional probabilities and thus completely determines the joint probability density on his own BBN. This is his expert model. After eliciting each expert on the probability and dependence structure of his BBN, all expert models are fully specified.

Then, evidence about some of the variables, often called seed variables, is used to update each expert model. After observing evidence $\tilde{x} = [x_1, \ldots, x_K]$ the likelihood functions of the observation given the expert models can be computed. [5] The likelihood then determines the weights for each expert. The BBNs produce new forecasts that are assumed to follow the posterior beliefs of the respective expert conditioned on the evidence.

6.2.2 Combining

After updating each expert model with the observed evidence, the Decision Maker constructs a joint distribution over all variables (which are the model nodes) $A_1, A_2, \ldots, A_n$ as a weighted combination of individual expert opinions. This weight should be equal to the relative probability that each expert model is correct given the observed evidence. For an expert system with E experts, M_j denotes the model for expert $j, j = 1, \ldots, E$, the Decision Maker's distribution, the probability-weighted aggregate prediction of variable A is given by:

$$P(A) = \sum_{j=1}^{E} P(A|M_j)P(M_j) \tag{6.1}$$

Here $P(A|M_j)$ is the probability of event A given that expert model M_j is correct. $P(M_j)$ is the probability that expert model M_j is correct and $\sum_{j=1}^{E} P(M_j)$ is set to one. The prior information state is the state before any observations have been made and is designated by the subscript 0. In this prior state the Decision Maker does not have any extra information to begin with regarding the performance of the expert or the likelihood of their models. All expert models are then considered equally likely and have probability $P^0(M_j) = \frac{1}{E}$ here E is again the total number of experts.

6.2.3 Weights

The likelihood of the evidence $\tilde{x}$ given expert model M_j can be calculated for any number of pieces of evidence $\tilde{x}$. Let the evidence consist of: $\tilde{x} = (A_1 = a_1), \ldots, (A_K = a_K)$ where A_k denotes model node k and K is the total number of observed model nodes. Then, the likelihood of $\tilde{x}$ given expert model M_j is: [11]

$$\begin{aligned} &P(\tilde{x}|M_j) = P_j^0[(A_1 = a_1)]P_j[(A_2 = a_2)|(A_1 = a_1)] \\ &\times \textstyle\prod_{k=2}^{K} P_j[(A_k = a_k)|(A_1 = a_1) \cap \cdots \cap (A_{k-1} = a_{k-1})] \end{aligned} \tag{6.2}$$

Each expert j is weighted according to the probability that expert model M_j is correct. This probability is determined as the likelihood of the observations given the specific expert model.

This likelihood that each expert model is correct after observing $\tilde{x}$ is updated using Bayes Rule:

$$P(M_j|\tilde{x}) = \frac{P(\tilde{x}|M_j)P^0(M_j)}{\sum_{h=1}^{J} P(\tilde{x}|M_h)P^0(M_h)} \tag{6.3}$$

where $P(x|M_j)$ is the likelihood function for the probability that the evidence $\tilde{x}$ could have occurred given model j and $P^0(M_j)$ is the prior weight for expert j.

6.3 Scoring Rule

In this chapter, we will show that the likelihood score is not a (strictly) proper scoring rule and therefore the experts are able to maximise their score by not express their true belief. Recall that a scoring rule is (strictly) proper if an expert receives the maximal expectation of his score under his model if (and only if) his stated assessment corresponds to his true belief. This means that the optimal assessment an expert can give is his true belief and experts cannot cheat to receive higher weights. To show that the likelihood is not a strictly proper scoring rule, we will illustrate that with the likelihood as scoring rule, the expert can always maximise his expected weight by giving an opinion not equal to his true belief.

Before we will show that the likelihood score is (strictly) improper, first let us consider the expert score. With E experts and K variables the score for expert e_i is given by:

$$\text{score}(e_i) = \frac{P(x|M_i)P^0(M_i)}{\sum_{h=1}^{E} P(x|M_h)P^0(M_h)} \tag{6.4}$$

where x is some observed evidence and $P^0(M_i)$ is the prior weight for expert e_i given by the Decision Maker. Without any extra initial information about the experts' performance, the prior weights are set to $\frac{1}{E}$ and thus will fall out of equation 6.4. therefore the expert score can be rewritten as:

$$\text{score}(e_i) = \frac{P(x|M_i)}{\sum_{h=1}^{E} P(x|M_h)} \tag{6.5}$$

To show that the likelihood is not a (strictly) proper scoring rule we will first assume that an expert expresses a different belief than his true belief. Then, we show that by choosing this different belief smartly the expert can maximise his expected score. Which means we have shown that each expert can always maximise his score by not telling his true belief. Therefore the likelihood is (strictly) improper:

To optimise his score an expert should optimise the numerator of the right hand side expression of equation 6.6. Let $g(\underline{x})$ denote the expert's true belief about variable $\underline{x} = x_1, x_2, \ldots, x_n$ and let $f(\underline{x})$ denote the distribution he reports, trying to maximise his weight. Furthermore, assume that $f \neq g$ and his prior weight is $\frac{1}{E}$. After observing some evidence $\underline{\tilde{x}}$ this weight will change into the likelihood of the observed evidence given the expert's assessments. The expectation of this score is the expected likelihood before observing any evidence. The expected likelihood is therefore the sum over all possible pieces of observations $\underline{\tilde{x}}$ of the likelihood of the particular observation times the probability of this observation given the expert model. In the discrete case, the expectation of the likelihood is:

$$EL(\underline{\tilde{x}}) = \sum_{i=1}^{n} L(\tilde{x_i})P(L(\tilde{x_i})) \tag{6.6}$$

Since f is the distribution the expert reports, the likelihood of $\underline{\tilde{x}}$ given the expert model, is $f(\underline{\tilde{x}})$ and the probability of this observation $\tilde{x_i}$ given the expert model, is $g(\tilde{x_i})$ which is the expert's true belief is g. The expected likelihood therefore becomes:

$$E(L) = \sum_{\underline{x}} f(\underline{x})g(\underline{x}) \tag{6.7}$$

The expert gives the distribution f over $\underline{x}$, but his true belief is that it is g. In order to maximalise his weight, the expert should maximise his expected likelihood given by the above expression 6.3. The optimal solution is:

$$\arg\ \max_f E(L) = 1(\underline{x})_{\{\mathrm{argmax}_x g\}} g \tag{6.8}$$

In words: the distribution that gives ones to the vector $\underline{x}_{max}$ that maximises the expert's true belief, maximises the likelihood and thus the expected weight. This shows that using the likelihood as a scoring rule is not a strictly proper scoring rule and that an expert can always receive equal or higher weight by not telling his true belief!

6.4 Properties Decision Maker

Within the BBN method, the Decision Maker has an important role. He specifies the structure of each BBN and can influence the outcome by giving prior weights to the experts. In this section we will show that the Decision Maker is not Bayesian. Followed by some general remarks about the role of the Decision Maker in Bayesian practice. To illustrate that although the Decision Maker is the combination of experts updated with Bayes' theorem, he is not Bayesian himself I repeatedly make use of this theorem. Recall:

Theorem 6.1 *Let A and B_j be sets. Conditional probability requires that*

$$P(A \cap B_j) = P(A)P(B_j|A) = P(B_j)P(A|B_j)$$

Therefore,

$$P(B_j|A) = \frac{P(B_j)P(A|B_j)}{P(A)} \tag{6.9}$$

An extension to multiple distinct events $B_1, \cdots, B_n$ is:

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n} P(B_i)P(A|B_i)} \tag{6.10}$$

6.4.1 Non-Bayesian Decision Maker

The Decision maker is not Bayesian. With this is meant that updating the Decision Maker with new evidence directly will not results in the same distributions as updating the experts individually and combining them. In general the Decision Maker can also be seen as an expert. With this method however he does not have the same properties as the real experts since he is not Bayesian.

Say we have E experts who report $g_1(\underline{x}), g_2(\underline{x}), \ldots, g_E(\underline{x})$ and assume that all experts receive prior weight $\frac{1}{E}$. Let us consider expert e_i who reports $g_i(\underline{x})$ After observing the first variable x_0 the weight expert e_i receives follows from 6.2 (the prior weights cancel out.)

$$score(e_i|x_0) = \frac{g_i(x_0)}{\sum_{h=1}^{E} g_h(x_0)} \tag{6.11}$$

Substituting 6.11 for the probability $P(M_j)$ that expert models are correct into 6.1 gives the Decision Maker's distribution for the next variable x_1 given x_0 as the weighted combination of the expert distributions:

$$DM(x_1|x_0) = \sum_{j=1}^{E} \frac{g_j(x_0)}{\sum_{k=1}^{E} g_k(x_0)} g_j(x_1|x_0) \tag{6.12}$$

Analogously, after observing the second variable x_1 and recalculating the weights, the Decision Maker's distribution for x_2 given x_1, x_0 will be:

$$DM(x_2|x_1, x_0) = \sum_{j=1}^{E} \frac{g_j(x_0, x_1)}{\sum_{k=1}^{E} g_k(x_0, x_1)} g_j(x_2|x_1, x_0) \tag{6.13}$$

Looking at the Decision Maker directly however and updating his distribution with a new observed variable x_1 using Bayes's rule (theorem 6.1) gives the following expression for the Decision Maker's distribution for x_2 given x_1, x_0:

$$DM(x_2|x_1, x_0) = \frac{DM(x_2, x_1|x_0)}{DM(x_1|x_0)} \tag{6.14}$$

The Decision Maker is Bayesian if equation 6.13 and 6.14 are the same. In order to show that the Decision Maker is not Bayesian we can rewrite equation 6.14 using 6.12 to:

$$\begin{gathered} DM(x_2|x_1, x_0) = \\ \left(\sum_{j=1}^{E} \frac{g_j(x_0)}{\sum_{k=1}^{E} g_k(x_0)} g_j(x_1, x_2|x_0)\right) \left(\sum_{j=1}^{E} \frac{g_j(x_0)}{\sum_{k=1}^{E} g_k(x_0)} g_j(x_1|x_0)\right)^{-1} \\ = \frac{\sum_{j=1}^{E} g_j(x_0) g_j(x_1, x_2|x_0)}{\sum_{j=1}^{E} g_j(x_0) g_j(x_1|x_0)} \end{gathered} \tag{6.15}$$

The last expression in 6.15 is clearly not equal to 6.13. This means that although the Decision Maker obtains his aggregated distribution by Bayesian techniques, he himself is not Bayesian.

6.4.2 Role of the Analyst

As mentioned before, the role of the Decision Maker in the BBN Method is considerable. This has both positive and negative effects. On the one hand, the Decision Maker has a large role in determining the structure of the Bayesian Belief Net and thus the dependence structure of the variables. In reality this structure may be as unknown to the DM as the distributions of the variables itself. However this is also something to obtain using Expert Judgement. On the other hand, to give weights to the various experts a priori gives the DM the opportunity to reflect his initial information about experts in the model. This means that an experienced Decision Maker can add valuable information to the analysis. It can be concluded that in practice the Decision Maker has an significant and difficult role in the BBN method and that an experienced Decision Maker can influence the performance of the method in a positive way. The Decision Maker could also assume some correlation between the experts and can express that in his choice of prior weights.

6.5 Performance

Comparing the likelihood weights with global weights in Excalibur on existing Expert Judgement data from TU Delft gives insight in how the likelihood score performs compared to a proper scoring rule. Note that we are not comparing the two methods itself but only the way of scoring. Therefore we consider the following two scoring rules:

Scoring Rule 6.2 *The likelihood of the expert's assessments given the realisations of the seed variables.*

Scoring Rule 6.3 *The product of the expert's calibration and relative information, both calculated for and then averaged over the seed variables.*

Recall from Chapter 2 that the Calibration is the statistical likelihood of the hypothesis that the realisations of the seed variables are sampled independently from distributions agreeing with the expert's assessments. The Information is the average Relative Information of the expert's probability distributions with respect to a background measure over all seed variables.

To compare the two scoring rules we first need a way to quantify the likelihood of the realisations of the seed variables given the expert models. In case of the data from TU Delft Expert Judgement studies, the expert models are expert distributions given by quantifying three or five quantile points.

In two recent studies, the Dikering and AOT Risk study, the experts were asked to quantify the five [0.050.250.500.750.95] quantiles of their distribution for each variable. The likelihood of the realisation of a seed variable given an expert opinion is simply calculated as the size of the percentile in which the realisation of the seed variable falls. For N seed variables this leads to a value of the likelihood of all seed variables. Normalising the likelihoods of the seed variables for all experts gives a weight and the Decision Maker obtains a distribution for each variable as a weighted combination of all expert distributions and can be compared with a Decision Maker using another weighting scheme.

Note that in this way, we will not reproduce results from the Bayesian Belief Net method. The purpose is the compare the two scoring rules given by 6.2 and 6.3.

Example Say an expert assesses the following five values for the percentiles of his distribution for seed variable i: $[-2.5 \quad -0.5 \quad 3 \quad 4.25 \quad 6]$ and the realisation of this seed variable is zero. Then the realisation thus falls in between the 0.25 and 0.50 quantile of the expert's distribution and has a likelihood of

$$L_i = (0.50 - 0.25) = 0.25$$

For N seed variables, the likelihood of an expert distribution is now the product of the N likelihood values $\prod_{i=1}^{N} L_i$ of the seed variables. With Excalibur, a software package developed at the TU Delft, the weights calculated using the calibration and relative information or the likelihood can be compared. Following Excalibur and the Classical Model, the performance of the Decision Maker is quantified by his calibration and relative information.

6.5.1 Results

We calculate the likelihood of seed variables as the size of the bin of the expert distribution in which the seed variable falls. This means however that we cannot compare studies where experts quantified a different number of quantile points. Since the possible likelihood values decrease with increasing number of bins.

Example When expert are asked to quantify the 0.05, 0.50 and 0.95 quantiles, the values of the likelihood computed as described above are either: 0.05, 0.45, 0.45 or 0.05. When experts are asked to quantify five quantile points however, the values of the likelihood can be: 0.05, 0.2, 0.25, 0.25 0.2 and 0.05. In case of a seed variables falling into the bin to the right of the median of the expert distribution, the likelihood scores are respectively 0.25 and 0.45 for five and three quantiles. This will lead to an unfair comparison. Note: our estimation for the likelihood score do get more accurate however with increasing number of quantiles.

First we will give results of the comparison for data from two recent studies at the TU Delft where experts assessed five percentiles of their distribution for each item, the Dike Ring and AOT-studies. The results show that the Decision Maker based on the improper lieklihood scoring rule 6.2 does not perform too bad. His calibration and relative information are at the most about

Table 6.1: Dikering Data

Result Comparison Excalibur			
Weights	Calibration	Mean Rel. Information	Product
Global	0.3956	0.6462	0.25
Likelihood	0.1676	0.5721	0.1

twice as small. Results in Expert Judgement studies in general show much higher differences between the experts than between our two Decision Makers. Comparing the two scoring rules in the same way, but for studies with three percentiles gives more results. Here are the results of the ten studies that can be found in table 6.3.

Figure 6.1 shows a scatter plot of the calibration of the Decision Maker when using scoring rule 6.2 against 6.3. Although more than half of the points lay under the diagonal, the rest lays above. This shows that in general the proper Decision Maker is better calibrated, but the difference is not very convincing. Figure 6.2 shows the relative information of the Decision Maker for both scoring rules. The two scoring rules give an equal Decision Maker in terms of relative information for one study. All the other studies show a more informative Decision Maker for the strictly proper scoring rule.

Table 6.2: AOT Risk Data

Result Comparison Excalibur			
Weights	Calibration	Mean Rel. Information	Product
Global	0.827	1.212	1.0
Likelihood	0.4742	0.7426	0.35

Table 6.3: Considered Expert Judgement Studies

EJ Studies with 3 Q's	
Estec-1	DSM-1
Acnepts	Eunrca_s
Gas95	Infosec
Opriskbank	Mont1
RETURNafter	BSWAAL
DCPWWLW2	WATERPOL
DSM2	GROND5
ESTEC2	ESTEC3
CARMA-GREECE	NH3EXPTS
SO3EPTS	EUR-DD
EUR-WD	EUR-INT
EUR-EAR	EUR-SOI MVBLBARR

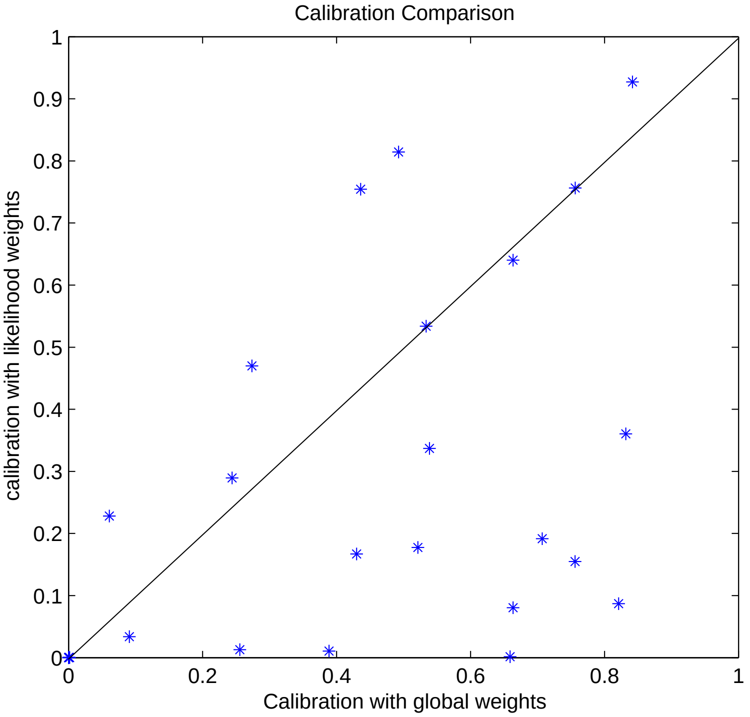

Figure 6.1: Scatterplot of calibration with global weights against likelihood weights.

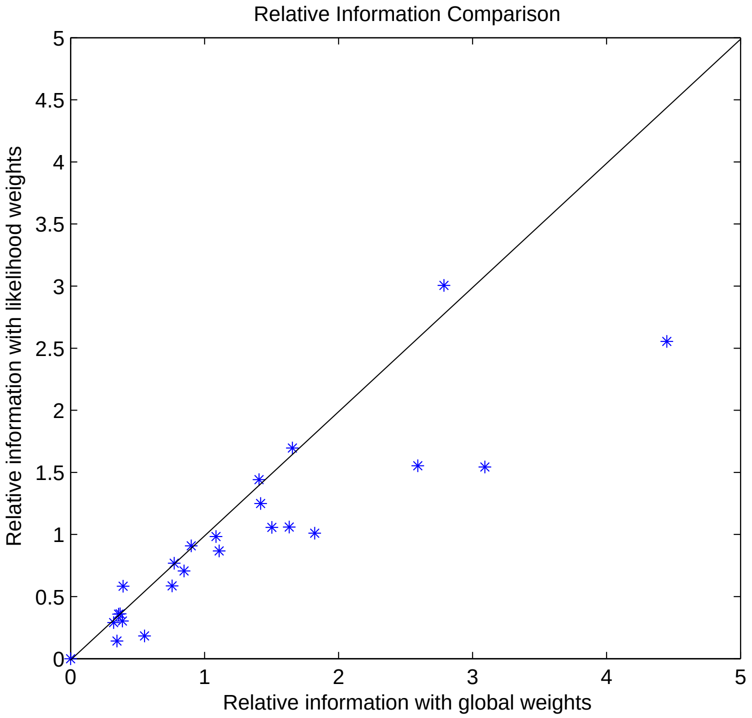

Figure 6.2: Scatterplot of relative information with global weights against likelihood weights.

6.5.2 Comparing with Equal Weights

From the results above could be concluded that performance-based weighting outperforms likelihood weighting. Note however that we have not compared the actual methods, but only the two underlying scoring rules. The actual BBN Method uses evidence on the seeds to update the expert models and is therefore expected to perform better than only the likelihood score which already does not perform too bad. Another scoring rule we can compare with likelihood weighting is equal weighting. This is the manner of weighting when there is no extra information on the experts' performance at all. In order for the BBN Method to be a useful method it should at least outperform equal weighting. If equal weighting appears to give better results there is no point in calculating the likelihood for each expert model and updating each BBN with evidence.

Table 6.4 compares the Calibration and Information of the equal weighting and likelihood weighting Decision Makers for the Expert Judgement data from the studies listed in table 6.3. In order to make the table more ledgible, the names of the studies that corresponds to the numbers $1, \ldots, 27$ are listed in appendix B. In more than half of the cases, the likelihood weight Decision Maker has a higher Correlation- a higher Information-score than the equal weight Decision Maker. The result is more convincing in the Relative Information graph. This indicates at least that the likelihood weights outperform equal weights on Information score whilst performance based global weighting outperformed likelihood weighting on Information.

Figures E.1 and 6.4 again illustrate that the likelihood Decision Maker is slightly better calibrated and more informative.

6.6 Conclusion

After illustrating that the likelihood is an improper scoring rule and the Decision Maker in the BBN method is not Bayesian himself the results using this improper scoring rule are not that bad at all. Especially considering the fact that the potential strength of the BBN Method, updating the expert models with evidence, is not even taken into account in our comparison. The likelihood score the BBN method uses, outperforms equal weighting and the likelihood Decision Maker in general has a calibration close to the performance-based global weight Decision Maker.

When comparing the the likelihood and performance-based Decision Makers one immediately notices that the calibration of both Decision Makers is not so different as their Information-scores. One explanation could be that, as Calibration, the likelihood measures how likely the observations are given the expert models. Likelihood does not tell us anything about the spread of the assessed

Table 6.4: Comparing equal with likelihood weights

	Calibration		Information		Product	
	likelihood	equal	likelihood	equal	likelihood	equal
1	0.75	0.43	1.06	1.42	0.80	0.61
2	0.08	0.53	0.91	0.81	0.07	0.43
3	0.29	0.28	1.54	1.51	0.45	0.42
6	0.76	0.56	0.87	0.30	0.66	0.17
7	0.81	0.12	1.06	0.72	0.86	0.09
8	0.19	0.71	1.70	1.01	0.33	0.72
9	0.17	0.34	0.59	0.32	0.10	0.11
10	0.64	0.53	1.01	0.75	0.65	0.40
11	0.09	0.01	0.71	0.17	0.06	0.00
12	0.53	0.64	0.18	0.29	0.10	0.18
13	0.00	0.00	0.00	0.14	0.00	0.00
14	0.00	0.36	1.25	1.10	0.00	0.39
15	0.93	0.50	3.01	0.62	2.79	0.31
16	0.15	0.06	3.01	2.90	0.47	0.17
17	0.00	0.97	0.36	0.15	0.00	0.14
18	0.47	0.13	0.98	0.53	0.46	0.07
19	0.18	0.18	0.36	0.29	0.06	0.05
20	0.23	0.29	1.55	0.97	0.35	0.28
21	0.34	0.15	2.55	2.09	0.86	0.31
22	0.00	0.00	0.14	0.17	0.00	0.00
23	0.01	0.00	0.58	0.65	0.01	0.00
24	0.36	0.11	0.77	0.56	0.28	0.06
25	0.01	0.07	0.30	0.16	0.00	0.01
26	0.00	0.00	0.29	0.97	0.00	0.00
27	0.03	0.22	1.44	0.55	0.05	0.12

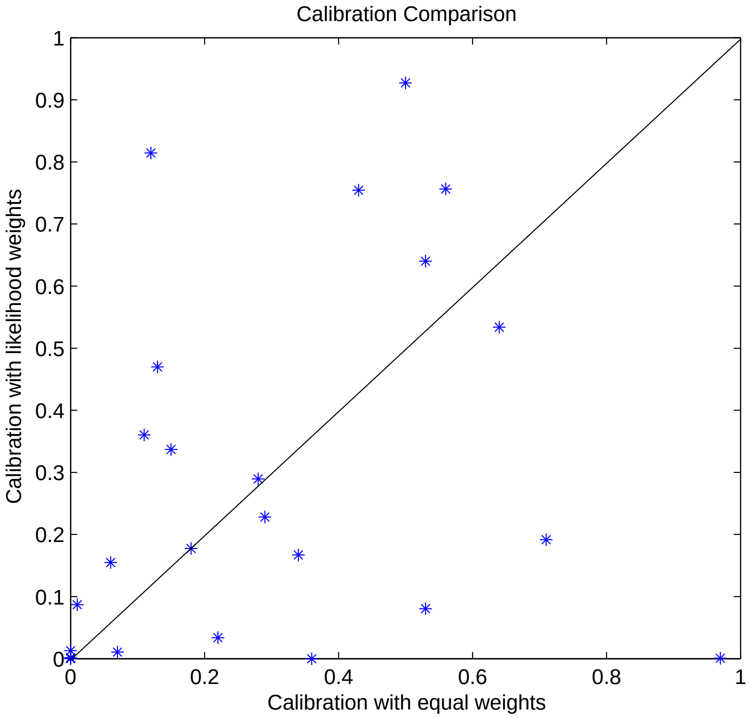

Figure 6.3: Scatterplot of calibration with equal weights against likelihood weights.

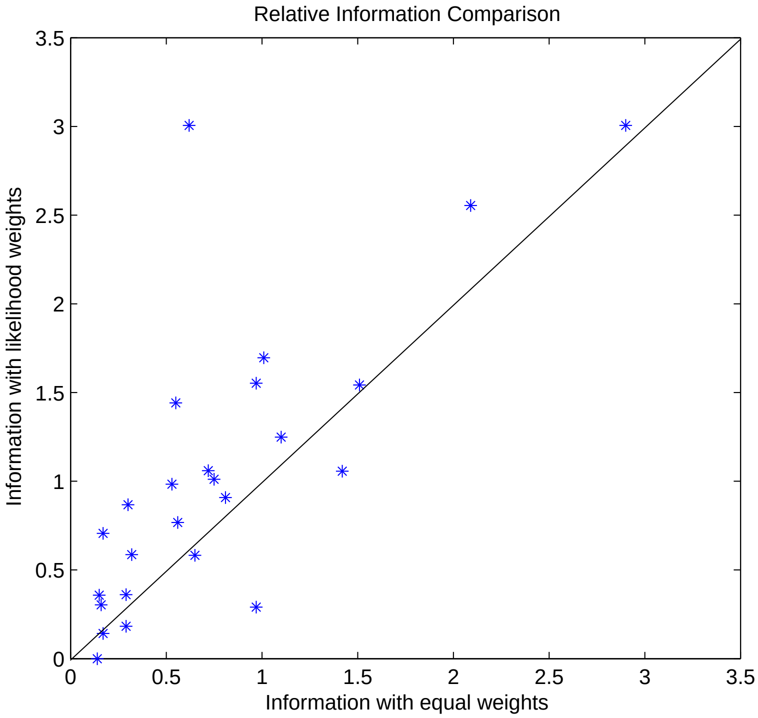

Figure 6.4: Scatterplot of relative information with equal weights against likelihood weights.

distributions. The Information of the likelihood Decision Maker is in general about twice as low as that of the performance-based Decision Maker.

6.6.1 Recommendations

The likelihood Decision Maker has a Calibration score not so far off that of the performance-based Decision Maker. The Information-score however is much better for the performance-based Decision Maker. One explanation that was offered is that the likelihood measure is similar to the Calibration in the sense that they both quantify the likelihood of the seed variables given the expert models. The likelihood however does not provide anything on the spread of the expert models. One way to improve the BBN-method therefore could be to include some measure that quantifies this spread, like the Relative Information.

Using the likelihood to weight the experts means using an improper scoring rule. However it is not clear whether this is a problem in real-life Expert Judgement practising. By improving the BBN method it can therefore become a very useful Expert Judgement tool.

Comparing with Social Networks

Huang and Cooke [13] have evaluating the performance of a Social Network Decision Maker. They used data from the TU Delft joined EU USNRC uncertainty analysis of accident consequence for nuclear power plants. This data includes ten expert panels and have resulted in performance based combinations of expert judgements. Recall from 2 that Social Networks theory views social relationships in terms of nodes and ties and focuses on relationships among social entities. Experts where weighted within these Social Networks by taking into account the number of scientific publications, experience and recommendations of a wide class of experts. In their study the authors use seven expert panels. Using the seven expert panels from EU USNRC data to compare the likelihood weight with Social networks could be a nice extension. However, the likelihood weights seem to outperform the equal weight whilst the SN Method does not seem perform better than equal weighting.

Chapter 7

Conclusions and Recommendations

This thesis has been about two different aspects of Expert Judgement.

I In the first part, we have a model to produce synthetic expert data. The most interesting results lead to some remarks and conclusions on the effect of Correlation and on Convergence of the Classical Model.

II The second part dealt with a new Expert Judgement method proposed by Small et al [5], the BBN Method. Although there are some thearetical comments to be made, evaluating the method on existing EJ data showed that the method does perform quite well.

7.1 Correlation

There are many possible forms of correlation that can have effect on the performance of an EJ method.

7.1.1 Dependence between Experts

In the first part of this thesis some illustrating results have been presented from applying the Classical Model onto synthetic expert data. The most interesting results concern the correlation effects. It is very well plausible that there exists some correlation between experts. An example where half of the experts are academics and the other half work in the field makes this clear. The academics might be correlated among each other as well as the other half. Correlation lowers the number of independent samples from the multinomial sample distribution and therefore delays the convergence to a Chi-square distribution. This means that the hypothesis test that determines the Calibration score can become invalid. My recommendation is that whenever the analyst suspects a high correlation between some (or all) experts he should include more seed variables to ensure Chi-square convergence and a valid hypothesis test. Also one should think about a way to correct the effective number of samples N when there is a high correlation between the expert distributions.

7.1.2 Dependence between Seeds

The Classical Model is based on the assumption that there is a large dependency between the seed variables and the unknown variables of interest. It is therefore also plausible to assume that the seed variables are not independent.

In Chapter 3, Modelling Experts, however we have not assumed any dependence between the seed variables or with the target variable. It is however interesting the research the effect of (un-)correlated seed variables on the performance of the Classical Model. Even more interesting would be to look at the correlation or dependence between the seed variables and target variables. One could for instance expect that when there is no dependence between the seed variables and the target variables, the Classical Model looses some of its value. Once again the EJ database from TU Delft could be very useful. With all this data at hand, one can just calculate the correlation or dependece between the seed variables and target variables.

7.2 Expert Characteristics

The main results from section 4.2 followed our expectations and thus our model to produce synthetic expert distributions provided us a useful model to evaluate for instance the effects of correlation and convergence. One fundamental question however remains. How does the way we model expert characteristics compare with real experts? In order to make the results presented in this thesis more relevant, one has to think of a way to quantify the expert characteristics for real experts.

Within our model to produce synthetic expert distributions, we used a fixed underlying bias. Suggested further research could investigate the situation where the underlying bias is not constant. What would be the influence of the seed variables in the case where experts have a much lower underlying bias for the seed variables as for the unknown variables?

The two questions posed above are closely related to the effect of the correlation. If there is no correlation between the seed variables and the target variable, it is more plausible to assume different underlying biases for the seed and target variables. On the other hand, the correlation between experts is the one characteristic that an analyst could estimate from a group of experts as in the example of academic and field specialists.

7.3 Convergence

Using the Bin-approach in Chapter 4 showed that the common number of around 10 experts might not always be enough to ensure convergence. Comparing the theoretical results with real data from Meng [15] confirms this.

One way to come to our conclusions on convergence of the Classical Model was to compare successive Decision Maker when adding experts. We compared the Decision Makers using their characteristics and the Relative Information. Using Relative Information however could be too strict of a comparison. Initial results on comparing the successive Decision Makers with respect to the seed variables show that we need even more experts to find a stable Decision Maker. These results can be found in Appendix **??**.

7.4 Evaluating the BBN Method

Except the theoretical disadvantages, the BBN Method seems to be a promising method. Considering the fact that the potential strength of the BBN Method, updating the expert models with evidence, is not even taken into account in our comparison it performs well. If it is possible to implement a measure of spread of the expert models it could provide a useful alternative for existing EJ methods.

7.5 Remarks

7.5.1 The Use of Numbers

Throughout this thesis we have used specified number of seed variables, experts and quantiles in our simulations and examples. The choice for using ten seed variables in most simulations is given by the fact that this is a common number of seed variables in practice. Most Expert Judgement studies at the TU Delft have used three quantiles. In this thesis however we have described the use of five quantiles. This is done in order to reduce the effect of the seed variables already a little bit. Using even more quantiles - which is easily done in modelling synthetic expert distributions - would not give justice to the real life practice of Expert Judgement. Eliciting expert on many quantiles will be too time consuming and moreover too hard for the experts. The number of experts has been one of the parameters that we have varied in order to infer something on convergence. In other simulations however, we used ten experts. This is again an amount given by the everyday practice of Expert Judgement.

7.5.2 Normalised vs. Unnormalised Weight

In Chapter 3 we have constructed the probability density function of the unnormalised weight of an expert depending on his underlying characteristics. The reason to look at the unnormalised weight as opposed to the normalised weight is that the unnormalised weight reflects only the characteristics of the expert under consideration. The normalised weight however will also take into account other experts and expresses more than the relation between the expert characteristics and his weight. It also includes the performance of the expert relative to the performance of other experts.

Bibliography

[1] Combining Expert Judgements: The Classical and Copula Method. James K. Hammit and Joshua T. Cohen. HCRA. Joint Satistical Mettings: Section on Physical and Engineering Sciences (SPES.)

[2] Assessing and Using Dependence in Expert Judgement Studies. Maarten-Jan Kallen. Technical University of Delft, January 19 2003.

[3] Using the Elliptical Copula to Combine Expert Assessments with a Full Dependence Structure. M.J. Kallen and R.M. Cooke. Technical University of Delft, June 2002

[4] Modelling errors with normal distribution.

[5] Site-Specific Updating and Aggregating of Bayesian Belief Network Models for Multiple Experts. Neil A. Stiber, Mitchell J. Small, and Marina Pantazidou. Society for Risk Analysis 2004.

[6] Experts in Uncertainty: Opinion and Subjective Probability in Science Roger M. Cooke. Oxford University Press, 1991

[7] Real World Applications of Bayesian Networks. Communications of the ACM, special issue 38(3), 1995

[8] Approximation bayesian Belief Networks by Arc Removal. Robert A. van Engelen. IEEE TRansations on Pattern Analysis and Machine Intelligence, Vol 19, No. 8, August 1997.

[9] Uncertainty Analysis with High Dimensional Dependence Modelling. Dorota Kurowicka, Roger M. Cooke. Delft Institute of Applied Mathematics, Delft University of Technology 2005.

[10] Advances in Bayesian Networks Series: Studies in Fuzziness and Soft Computing, Vol. 146 Gmez, Jos A.; Moral, Serafin; Salmern, Antonio (Eds.) Springer 2004.

[11] Sequential Updating of Conditional Probabilities on Directed Graphical Structures. D.J. Spiegelhalter and S. Lauritzen. Networks 20, 579-605, 1990.

[12] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. Bayesian Data Analysis. Chapman and Hall/CRC, Boca Raton, 1995.

[13] Review of Expert Judgement methods and Application of Social Network Theory. X. Huang and R.M. Cooke. Delft Institute of Applied Mathematics, Delft University of Technology 2005.

[14] Combining Probability Distributions from Experts in Risk Analysis. R.T. Clemen and R.L. Winkler. Risk Analysis 2, Vol 19, 187-203, 1999.

[15] Performance Based Expert Aggregation, Chunfang Meng, technical University Delft, 2005.

[16] http://mathworld.wolfram.com/MultinomialDistribution.html

[17] The Assessment of Probability Distributions from Expert Opinions with an Application to Seismic Fragility Curves. A. Mosleh and G. Apostolakis. Risk Analysis, Vol6. No. 4, 447-461, 1986.

[18] Models for the Use of Expert Opinions.A. Mosleh and G. Apostolakis. presented at the workshop on low-probability high-consequence risk analysis, Society for Risk Analysis, Arlington VA, June 1982.

[19] Gaussian Influence Diagrams. Ross D. Shachter and C. Robert Kenley. Management Science Vol 35, No. 5, May 1989.

[20] The Role of Expert Judgement in Hazardous Factors Influence Prognosis: Parametric Elicitation Technique. Victor G. Krymskt, Roger M. Cooke and Andrey R. Yunusov. Proceeding 9th Anual Conference Risk Analysis: Facing the New Millenium, 1999.

Appendix A

Theoretical explanation Backgrounds

Let f, g and h be the probability functions of respectively an expert distribution, the normal background and the uniform background. Then the relative information of the expert distribution and both backgrounds can be computed by:

$$\begin{aligned} RI(f,g) &= \int_l^u \tfrac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(\tfrac{-(x-\mu_1)^2}{2\sigma_1^2}\right) \left(\ln(\tfrac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(\tfrac{-(x-\mu_1)^2}{2\sigma_1^2}\right)) - \ln(\tfrac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(\tfrac{-(x-\mu_2)^2}{2\sigma_2^2}\right))\right) \mathrm{dx} \\ &= \int_l^u \tfrac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(\tfrac{-(x-\mu_1)^2}{2\sigma_1^2}\right) \left(\ln(\tfrac{1}{\sqrt{2\pi\sigma_1^2}}) - \tfrac{(x-\mu_1)^2}{2\sigma_1^2} - \ln(\tfrac{1}{\sqrt{2\pi\sigma_2^2}}) + \tfrac{(x-\mu_2)^2}{2\sigma_2^2}\right) \mathrm{dx} \end{aligned} \tag{A.1}$$

and

$$\begin{aligned} RI(f,h) &= \int_l^u \tfrac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(\tfrac{-(x-\mu_1)^2}{2\sigma_1^2}\right) \left(\ln(\tfrac{1}{\sqrt{2\pi\sigma_1^2}}) - \tfrac{(x-\mu_1)^2}{2\sigma_1^2} - \ln(\tfrac{1}{u-l})\right) \mathrm{dx} \\ &= \int_l^u \tfrac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(\tfrac{-(x-\mu_1)^2}{2\sigma_1^2}\right) \left(\ln(\tfrac{1}{\sqrt{2\pi\sigma_1^2}}) - \tfrac{(x-\mu_1)^2}{2\sigma_1^2} + \ln(u-l)\right) \mathrm{dx} \end{aligned} \tag{A.2}$$

Subtracting the second equation from equation A.1 gives us:

$$\int_l^u \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(\frac{-(x-\mu_1)^2}{2\sigma_1^2}\right) \left(-\ln(\frac{1}{\sqrt{2\pi\sigma_2^2}}) + \frac{(x-\mu_2)^2}{(0.5(u-l))^2} - \ln(u-l)\right) \mathrm{dx} \tag{A.3}$$

Substituting the parameters of the normal background distribution

$$\begin{aligned} \sigma_2^2 &= (0.25(u-l))^2 \\ \mu_2 &= 0.5(u+l) \end{aligned} \tag{A.4}$$

into equation A.3 gives:

$$\int_{l}^{u} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(\frac{-(x-\mu_1)^2}{2\sigma_1^2}\right)\left(-\ln(\frac{1}{\sqrt{2\pi(0.25(u-l))^2}}) + \frac{(x-0.5(u+l))^2}{(0.5(u-l))^2} - \ln(u-l)\right) \mathrm{dx} \tag{A.5}$$

All we need to show now is that this last equation is bigger than zero.

Appendix B

Names EJ Studies

Table B.1: Comparing equal with likelihood weights

1	estec1	2	dsm1
3	acne	4	aotrisk
5	dikering	6	aseed
7	gas95	8	infosec
9	opriskbank	10	mont
11	returnafter	12	bswaal
13	dcpwwlw2	14	waterpolo
15	dsm2	16	grond5
17	estec2	18	estec3
19	carma-greece	20	nh3expts
21	so3expts	22	eurdd
23	eurwd	24	eurint
25	eurear	26	eursoi
27	mvblbarr		

Appendix C

Expert Judgement studies by the TU Delft

The following table lists all Expert Judgement studies performed by the Delft Institute of Applied Mathematics at the Delft University of Technology. It gives both the scores of the performance-based and equal-weight Decision Maker as of the best expert. It total 21942 elicitations were made in the pas years, resulting in an extensive database of Expert Judgement data.

Appendix D

Bayesian Belief Net Software

The last table gives an overview of available software for updating Bayesian Belief Nets coming from http://www.cs.ubc.ca/ murphyk/Software/BNT/bnsoft.html

What do the headers in the table mean?

- Src = source code included? (N=no) If so, what language?
- API = application program interface included? (N means the program cannot be integrated into your code, i.e., it must be run as a standalone executable.)
- Exec = Executable runs on W = Windows (95/98/NT), U = Unix, M = Mac, or - = any machine with a compiler.
- Cts = are continuous (latent) nodes supported? G = (conditionally) Gaussians nodes supported analytically, Cs = continuous nodes supported by sampling, Cd = continuous nodes supported by discretization, Cx = continuous nodes supported by some unspecified method, D = only discrete nodes supported.
- GUI = Graphical User Interface included?
- Learns parameters?
- Learns structure? CI = means uses conditional independency tests
-
- Utility = utility and decision nodes (i.e., influence diagrams) supported?

- Free? 0 = free (although possibly only for academic use). \$ = commercial software (although most have free versions which are restricted in various ways, e.g., the model size is limited, or models cannot be saved, or there is no API.)

- Undir? What kind of graphs are supported? U = only undirected graphs, D = only directed graphs, UD = both undirected and directed, CG = chain graphs (mixed directed/undirected).

- Inference = which inference algorithm is used? jtree = junction tree, varelim = variable (bucket) elimination, MH = Metropols Hastings, G = Gibbs sampling, IS = importance sampling, sampling = some other Monte Carlo method, polytree = Pearl's algorithm restricted to a graph with no cycles, none = no inference supported (hence the program is only designed for structure learning from completely observed data)

- Comments. If in "quotes", I am quoting the authors at their request.

Appendix E

Performance DM's wrt. the Seeds

The following figure shows the Calibration and Relative Information of the Decision Maker when adding expert based on only one seed variable. Since it is only based on one seed, this is a very preliminary result.

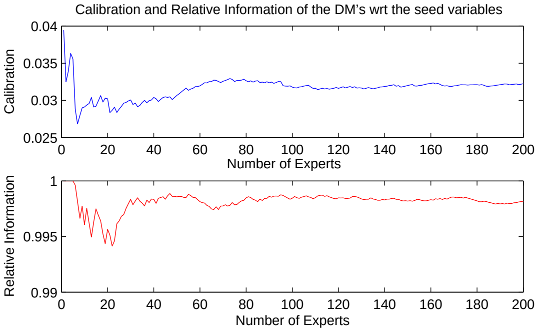

Figure E.1: Performance of the DM with respect to the seed variables