

Performance Based Experts Aggregation

Chunfang Meng

September 23, 2005

Contents

1	Introduction and Summary	3
1.1	Summary of Conclusions	3
1.2	Recommendations	4
1.3	Outline of the Report	4
2	Classical Model	4
3	Experts' distributions in PEGASOS	4
4	Scoring rules	6
4.1	Calibration (quantile tests)	6
4.2	Calibration (Discrete test)	7
4.3	Global Calibration Score	7
4.4	Information	8
4.4.1	Information of Quantile Test	8
4.4.2	Information of Discrete Test	8
4.5	Performance weight decision maker	10
4.6	Noisy calibration variables	10
4.6.1	Instrument Noise Example	10
4.6.2	Kraan 1, [5, Kraan](3.6)	11
4.6.3	Kraan 2 [5, Kraan](3.9)	12
4.6.4	Another Approach by the Author	12
4.6.5	Scoring with sample distributions	13
5	Elicitations of Subproject 1	14
5.1	Seismic Source Definition	14
6	Experts' Combinations and Calibration Subjects	15
6.1	Data Set Description	15
7	Aggregation Based on Synthetic Seed Variables	16
7.1	Groen's Model and Dependence Structure	18
7.2	Seed Variable Sampling	19
7.2.1	Choice of <i>rec</i> and <i>freq</i> Values	19
7.3	Performance Based Decision Maker	20
8	Case Study	21
8.1	Multiple Seismic Events	21
8.2	Dependence Assumption Effect	27
8.2.1	Unify the Sampling Randomness	27
8.2.2	Impact of the Joint Normal Dependence	29

8.2.3	Impact of the Vine Dependence	29
8.2.4	Comparison between Joint Normal and Vine	29
8.3	Noise Effect	31
8.3.1	Kraan 1	31
8.3.2	The Author's Approach	32
8.4	Summary of the Matlab Code Functions	36
9	DM Convergence	36
9.1	Measure of Convergence	36
9.2	DM w.r.t. Number of Seed Variables	37
9.3	DM w.r.t. Number of Seed Experts	39
9.3.1	Fix the Intrinsic Range	39
10	Conclusion	43
10.1	80 Experts' Performances	43
10.2	Different Decision Makers	43
10.3	Impact of the Optimistic Dependence Assumptions and Noise term	43
10.4	Performance DM Convergence	43
11	Appendix A	44
11.1	Figure (26 to 31) are rang graphs corresponding to Figure (15)	44
11.2	Table (21) is the Excalibur counterpart of Table (4) from Excalibur	44
12	Appendix B	51

1 Introduction and Summary

This study is concerned with Probabilistic Seismic Hazard Analysis for Swiss Nuclear Power Plant Sites (PEGASOS Project), [4, Pegasos, 2004]. In this project experts' opinions and equal weight decision makers were used. The idea of the post hoc calibration is to make assessment on experts' performance based on updated observations and, if possible, propose non-equal weight decision maker.

Post-hoc calibration involves finding observations of quantities for which probabilistic predictions exist. This can be done at different 'aspiration levels', depending on the type of calibration variables which can be found:

- 1 Find observations of quantities predicted probabilistically by individual experts,
- 2 Find observations that can be predicted by panels, that is aggregations of experts in a given area,
- 3 Find observations which are predicted probabilistically by the combination of several or all expert panels.
- 4 Develop synthetic seed variables from plausible physical models.

The first possibility requires the most extensive set of calibration variables, and would allow us to construct performance based combinations of the experts in each panel. The second possibility would not allow to re-combine the expert assessments, but would enable us to compare the performance of different expert panels. The third possibility involves the least extensive set of calibration variables, and enables us only to validate the performance of the PEGASOS decision maker.

The following remarks relate to the feasibility of implementation of these aspiration levels

- The calibration variables may be either direct observations of variables for which probabilistic predictions exist. Alternatively, they may also be observations of variables which can be *predicted*, perhaps with the help of supplementary modelling assumptions, from the PEGASOS model.
- Ideally the calibration variables are 'deterministic' in the sense that their values are known with certainty. However, it is also possible to calibrate experts with 'uncertain observations' as long as the uncertainty in the observation is known, is independent of the experts' uncertainty and is small relative to the experts' uncertainty. [5, bernds thesis].
- The probabilistic predictions may be either in the form of quantiles or discrete probabilities. The scoring rules used in cases are conceptually identical but differ in implementation.
- At the fourth and lowest level of calibration we can generate plausible sets of seed variables and use these to determine
 - i whether there are significant differences in the experts' performance relative to these values
 - ii whether the equal weight combination of experts yields reasonable and representative performance
 - iii how many seed variables and/or how many experts are necessary to produce stable results

1.1 Summary of Conclusions

The report assesses the feasibility of post-hoc calibration at each of the above three aspiration levels. Real calibration variables could in principle be applied in seismic risk studies. For example,

recent measurements in California could have been used in this context. However, without re-interviewing the experts, we conclude that it is not possible to calibrate the PEGASOS experts with actual measurements, noisy or otherwise.

Aspiration level [4] is feasible. The model of Groen [Groen, 2004] enables calibration of combinations of experts with plausible synthetic data, and thus allows us to answer the three questions 1..3 above. More precisely, we have 4, 5, and 4 experts / expert panels in SP1, SP2, SP3. This yields 80 different combinations, and variables predicted by each of these 80 combinations may be generated from Groen's model and used as synthetic calibration data. Moreover, we can study the effect of using a "typical" number of experts (5), versus using as much as 80 experts.

Groen's model can be used either with or without dependencies in the input variables. The dependence structure is elaborated in [Meng, 2005].

The main conclusions of this study are

- 1 There are very significant differences between the experts' performance. The performance based decision maker (DM) yields good performance.
- 2 The equal weight decision maker returns poorer performance both in calibration and information, comparable with most expert judgement studies.
- 3 The decision maker stabilizes with about 10 experts and 15 seed variables.

1.2 Recommendations

Based on the results with synthetic data, the following recommendations suggest themselves:

- 1 since (i) PEGESOS involves potentially 80 different expert combinations for computing seismic hazard curves, since (ii) the performance of these combinations is quite different, and since (iii) the equal weight combination yields inferior performance, it would be advantageous to obtain real calibration variables.
- 2 Real calibration variables would require re-interviewing the experts.

1.3 Outline of the Report

The features of PEGESOS and basic notions of structured expert judgment are presented in sections 2 - 4. Sections 5, 6, 7 and 78 describe the data available for the post-hoc calibration study. Section 9 examines the question of convergence. Section 10 gathers conclusions. Two appendices provide additional figures and tables.

2 Classical Model

Here we review classical model, [1, Roger Cooke, 1991], as a non-equal weight combination of experts' opinions. The experts' performance is determined by their calibration and informativeness scores. The model treats experts' assessments as statistical hypotheses and scores calibration as the p-value at which we would reject these hypotheses, under the assumption that the calibration variables are independent realizations of the experts' distributions. The informativeness of the experts are defined as the relative information of the experts' distributions with respect to user specified background measure. The theory of strictly proper scoring rules is used to combine calibration and information scores, and thus to form the so-called performance based decision maker. Some details of this model are provided in Section (4).

3 Experts' distributions in PEGASOS

From now on we refer expert panels(teams) as experts. The experts' elicitations were done according to the level 4 study of Probabilistic Seismic Hazard Analysis (PSHA) recommended by

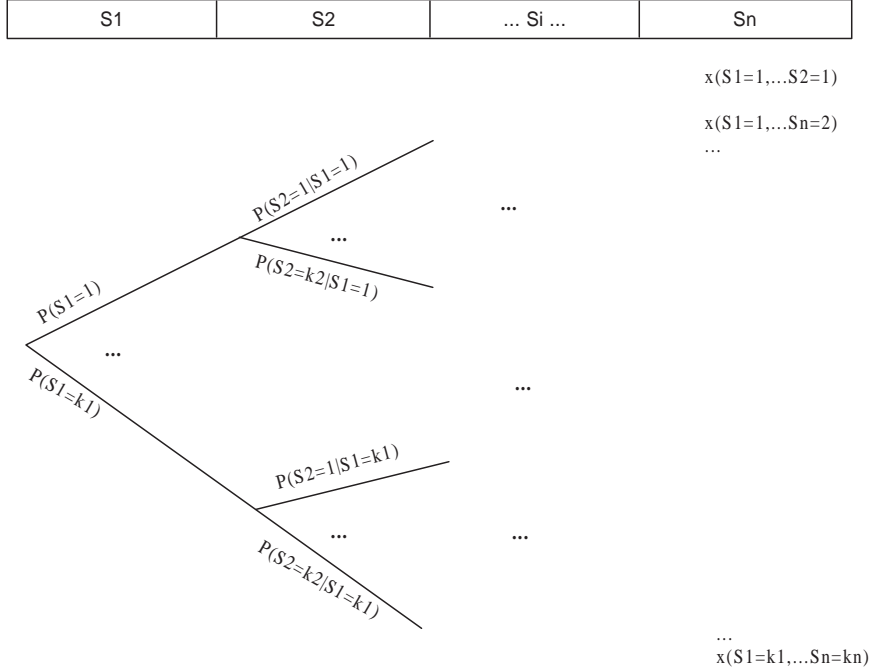


Figure 1: n -level logic tree

Senior Seismic Hazard Analysis Committee (SSHAC), [6, R.J.Budnitz, 1997]. During the elicitation experts sometimes gave their subjective (discrete or continuous) distribution over unknown quantities. While in many cases the experts put their probabilities (sometimes conditional probabilities or split fractions) over the branches of a logic tree structure. Suppose for example that experts are asked to assess a unknown quantity X which is at the n^{th} level of a logic tree. At the i^{th} level or *node* of the tree there are k_i branches emerging from each $i - 1$ level branch. In Figure (1) S_i denotes the i^{th} node of the tree. $S_i = j$ means that the j -th branch is chosen at the i -th node. The expert's mass function over the values of x is just the probability of the path leading to x , where x takes values $\{1, 2, \dots, \prod_{i=1}^n k_i\}$:

$$\begin{aligned}
 P(X = 1) &= P(S_1 = 1)P(S_2 = 1|S_1 = 1)\dots \\
 &\quad P(S_n = 1|S_1 = 1, \dots, S_{n-1} = 1) \\
 &\quad \dots \\
 P(X = \prod_{i=1}^n k_i) &= P(S_1 = k_1)P(S_2 = k_2|S_1 = k_1)\dots \\
 &\quad P(S_n = k_n|S_1 = k_1, \dots, S_{n-1} = k_{n-1}) \\
 \\
 P(X = 1) &= P_1; \\
 &\quad \dots = \dots \\
 P(X = \prod_{i=1}^n k_i) &= P_N
 \end{aligned}$$

A discrete distribution example of X is shown in Fig 2. The elicitation results from this

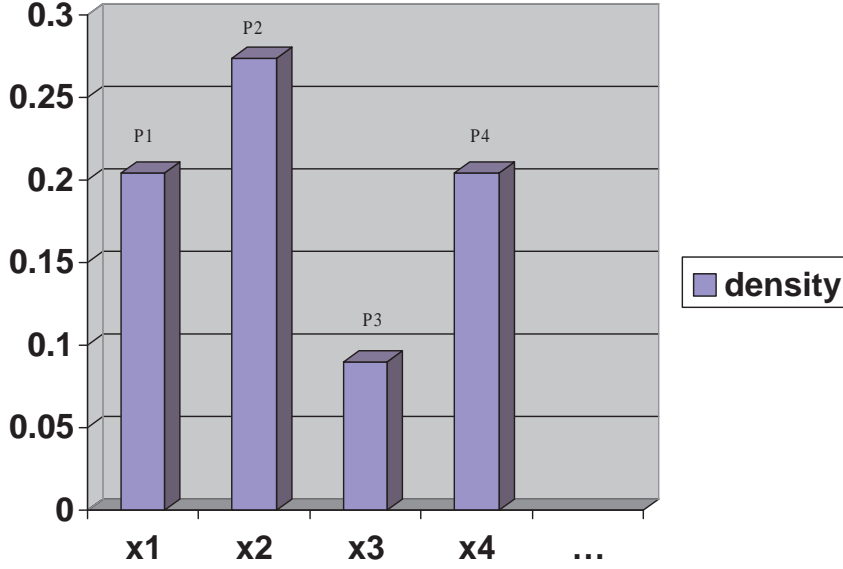


Figure 2: A density example of X

logic tree can be seen as discrete distributions assessed by experts. From the density of X the $\{q_i\}, i = 1, \dots, k$, quantiles are calculated as:

$$Q(q_i) = X_k \mid \sum_{j=1}^k P_j \leq q_i < \sum_{j=1}^{k+1} P_j; \quad (1)$$

If the elicitation results are continuous distribution of an unknown X , the corresponding quantiles are:

$$Q(q_i) = F_X^{-1}(q_i);$$

Where $F_X^{-1}(\cdot)$ is the inverse cumulative distribution function of X .

In this project, the expert distributions are discrete. The weights assigned to the different branches of the logic tree represent a discrete distribution which in some cases is understood as an approximation of a continuous distribution

The quantile intervals $[\min\{x_i\}, Q(q_1)], \dots, [Q(q_k), \max\{x_i\}]$ are used to calculate calibration and information scores.

4 Scoring rules

4.1 Calibration (quantile tests)

Suppose there are N elicitation quantities which are observed post hoc with certainty. An expert has given a distribution, continuous or discrete, for each of them. The calibration and information scores are calculated as [1, Roger Cooke].

Recall the quantile intervals $[\min\{x_i\}, Q(q_1)], \dots, [Q(q_k), \max\{x_i\}]$ in last section. Let the number of observations which fell into the $[0, q_1]$ quantile interval be s_1N , let the number of observations which fell into $(q_1, q_2]$ quantile interval be s_2N , etc. The empirical mass function is $s = (s_1, \dots, s_k + 1)$, and the hypothetical mass function is $p = (q_1, q_2 - q_1, \dots, q_k - q_{k-1}, 1 - q_k)$. Relative

information measures the 'degree of surprise' we should have if we believed p and subsequently learned s .

$$I(s; p) = \sum_1^{k+1} s_i \ln\left(\frac{s_i}{p_i}\right) \quad (2)$$

This value is nonnegative and takes minimum 0 if and only if $s = p$. If the hypothesis density and the empirical density are close, the relative information should be close to 0 which means the expert's performance is good. For large N the distribution of $2N$ times of the relative information is approximately χ^2 distribution with k degrees of freedom, when q is generated by independently sampling p .

$$P(2NI(s; p) \leq x) \approx \chi_k^2(x); \quad (3)$$

where χ_k^2 is the distribution function of the χ^2 variable with k degrees of freedom. The calibration of expert e is defined as the probability of getting a relative information greater than or equal to that actually obtained under the expert's hypothesis that the true distribution is (p_1, \dots, p_{k+1}) . It can be interpreted as the likelihood of the hypothesis:

$$C(e) = 1 - \chi_k^2(2NI(s; p)) \quad (4)$$

Obviously an empirical distribution s equal to the hypothetical distribution p gives the maximal calibration 1.

In most of the cases, we choose $q = (0.05, 0.5, 0.95)$ then $p = (0.05, 0.45, 0.45, 0.05)$ and $k = 3$. Note, however, than any other choice of q could also be used.

The experts might not have the same number of post hoc observations. Since in PEGASOS they don't have to assess the same things. The number of realizations determines the *power* of the statistical test applied to the expert-hypotheses. If the number of realizations is very large, small differences between the experts probabilities p and the sample distribution q will be detected as statistically significant. If the number of realizations is small, then these small differences may not be detected as statistically significant. In comparing experts who have assess different numbers of calibration variables, we therefore scale the "effective number" of calibration variables for each expert to be equal to the smallest number of calibration variables assessed by some expert. This is accomplished by adjusting the number N in equation (3). For comparisons across studies it is common to use 10 effective calibration variables.

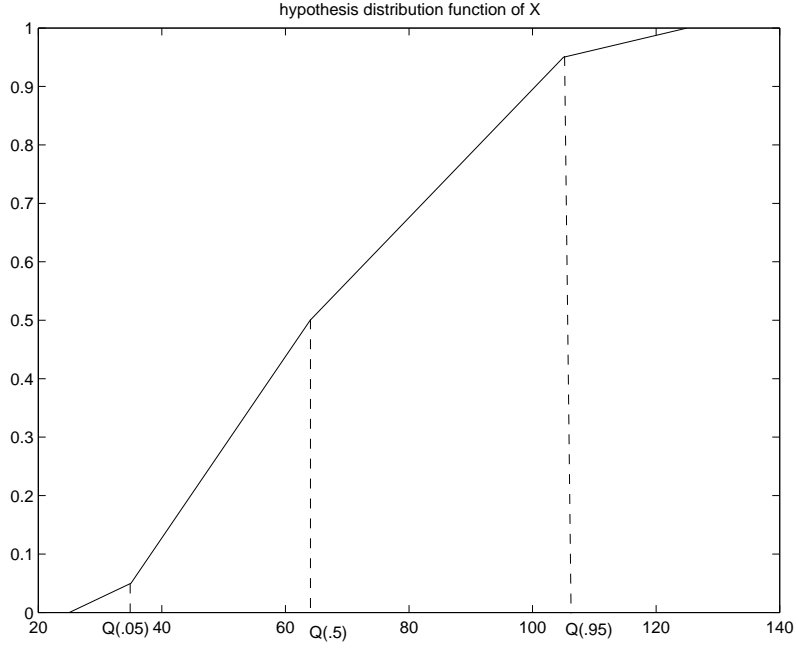
4.2 Calibration (Discrete test)

In the PEGASOS project some elicitation probabilities are over a small number of possible values, say R . In such cases the experts do not assess quantiles of an uncertain quantity with a continuum of possible values, rather they assess a probability vector $p = (p_1, \dots, p_R)$ over the R possible outcomes. If we have several realizations corresponding to the same assessment vector p , then we can score these as before using the asymptotic χ^2 distribution. However, we will not generally have multiple realizations. In this case we can still apply the scoring rule theory, but we cannot appeal to the asymptotic distribution. Instead, we must simulate the distribution of an appropriate scoring variable. For details see [1, Cooke 91] Chapter 8, 10.

Quantile tests are preferred over discrete tests [1, Cooke, 1991] since the former require fewer realizations. Moreover, the scaling required to equalize the effective number of calibration variables is more difficult for discrete tests.

4.3 Global Calibration Score

For each expert (team) we have a χ^2 variable $2N'I(s, p)$ from the quantile test with 3 degrees of freedom, recall 4.1. Suppose we have a χ^2 scoring variable R_k from the discrete variables with k degrees of freedom. Under the hypothesis that the experts give accurate probability assessments, with independent realizations, we may treat these two χ^2 scoring variables as independent. Hence their sum will be again be χ^2 with the degrees of freedom equaling the sum of the degrees of freedom of the summands.

Figure 3: cumulative distribution of X

4.4 Information

4.4.1 Information of Quantile Test

Another criterion to measure the quality of an expert opinion is informativeness. To measure it a background measurement, usually uniform or log-uniform, is assigned to each observed variable over an intrinsic range. The intrinsic range is determined by adding a $k\%$ margin to the smallest interval containing all quantiles and observation, where k is selected with default value 10. After choosing the intrinsic range for unknown X , we substitute the interval $[\min\{x_i\}, \max\{x_i\}]$ by the intrinsic range, say $[l, u]$, in definition of back ground measurement. When the back ground measurement is uniform, the distributions within the quantile intervals are uniform.

Plot $[\min\{x_i\}, Q(0.05), Q(0.5), Q(0.95), \max\{x_i\}]$ versus $[0, 0.05, 0.5, 0.95, 1]$. The expert's interpolated cumulative distribution function and interpolated density functions of X is shown in Fig (3) and Fig (4). This distribution is proven to be minimal information distribution w.r.t uniform back ground measure shown as Fig (5), [5, Kranss].

Where $r_i, i = 1, \dots, 4$ are back ground measures of the corresponding intervals. The relative information of expert e is given as [1, Cooke 1991]

$$I(e) = \sum_1^4 p_i \ln\left(\frac{p_i}{r_i}\right), \quad (5)$$

where $p = (0.05, 0.45, 0.45, 0.05)$ is the expert's probability. Overall informativeness for each expert is the average of the information scores over all observed variables. This average is proportional to the relative information in the expert's joint distribution over all variables under the assumption that the variables are independent. Information scores are always positive. When calibrations are equal, experts with high information scores are preferred.

4.4.2 Information of Discrete Test

Scoring informativeness for discrete tests is different than for quantile tests. We do not appeal to an intrinsic range. Instead, for assessed mass function $p = (p_1, \dots, p_R)$ we use the relative

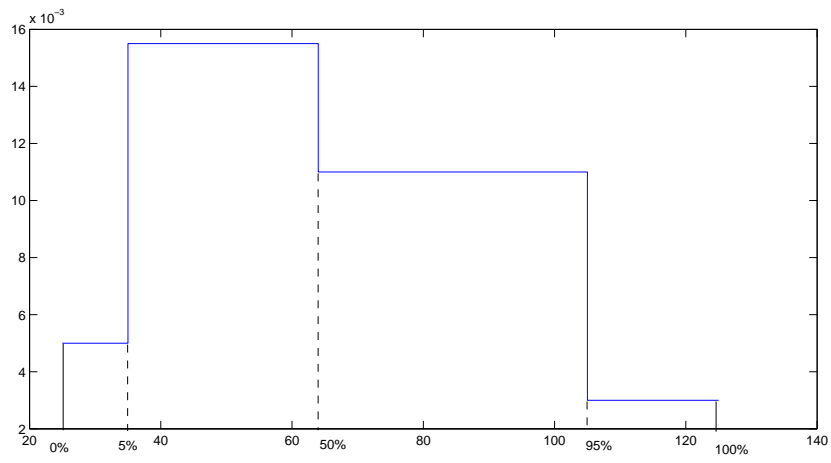


Figure 4: hypothesis distribution density of X

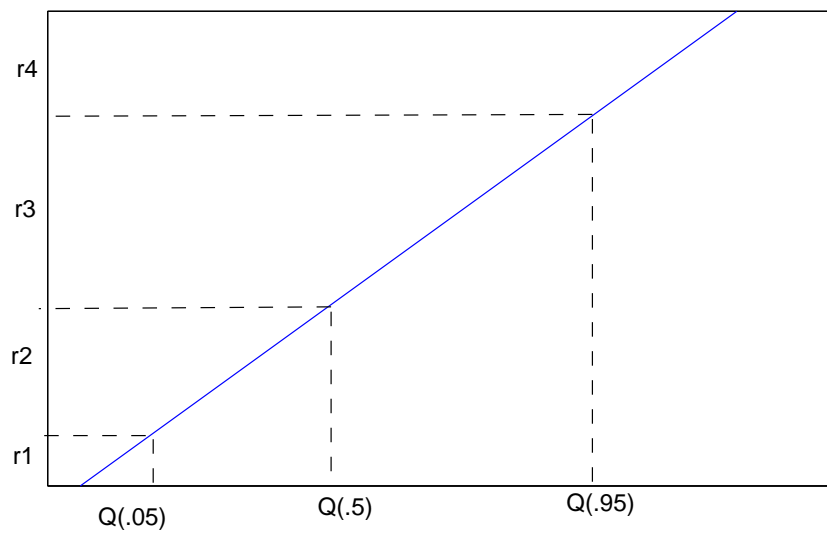


Figure 5: uniform back ground distribution of X

information of p with respect to the uniform distribution $u = (\frac{1}{R}, \dots, \frac{1}{R})$.

For each calibration variable we have an information score. The global information score for the expert e is defined as the average of the information scores over all variables :

$$Inf(e) = \frac{\sum_X I_X(e)}{\#X} \quad (6)$$

4.5 Performance weight decision maker

The performance weight of an expert is determined by combining the calibration and information scores. We usually make weight of each expert proportional with the product of these two scores. Usually, calibration is very sensitive over experts and information is a relatively slow function. The product of these two scores is dominated by calibration. The information score is to differentiate almost equally well calibrated experts.

To make the scoring rules weakly asymptotic proper, see [1, 9.4 Cooke], we set a cut-off level α such that experts with calibrations lower than α are given 0 weights. Then the unnormalized weight of the expert e is, recall $GC(e)$ and $Inf(e)$ in 4.3 and 4.4.2,

$$w(e) = GC(e)Inf(e)1_{\{GC(e) \geq \alpha\}} \quad (7)$$

After normalization we have a global weight for each expert

$$\tilde{w}(e) = \frac{w(e)}{\sum_1^{\#expert} w(i)}. \quad (8)$$

To determine the weights we have to choose the cut-off level α . The decision maker's distribution density is weighted sum over experts' distribution density. The decision maker depends on the cut-off level. The way to choose α is to treat the decision maker as another expert and calculate its unnormalized weight. The optimal α is chosen to maximize the decision maker's weight

$$\alpha = argmax_{\alpha}(\tilde{w}(DM)). \quad (9)$$

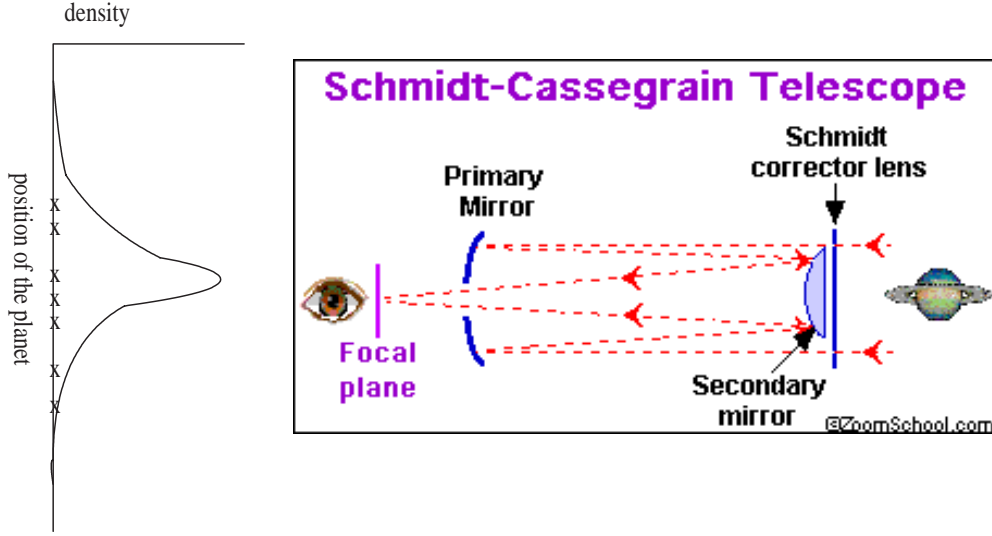
4.6 Noisy calibration variables

In the foregoing, the observations were assumed to be certain. The method for dealing with uncertain or noisy observations of calibration variables is described in [5, Kraan, 2002].

In some cases the repeatability of the observation results is not high. The variation of the observations could be due to the noise of instruments.

4.6.1 Instrument Noise Example

Suppose we observe the position of a planet using a telescope.



Due to instrument noise, we observe some scattered points on the screen. Denote the positions of the points as a sample set $Z^* = \{z_i^*\}$, where $i = 1 \dots n$. Experts are asked to give their uncertainty about the true position of the planet in term of quantile values without taking the instrument noise into consideration. Denote the elicitation variable by Y .

Let's change the elicitation subject. Suppose we do not ask the uncertainty of the planet position directly. We ask what we would observe using this type of telescope. Denote this elicitation variable by Z . Z can be modeled by:

$$Z = Y + \epsilon; \quad (10)$$

where the noise term ϵ is independent of Y . Usually ϵ is assumed to be normally distributed. The relation between distribution densities f_Z and f_Y is given by the a convolution equation:

$$f_Z = f_Y * f_\epsilon. \quad (11)$$

In [5, Kraan], two approaches are made to approximate f_Z for given quantile elicitation of Y .

4.6.2 Kraan 1, [5, Kraan](3.6)

First approximate the minimal relative information density function. It is given by the solution of a Convex Programming (CP) problem.

$$\min_{f_Y} \sum_{l=1}^{N_Y} f_Y(y_l) \ln \frac{f_Y(y_l)}{f_U(y_l)} \quad (12)$$

$$\text{s.t. } Af_Y = b \quad (13)$$

Where

- matrix $A \in \mathbf{R}^{(K+1) \times N_Y}$,
- f_U is background measure of Y .
- N_Y is the number of discretized points of Y ;
- The entry of A , $a_{s,t} = \mathbf{1}_{I_s}(y_t)$;
- K is the number of quantiles elicited from experts;
- I_s is the s -th quantile interval of Y ;
- b is ranges of the quantile intervals, say $[0.05, 0.45, 0.45, 0.05]$.

Note. the constraint of this CP problem (15) is the discretized form of

$$\int_{q_s}^{q_t} f_Y(y)dy = t - s. \quad (14)$$

It is proven that if the background measure f_U is uniform, the solution f_Y is piece wise uniform over each quantile interval.

Convolute the density f_y and given f_ϵ , then we have f_Z .

4.6.3 Kraan 2 [5, Kraan](3.9)

We may approximate the minimal information distribution f_Z w.r.t. the background measure. It can be formulated by another CP problem

$$\min_{f_Y} \sum_{l=1}^{N_Y} f_Y(y_l) \ln \frac{f_Y(y_l)}{f_U(y_l)} \quad (15)$$

$$\text{s.t.} \quad Af_Y = b \quad (16)$$

$$\text{s.t.} \quad f_Z(z_m) = \sum_{l=1}^{N_Y} f_Y(y_l) f_\epsilon(z_m - y_l) 1_{A_\epsilon}(z_m - y_l); \quad (17)$$

where

- N_Z is the number of discretized points of Z ;
- A_ϵ is the support range of ϵ .

The *constraint* of this CP problem is the discretized form of (11).

4.6.4 Another Approach by the Author

From f_Z produced by Kraan, the quantile points of Z is calculated. We change the calibration subject form Y to Z . The seed variables are from the noisy observation, recall the set Z^* . Let the calibration of Z given Z^* be $Cal(Z|Z^*)$. The noise term ϵ is implicitly contained in Z .

In this study, we can try a different method. The aim is to keep Y unchanged and calibrate it for given observation, denoted by $Cal(Y|Z^*)$.

First interpolate sample distribution. For ordered given n observation samples $Z^* = (z_1^*, \dots, z_n^*)$, the quantiles can be calculated as

$$q_k = Z_i \quad \text{s.t.} \quad i/n \leq k < (i+1)/n.$$

The observation distribution density f_{Z^*} can be interpolated as piece wise uniform within each quantile interval, $I_k = [q_k, q_{k-1})$.

Comment: It looks like contradiction. If we knew the distribution of elicitation subjects in first place what do we need experts for?

The fact is we may frequently observe events with high recurrence rates /year. They are mostly earthquakes can hardly cause any damage. The purpose of using experts is to predict the damage caused by large earthquakes. For such earthquakes the observations are very rare in Switzerland.

We do not need experts to predict anything with high recurrence. However, we do need something observable to calibrate experts and weight them based on their performance. The seed variables will probably come from some low magnitude events with high yearly recurrence rates.

In many cases there is only one observation point for each uncertain variable. The sample distribution can hardly be approximated. In such cases, we can treat samples as deterministic. If the noise of the those observations are assessed independently we may turn to Kraan 1 or 2. For some cases we have a distribution rather than scattered points through an observation. The next section is to handle such observations.

Now, try to filter the noise out of the observations. Denoting the observation variable without noise as Y^* , the Equation (11) holds as:

$$Z^* = f_{Y^*} * f_\epsilon$$

Approximating the distribution of Y_j^* is another CP problem given by:

$$\min_{f_{Y^*}} \sum_{l=1}^{N_Y} f_{Y^*}(y_l) \ln \frac{f_{Y^*}(y_l)}{f_U(y_l)} \quad (18)$$

$$\text{s.t.} \quad f_{Z^*}(z_m^*) = \sum_{l=1}^{N_Y} f_{Y^*}(y_l) f_\epsilon(z_m^* - y_l) 1_{A_\epsilon}(z_m^* - y_l); \quad (19)$$

where

- N_Y is the number of discretized points of Y^* ;
- f_U is the back ground measure of Y^* ;
- f_ϵ is the PDF of ϵ ;
- $m = 1, \dots, N_Z$, where N_Z is the number of observed samples;
- A_ϵ is the support range of ϵ .

The constraint of this CP problem can be formulated as:

$$A f_{Y^*} = b; \quad (20)$$

where $A \in \mathbf{R}^{N_z \times N_y}$ with (s, t) entry :

$$a_{s,t} = f_\epsilon(z_s^* - y_t^*) 1_{A_\epsilon}(z_s^* - y_t^*).$$

The s -th component vector b is

$$b_s = f_{Z^*}(z_s^*).$$

Choose $N_z \leq N_y$ such that the constraint will not be over-determined.

From f_{Y^*} and set Z^* , The CDF F_{Y^*} and F_{Z^*} are derived. The noise-free set $Y^* = \{y_i^*\}$ can be derived as:

$$y_i^* = F_{Y^*}^{-1}(F_{Z^*}(z_i^*)) \quad (21)$$

Note, if the noise is small, then F_{Y^*} and F_{Z^*} are almost identical thus $y_i^* \approx z_i^*$. The set Y^* is a function of Z^* , written as $Y^*(Z^*)$. Use them as the noise-free seed variables to calibrate the elicitations of Y for given noisy observations Z^* .

$$\text{Cal}(Y|Z^*) = \text{Cal}(Y|Y^*(Z^*)). \quad (22)$$

4.6.5 Scoring with sample distributions

There are cases such that for each seed variable j we have n observed distribution $F_{Y_j^*}$. Let

$$y_j(q_j) = F_{Y_j^*}^{-1}(q_j), \quad (23)$$

where $q = \{q_j\} \sim U[0, 1]^N$ are *rank* values of Y_j^* ; N is the total number of uncertain seed variables. Let

$$s_k(q) := \frac{\sum_j 1_{I_{j,k}}(y_j(q_j))}{N};$$

where $I_{j,k}$ is the k -th quantile interval of the j -th elicitation. The relative information of the observation w.r.t elicitation is a function of q :

$$RI(S(q), P) = \sum_{i=1}^K s_i(q) \ln\left(\frac{s_i(q)}{p_i}\right); \quad (24)$$

where $K = 4$ and $\{p_i\} = (0.05, 0.45, 0.45, 0.95)$.

The calibration score is

$$C(S(q), p) = 1 - \chi_k^2(2NI(s(q); p)). \quad (25)$$

The expectation of the calibration score is

$$EC(S, p) := \int_q C(S(q), p) dF(q); \quad (26)$$

Where $dF(q) = dq = dq_1 \dots dq_N$. Sample q from a joint uniform distribution $U[0, 1]^N$. For each sample q , (24) can be calculated. Substitute $dq = 1/R$, where R is the number of samples of q . The (26) becomes average value over all the samples.

$$EC(S, p) = \frac{\sum_q C(S(q), p)}{R}. \quad (27)$$

5 Elicitations of Subproject 1

The Expert elicitions in PEGASOS happened during three PEGASOS subprojects, source characteristic (SP1), ground motion characteristic (SP2) and site response (SP3). Experts worked as teams (panels) to give their (conditional) probabilities over unknown quantities as shown in Figure(1)

The SP1 Expert elicitions are documented as [4, PEGASOS final report volume 4]. There are 4 experts teams (panels) consisting of several experts, *EG1a*, *EG1b*, *EG1c* and *EG1d*. Each of the teams carried out 4 tasks

- Seismotectonic Framework,
- Seismic Source Definition,
- Maximum Earthquake Magnitude,
- Earthquake Recurrence Parameters.

The seismotectonic framework is to set up a common state of knowledge within each team. Based on the the framework the teams carried out the next 3 tasks.

Among the next three tasks, Seismic Source Definition offers the best hope for calibration variables. There are possibilities that those unknown quantities can be observed with good accuracy. The latest effort made by Berkeley Seismological Laboratory (BSL) shows that right after an earthquake happened, the source geometry can be determined very fast and precisely. A standardized routine was described at <http://seismo.berkeley.edu>.

5.1 Seismic Source Definition

Many elicitations were done to give probabilities of unknown seismic source parameters. Here we list some of those whose values may be observable post hoc.

For the 4 expert teams the potential calibration (seed) variables could be:

- Fault orientation [4, Final Report vol4]:

- Tab. 2: Style of Faulting and Rupture Orientation for EG1a Sources.
- Tab. A-1: Earthquake Rupture Parameters for EG1b.
- Tab. A1-2: Rupture orientation, style of faulting, and source boundary conditions for EG1c.
- Tab. A-2: Styles of faulting for EG1d source. zones.
- Earthquake depth distribution for sources,[4, Final Report vol4]
 - Fig 15 17 from EG1a;
 - Fig 28, Fig 29 from EG1b
 - Fig. A1-7,Tab. 6: Depth distributions used, by tectonic region from EG1c
 - Tab. A-3: Focal depth distributions

Elicitation results are in a different format with different notation. Experts had consensus on the existence of seismic sources/faults within each team while the inter-team differences were significant.

The results from BSL cannot be used to calibrate experts in PEGASOS directly since the events did not happen in Switzerland. If there is an event Switzerland, the event-fault geometry can be determined precisely.

We could also perform bootstrap calibration if there are sources in regions similar to Switzerland:

- 1 First go through the elicitations and find some common sources shared by all teams.
- 2 Give each source *deterministic* parameters as our imaginary observations.
- 3 Using the scoring rules, compute the performance based decision maker.
- 4 Once *real* observations are available, we can use them instead of the imaginary observations.

The fact is that earthquakes do not happen often in Switzerland. If there is simply no event observed, then hopefully the fault uncertainty can still be narrowed down as we keep collecting data from field studies. As the source parameters become more and more certain, they can be used as uncertain observations to calibrate the experts. The idea is to use any valuable information about source geometry.

6 Experts' Combinations and Calibration Subjects

6.1 Data Set Description

The experts' data set represents the 80 single combinations (permutations) of expert opinions for one plant, see Figure (6). The combinations were extracted from the combined logic tree model for one of the sites. For each combination the PEGASOS hazard model (logic tree) can compute a set of hazard curves for different spectral ordinates, $freq(Hz)$ which can be represented as hazard spectra (plot of spectral acceleration, versus frequency and probability of exceedance).

A single combination of expert opinion (from the 80) for a fixed probability of exceedance (fixed values of magnitude and distance) can be approximated as a lognormal distribution. This follows from the basic SSHAC seismic hazard model. Seismologists still discuss whether or on what level this distribution shall be truncated. In the PEGASOS-project an upper limit truncation was performed.

For each combination the PEGASOS algorithm (logic tree) can compute a set of hazard curves for different seismic spectra, $freq(HZ)$.

$$PAGASOS(80 \text{ experts}|freq) \Rightarrow 80 \text{ hazard curves}|freq; \quad (28)$$

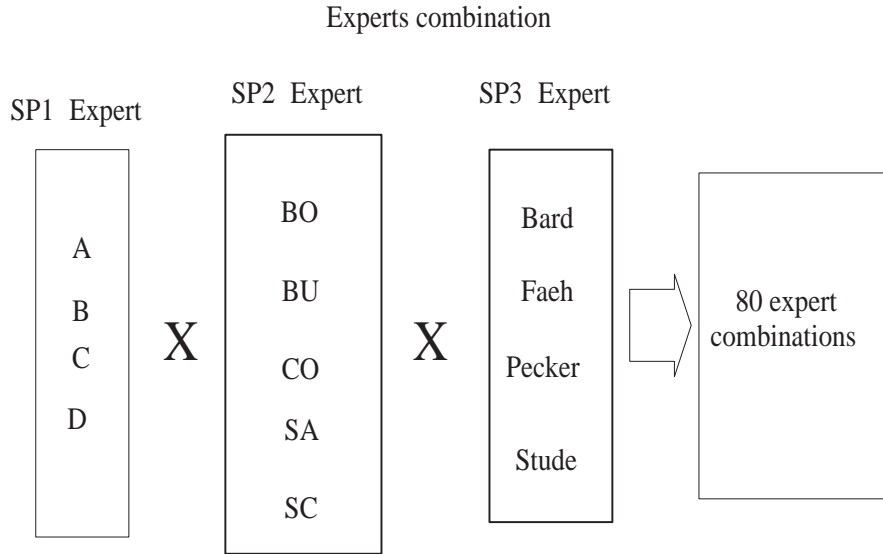


Figure 6: 80 experts

where $\text{freq} = [0.5, 1, 2, 4, \dots](\text{HZ})$.

The curves are given by ASCII files. The file names are combinations of the expert's names in the three subprojects with some parameters, as: $\langle \text{Type} \rangle - \langle \text{SP1 - Group} \rangle - \langle \text{site} \rangle - \langle \text{SP2 - Expert} \rangle - \langle \text{rec.index} \rangle - \langle \text{SP3 - Expert} \rangle - \langle \text{Bodenbewegung} \rangle .\text{asc}$, see [7][Soil Hazard Resultate].

The x -coordinate of the curves is spectral acceleration. The y -coordinate is yearly exceedance probability. Each data file contains five quantile curves.

The plots of some of those hazard curves are shown as Figure(7): Each of the hazard plots is for one earthquake spectral ordinate, $\text{freq}(\text{HZ})$. There are in total 11 spectral ordinates. Plotting the $\text{freq}(\text{HZ})$, acceleration (g) and exceedance probability ($/\text{year}$) together, we have the surface plot shown as Figure(8). Note, only 0.05 and 0.95 quantile surfaces are plotted.

Suppose there is a plane vertical with z axis. It intersects with the 5 quantile surfaces. Since z (exceedance probability) is strictly decreasing along the y (acceleration) axis, we have exactly 5 intersection curves. Project the curves into the plane of $\text{acceleration}(g)$ *v.s.* $\text{freq}(\text{HZ})$, then we have quantile curves of *hazard spectral plot* shown as Figure(9).

The hazard spectral curves are for each exceedance probability. Each realization point in a quantile interval of the surface plot will end up in the same quatnile interval of the hazard spectral plot. Calibration of the hazard surfaces can be done by calibrating the hazard spectral curves.

In next sections we try to answer following questions

- Do experts' performances differ significantly?
- Does performance DM differ far from equal weight DM?
- Effect of the noise term.
- When does the performance DM converge as we add up the number of seed variables and experts?

7 Aggregation Based on Synthetic Seed Variables

It is important to see how experts and DM perform based on different seed variables (realizations). If we can build a simple model to simulate the seismic hazard, then we may sample synthetic seed

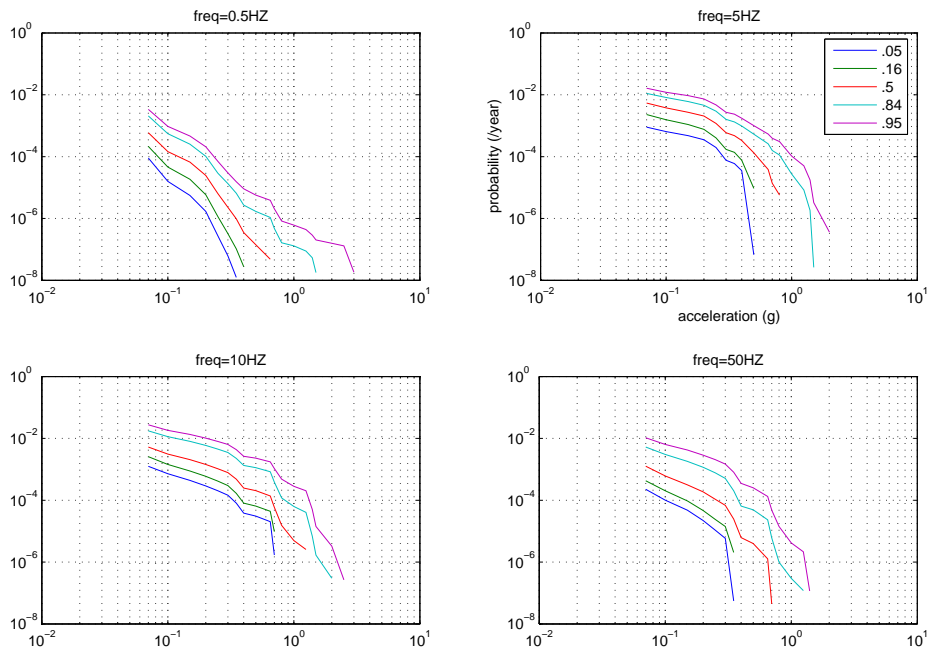


Figure 7: hazard curves from an expert

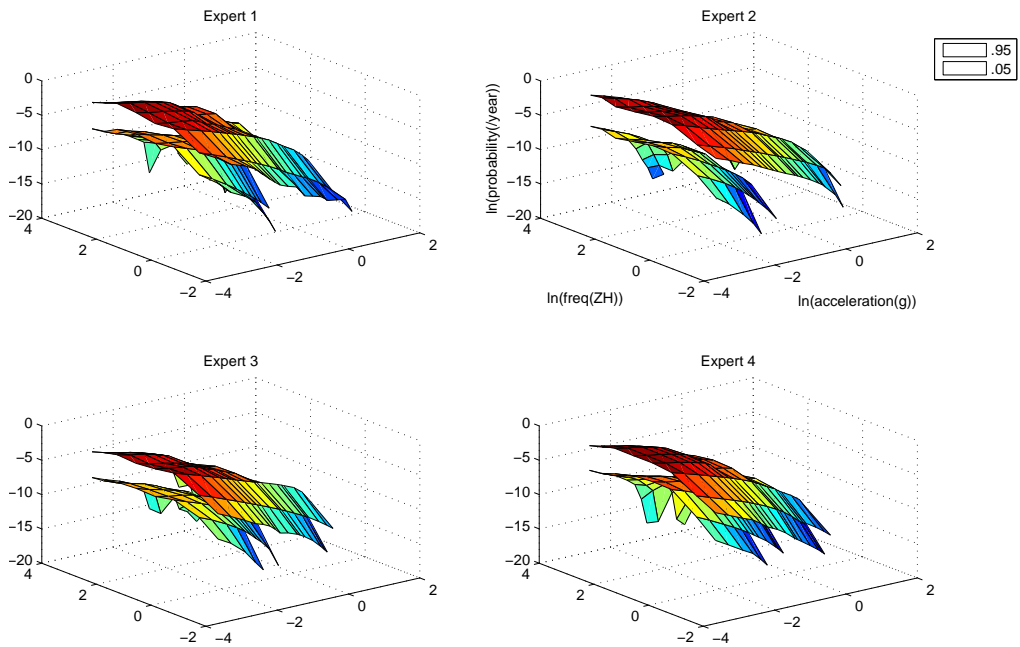


Figure 8: hazard surface plot of 4 experts

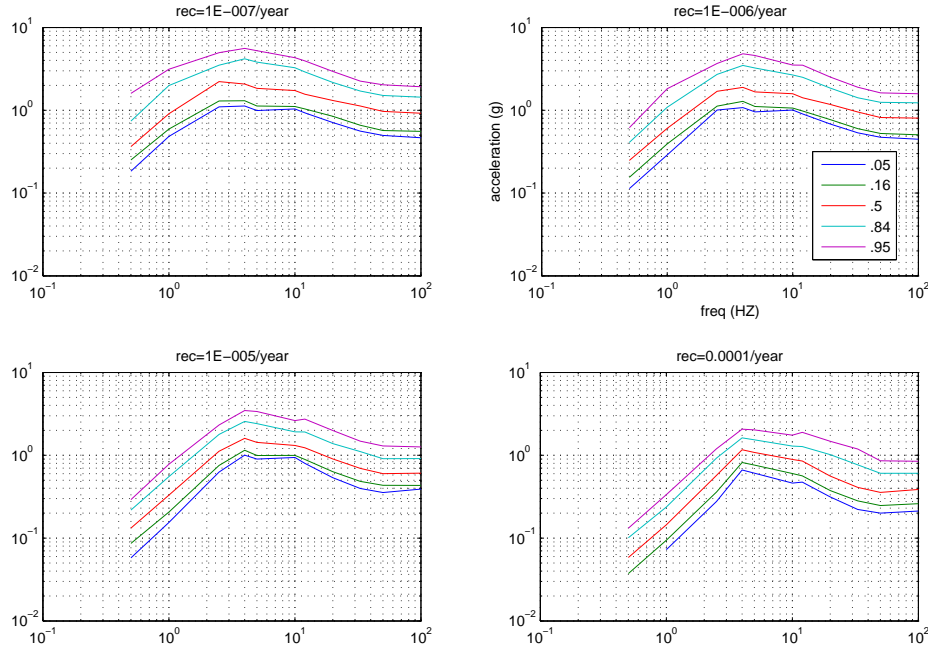


Figure 9: hazard spectral curves of an expert

variables out of the the model. Using those variables, we can perform the scoring rules on the experts. An interesting result could be that for each different seed variable set, the experts' performances vary significantly. This means the performance based DM (per-DM) could be very different with equal weight DM (ep-DM). If the per-DM has largely better performances than eq-DM, then the effort to find (real) seed variables is worth to make.

7.1 Groen's Model and Dependence Structure

A Monte Carlo procedure was built by Dr. Frank Groen, [2][Groen, 2004], to simulate the seismicity for each nuclear site in Switzerland. Each hazard curve is the result of a separate Monte Carlo procedure. After many hazard curves are calculated, quantile curves are approximated from the the curve set.

In his approach the model inputs are randomly sampled from given distributions. In my previous study, [8][Meng 2005], I introduce a plausible dependence structure and showed its effects.

The task in that study included two parts:

- Build dependence structure between input variables.
- Use dependent sampling procedure to sample inputs.

After that, I use Groen's statistic routine to estimate ground motion hazard curves.

The objective of that effort is to show how the hazard curves behave due to the dependence between random inputs. There are two dependence structures used as analogs, vine and joint normal. Ideally, under the same correlation assumptions, the choice of the dependence structure does not make much of difference.

It shows that the dependence specification may have an impact on the hazard curves and it can reduce the uncertainty bandwidth of the curves. Due to some optimistic assumptions made by the author, the hazard curves are below the original Groen's outcomes.

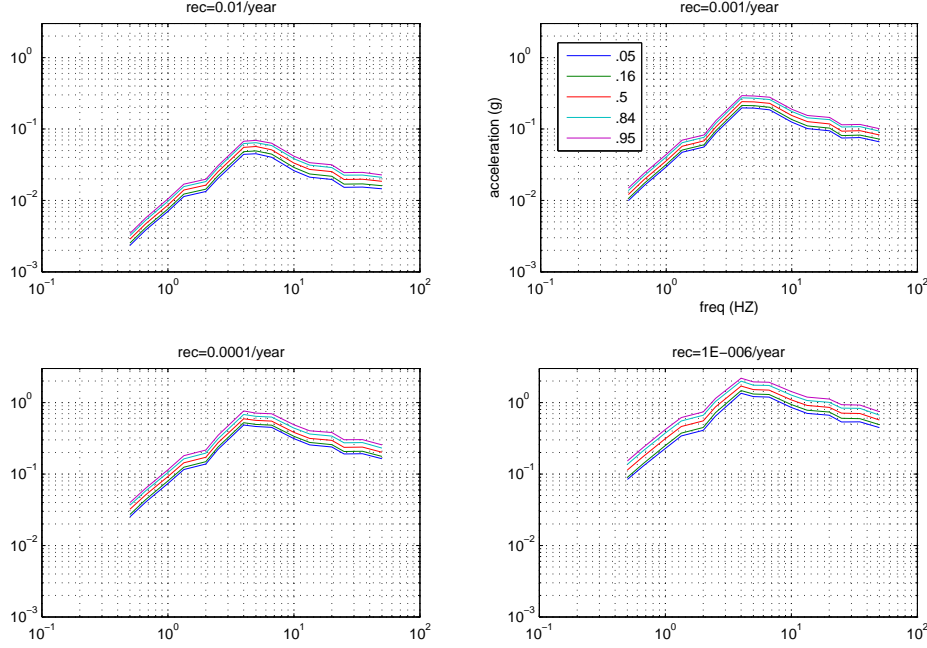


Figure 10: hazard spectra without dependence

The results from the simulation (Groen with/without dependence) include quantile curves of hazard spectral plots for different exceedance probabilities. Some of them are shown in Figure(10). These plots can be seen as the counterpart of the PEGASOS results, Figure(10). From them we can sample some synthetic seed variables.

7.2 Seed Variable Sampling

The sample procedure is described as:

- Choose an event recurrence rate $rec(/year)$ and an observed spectral ordinate $freq(HZ)$, the choice see next section;
- In Groen's hazard spectral plot of the chosen rec , draw a vertical line at the $freq$ value intersecting with the 5 quantile curves by 5 quantile points;
- Interpolate the distribution function as piece wise uniform within quantile intervals. This distribution is proven to be the minimal informative distribution w.r.t. the uniform background measure.
- Sample random variable U from a uniform distribution over $[0, 1]$.
- Through the inverse transform, $X = F^{-1}(U)$, find the synthetic seed variable.

Note, the sample distributions are interpolated in log -scale.

7.2.1 Choice of rec and $freq$ Values

To make the synthetic seed variables realistic, the ideal choice of the rec value should be as big as possible, say 10^{-1} . Thus, for average 10 years we may have 1 local (in Switzerland) event. For this event we may record earthquake time histories at different sites and develop an observation

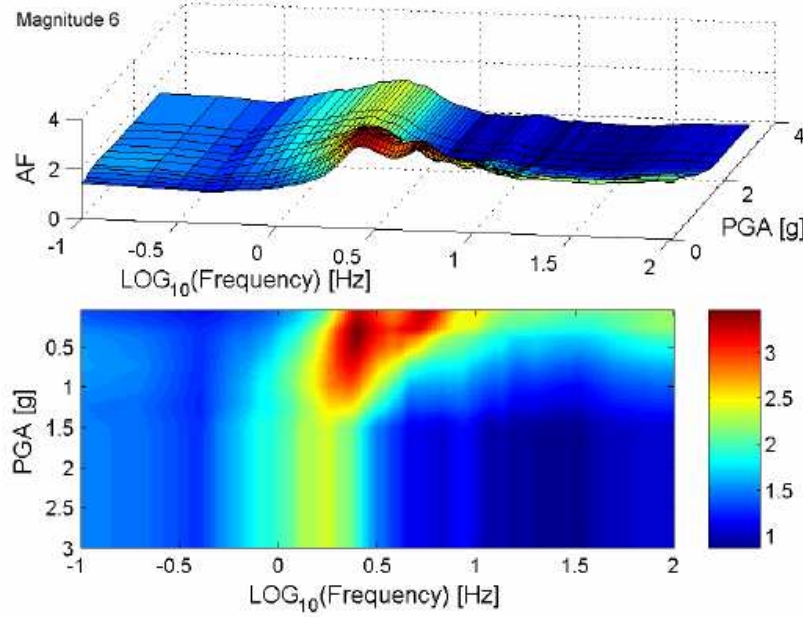


Figure 11: amplification factor (AF) from SP3 report

set of hazard spectra for different sites spectral acceleration versus frequency. These values are realizations of the PEGASOS prediction. If we can identify 100 independent seismic observation sites with the similar seismicity, we may have 10 global events each year. Unfortunately, the current data from PEGASOS only includes magnitude values larger than 5.0. The *rec* ranges assessed by different experts are different. If all the experts are to be calibrated, the biggest choice of *rec* value is 10^{-4} . For up to $rec = 10^3/\text{year}$, the extraction of data showed that some expert combinations turned out to be 0.

Note, the hazard spectra are developed based on the solution of the vibration equation for a single degree of freedom system with a specified damping (in our case 5%) with a natural frequency corresponding to *freq*(HZ) in this study.

The choice of *rec* and *freq* values is also limited by something else. In Groen's approach, the site response (SP3) is not taken into consideration. This results a significant discrepancy between Groen and PEGASOS. In SP3, experts gave assessments about the amplification effect of ground motion from rock to soil. We found that for large ground motions at some *freq*(HZ) values, Groen's results and PEGASOS match quite well. This may suggest that the amplification effect is not significant. The blue area in Figure (11) quoted from [4][PEGASOS vol.6] confirms this suggestion. To avoid the amplification effect, we need to sample seed variables with large ground motion. This means we can only choose low exceedance recurrence rate (*rec*/year). This compromise could be avoided by encoding the site response into the simulation.

Now, the sample values from Groen (with or without dependence) can be used to calibrate experts directly.

7.3 Performance Based Decision Maker

Based on the synthetic seed variables the calibration and information scoring rules are applied on the experts' curves. The performance weight of an expert is proportional to the product of his calibration and information scores. The decision maker (DM) distribution is the weighted average over the experts' distributions. The DM quantiles are derived from its distribution. The scoring rules are also applied on the DM's distribution to calculate its unnormalized weight.

	SP 1	SP 2	SP 3
EX 1	EG1a	BO	Bard
EX 2	EG1b	CO	Faeh
EX 3	EG1c	SA	Peck
EX 4	EG1d	BO	Studer
EX 5	EG1a	CO	Studer

Table 1: expert list

Set the optimal *cutoff level* α , see section 4.5, such that the DM's unnormalized weight is maximized.

8 Case Study

To demonstrate the decision making procedure, We arbitrarily choose 5 combinations out of 80 as our *experts*, see Table (1).

A cluster of Matlab programs were written by the author to implement expert judgment on the experts. The results were checked by a program called Excalibur, [9][Exalibur 2000], developed in Delft University of Technology, see <http://delta.am.ewi.tudelft.nl/risk/>. The data flow is shown in Figure (12).

8.1 Multiple Seismic Events

Suppose there are 3 events. All the events are with yearly exceedance probability $rec = 10^{-6}/\text{year}$. For each event we sample the seed variables at 5 different spectra (*freq*) values. Thus, there are in total 15 synthetic seed variables. The sampling method for a given distribution, recall Section (7.2), is to uniformly sample $U \sim U[0, 1]$ and do inverse transform as Figure (13). Through this method, a set of seed variables are sampled as Figure (14). These seed variables are sampled from Groen's results without optimistic dependence assumptions. The corresponding seed list and scoring results (including weights) are in Table (2). Where, DM¹ is the performance based decision maker and eq_DM is the equal weight decision maker. The Matlab code and Excalibur produce the same DM and eq_DM².

For each expert/DM there is a data-fit hazard spectra plot, see Figure (15). In the Figure (15), at each seed variable draw a vertical line intersecting with the curves by 5 quantile points, see Figure (16). The quantile curves are given by discrete points. Linear interpolations are made to calculate the positions of the points in Figure (16). From the quantile points and seed variables, the Matlab code generates Excalibur input files.

Note, in the Matlab code the aggregations are done at the discrete curve points. Thus, it can produce quantile curves in the same format as the inputs. However in Excalibur, the aggregations are done at the seed variable points. This explains the slight difference between the Matlab and Excalibur results.

Through Excalibur, we can see the range graphs of the input/output data. Figure (26-31), see Appendix, show the range graphs for the experts and DM. From those graphs, we see how PE-GASOS and Groen's results fit each other. Note, the intrinsic ranges of the graphs are calculated as Section (4.4.1).

The seed variables are randomly sampled from the simulation results. For each sampling trial we have a different seed variable set therefor we have a different score table and DM. Sample seed

¹For sake of simplicity, we denote the performance based decision maker as DM through out this report

²Note, DM has better performance than eq_DM in both calibration and information scores

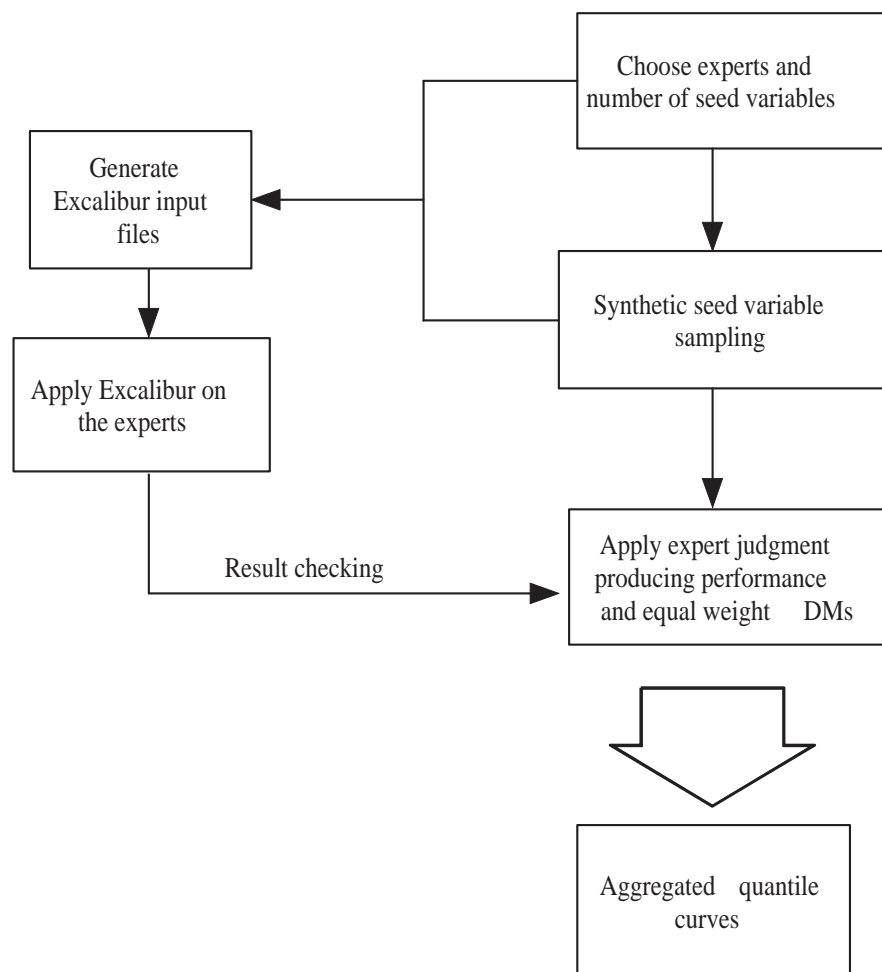


Figure 12: data flow of the Matlab code

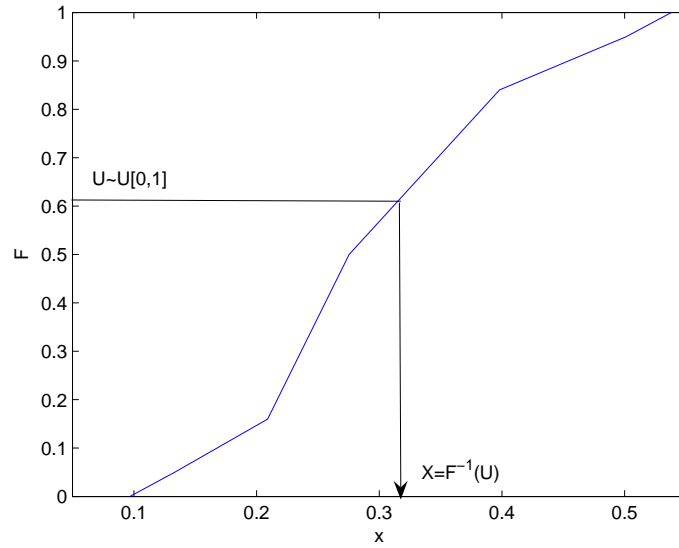


Figure 13: sample by inverse uniform transform

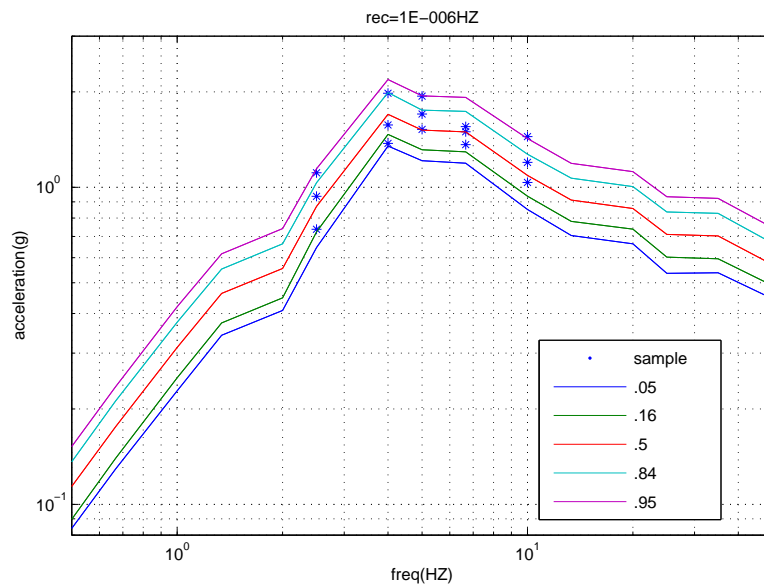
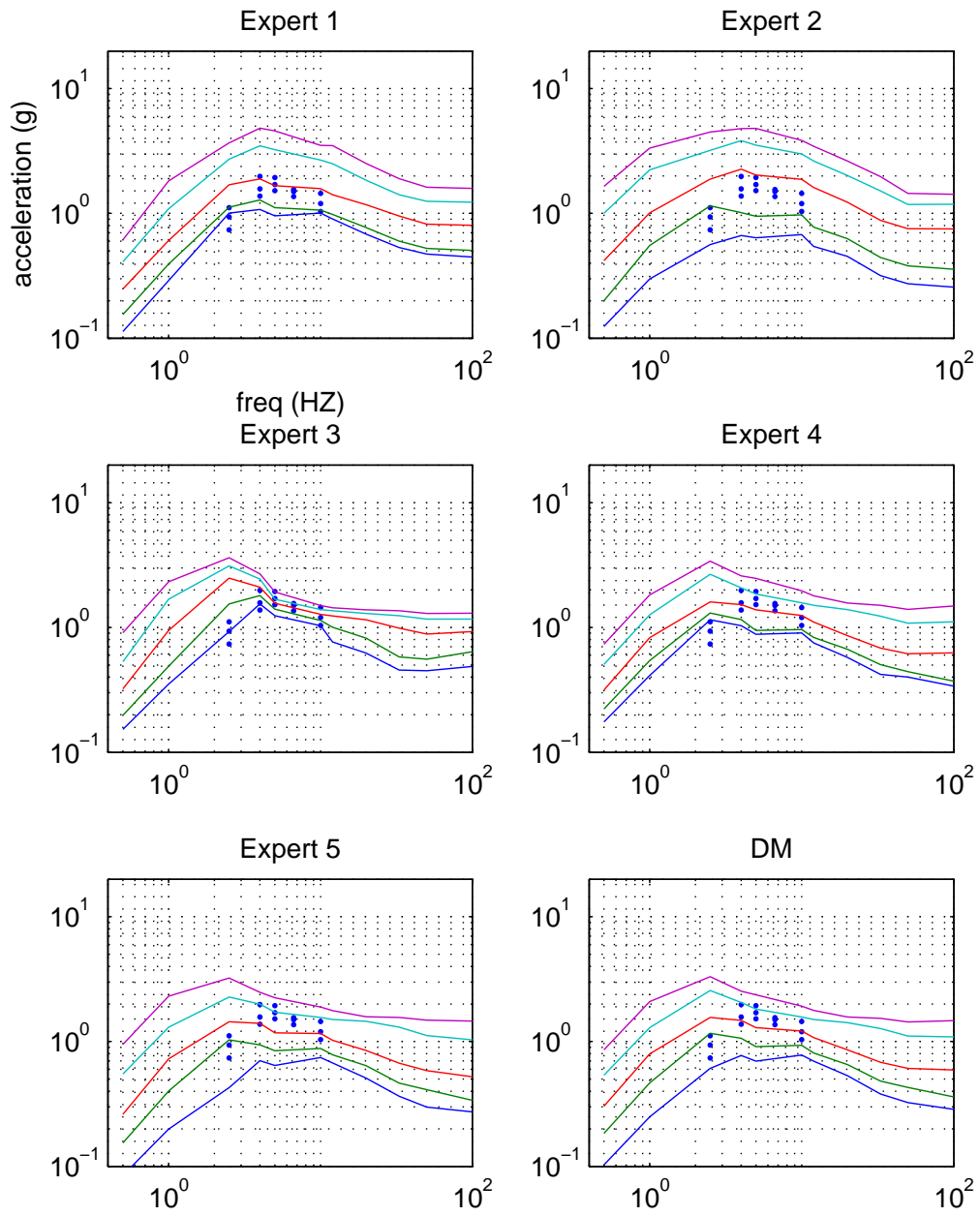


Figure 14: seed variables sampled from Groen

Figure 15: data fit plot ($rec = 10^{-6}/year$)

seed ID	value	full name
seed1	0.73774	acc(g) rec=1E-006/year,freq=2.5HZ
seed2	1.3767	acc(g) rec=1E-006/year,freq=4HZ
seed3	1.5227	acc(g) rec=1E-006/year,freq=5HZ
seed4	1.5543	acc(g) rec=1E-006/year,freq=6.6667HZ
seed5	1.1992	acc(g) rec=1E-006/year,freq=10HZ
seed6	0.9372	acc(g) rec=1E-006/year,freq=2.5HZ
seed7	1.5739	acc(g) rec=1E-006/year,freq=4HZ
seed8	1.7039	acc(g) rec=1E-006/year,freq=5HZ
seed9	1.3634	acc(g) rec=1E-006/year,freq=6.6667HZ
seed10	1.0366	acc(g) rec=1E-006/year,freq=10HZ
seed11	1.1119	acc(g) rec=1E-006/year,freq=2.5HZ
seed12	1.98	acc(g) rec=1E-006/year,freq=4HZ
seed13	1.9396	acc(g) rec=1E-006/year,freq=5HZ
seed14	1.4981	acc(g) rec=1E-006/year,freq=6.6667HZ
seed15	1.4464	acc(g) rec=1E-006/year,freq=10HZ

	expert	Calibration	Information	weight		Id	Calibr.	Mean relative information	UnNormaliz weight	Normaliz.w without DM
	1	0.12111	0.37253	0.075177		EX1	0.1211	0.3725	0.04512	0.07516
Matlab	2	0.00020562	0.10934	0	Excalibur	EX2	0.000206	0.1093	0	0
	3	0.24682	1.3216	0.54352		EX3	0.2469	1.322	0.3263	0.5436
	4	0.045117	0.90371	0.067939		EX4	0.04509	0.9037	0.04075	0.06788
	5	0.23533	0.79915	0.31336		EX5	0.2354	0.7991	0.1881	0.3134
	DM	0.5019	0.74043	NaN		DM	0.5018	0.7301	0.3664	
	eq_DM	0.18684	0.37214	NaN		eq_DM	0.1868	0.3659	0.06835	

Table 2: Seed set sampled from Groen without dependence; scores and weights of the experts and DM/eq_DM.

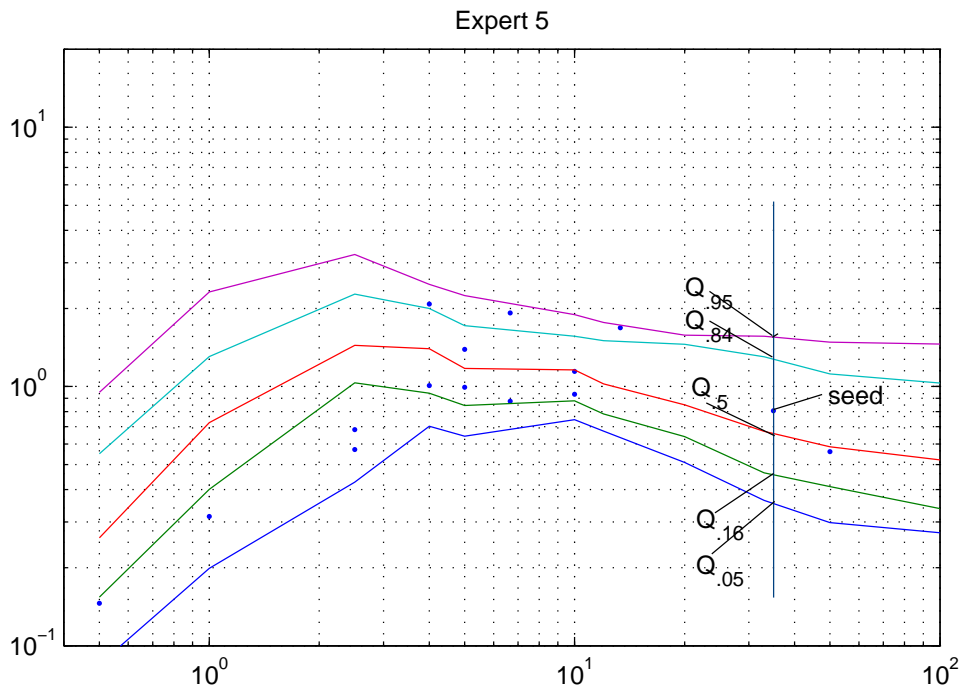


Figure 16: seed and quantile points of Expert 5

seed ID	trial1	trial2	trial3	trial4	trial5	trial6	full name
seed1	0.86417	0.64512	0.87714	1.1578	0.76286	0.66225	acc(g) rec=1E-006/year,freq=2.5HZ
seed2	1.8738	1.7831	2.0504	1.9699	1.5551	1.6134	acc(g) rec=1E-006/year,freq=4HZ
seed3	1.7287	1.3746	1.3697	1.583	1.8474	1.8003	acc(g) rec=1E-006/year,freq=5HZ
seed4	1.9457	1.5304	1.5681	1.1661	1.6764	1.4219	acc(g) rec=1E-006/year,freq=6.6667HZ
seed5	1.1978	1.0279	1.378	1.1218	0.85131	1.0095	acc(g) rec=1E-006/year,freq=10HZ
seed6	0.17466	0.19728	0.17291	0.1938	0.21263	0.16882	acc(g) rec=1E-007/year,freq=0.5HZ
seed7	0.33317	0.53837	0.31751	0.64675	0.52061	0.32899	acc(g) rec=1E-007/year,freq=1HZ
seed8	1.1506	1.7264	1.7256	1.2665	1.6991	1.3012	acc(g) rec=1E-007/year,freq=13.34HZ
seed9	1.0672	1.0914	0.91245	1.0933	0.73896	0.66815	acc(g) rec=1E-007/year,freq=35HZ
seed10	0.59736	0.80202	0.88163	0.65573	1.0321	0.6469	acc(g) rec=1E-007/year,freq=50HZ
seed11	0.42304	0.68773	0.42144	0.49598	0.54319	0.41951	acc(g) rec=1E-005/year,freq=2.5HZ
seed12	1.4294	0.91165	1.1862	1.4333	1.2884	1.2641	acc(g) rec=1E-005/year,freq=4HZ
seed13	1.313	1.1257	0.93743	0.88212	0.83918	0.95073	acc(g) rec=1E-005/year,freq=5HZ
seed14	1.0498	1.0324	1.1472	0.9565	0.95833	1.0919	acc(g) rec=1E-005/year,freq=6.6667HZ
seed15	0.79849	0.6794	0.64583	0.63048	0.62598	0.59611	acc(g) rec=1E-005/year,freq=10HZ

Table 3: synthetic seed samples

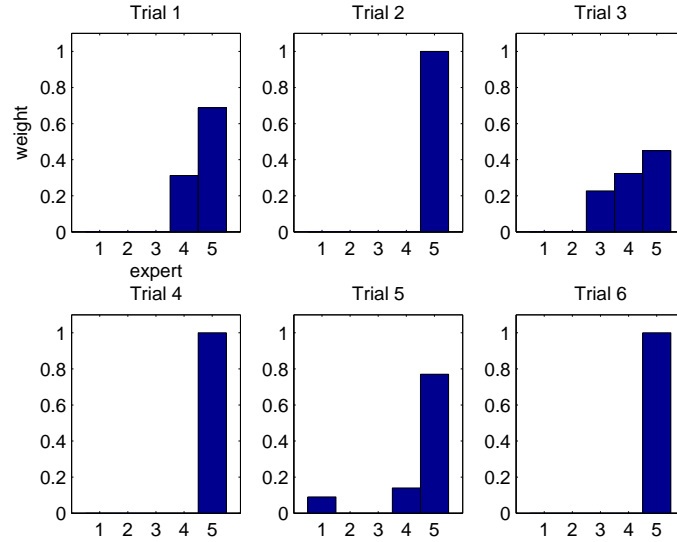


Figure 17: weight plot for the 6 DMs

variables from different *rec/year* and *freq(HZ)* values, see table (3).

The corresponding weight plot of the 6 DMs is shown in Figure(17). From the figure, we can see the differences between DMs.

The corresponding scores are shown in Table (4).

These trial results show

- For each expert, the information scores are almost identical across the trials.
- The calibration of an expert varies over the trials roughly with factor 4.
- Expert 5 has non-zero weight in all the trials. In half of the trials it has all the weight. This suggests Expert 5 is the best performance expert.
- The DM is well calibrated and reasonably informative in most of the trials.
- In some trials, the eq_DM is well calibrated, see trial 1 and 3. In other trials, it is badly calibrated. Its information score is always below the DM.

Note, all the trials are checked by Excalibur, see Table (21) in Appendix.

	expert	Calibration	lformation	weight		expert	Calibration	lformation	weight
	1	0.00043367	0.42665	0		1	0.0053765	0.42665	0
Matlab	2	0.0006228	0.13197	0	Matlab	2	0.0001699	0.13197	0
trial1	3	0.0013324	0.9831	0	trial2	3	5.2237e-005	0.9831	0
	4	0.099605	0.77976	0.3121		4	0.022632	0.77976	0
	5	0.25047	0.68347	0.6879		5	0.19709	0.68347	1
	DM	0.49196	0.65817	NaN		DM	0.19709	0.68023	NaN
	eq_DM	0.17562	0.34564	NaN		eq_DM	0.063968	0.34564	NaN
	expert	Calibration	lformation	weight		expert	Calibration	lformation	weight
	1	0.0022248	0.42679	0		1	0.0066959	0.42665	0
Matlab	2	0.00081815	0.13209	0	Matlab	2	0.0039772	0.13197	0
trial3	3	0.002793	0.98325	0.22657	trial4	3	8.5811e-005	0.9831	0
	4	0.0050285	0.77992	0.32357		4	0.034868	0.77976	0
	5	0.0079766	0.68356	0.44986		5	0.18684	0.68347	1
	DM	0.4777	0.63095	NaN		DM	0.18684	0.68023	NaN
	eq_DM	0.18494	0.34577	NaN		eq_DM	0.030461	0.34564	NaN
	expert	Calibration	lformation	weight		expert	Calibration	lformation	weight
	1	0.0037512	0.42665	0.090256		1	0.00084037	0.42667	0
Matlab	2	0.0015935	0.13197	0	Matlab	2	0.00081815	0.13198	0
trial5	3	4.2446e-005	0.9831	0	trial6	3	3.8919e-006	0.98312	0
	4	0.0031768	0.77976	0.1397		4	0.0019405	0.77977	0
	5	0.019979	0.68347	0.77005		5	0.36631	0.68348	1
	DM	0.049272	0.56747	NaN		DM	0.36631	0.68023	NaN
	eq_DM	0.020309	0.34564	NaN		eq_DM	0.028025	0.34565	NaN

Table 4: score table for the 6 trials

8.2 Dependence Assumption Effect

The seed variables above are sampled from the original [2][Groen] model. We may also sample seed variable sets from Groen plus optimistic dependence (vine/joint normal) assumptions. Based on the seed variables sampled from different sources, we compare the scores of the experts and DM. The objective is to see the ultimate impact of the dependence assumptions on the DM.

8.2.1 Unify the Sampling Randomness

Let us recall the sampling procedure mentioned in Section (7.2) and the inverse uniform transform in Figure (13).

Suppose there are 2 distributions, F_Y and F_Z , as 2 sample sources. From them we sample seed variables. The way to compare these sources is to use the same uniform set, U values in Figure (13), to sample seed variable sets of Y and Z . Thus, we can compare these two sets and, based on them, the performances of the experts and DMs.

Short hand notations:³

- Denote Groen without dependence assumption as (seed variable) Source 1;
- Groen with the joint normal dependence as Source 2;
- Groen with the vine dependence as Source 3.

Recall the data flow diagram, Figure (12). It includes two parallel procedures using Excalibur and the Matlab code. These two procedures are functionally identical. An easy way to compare the different sample sources is to redesign the data flow as Figure (18).

³Details of the dependence assumptions see report [8][Meng, 2005]. Different seed variable sources are compared in terms of experts/DM's calibration and information scores.

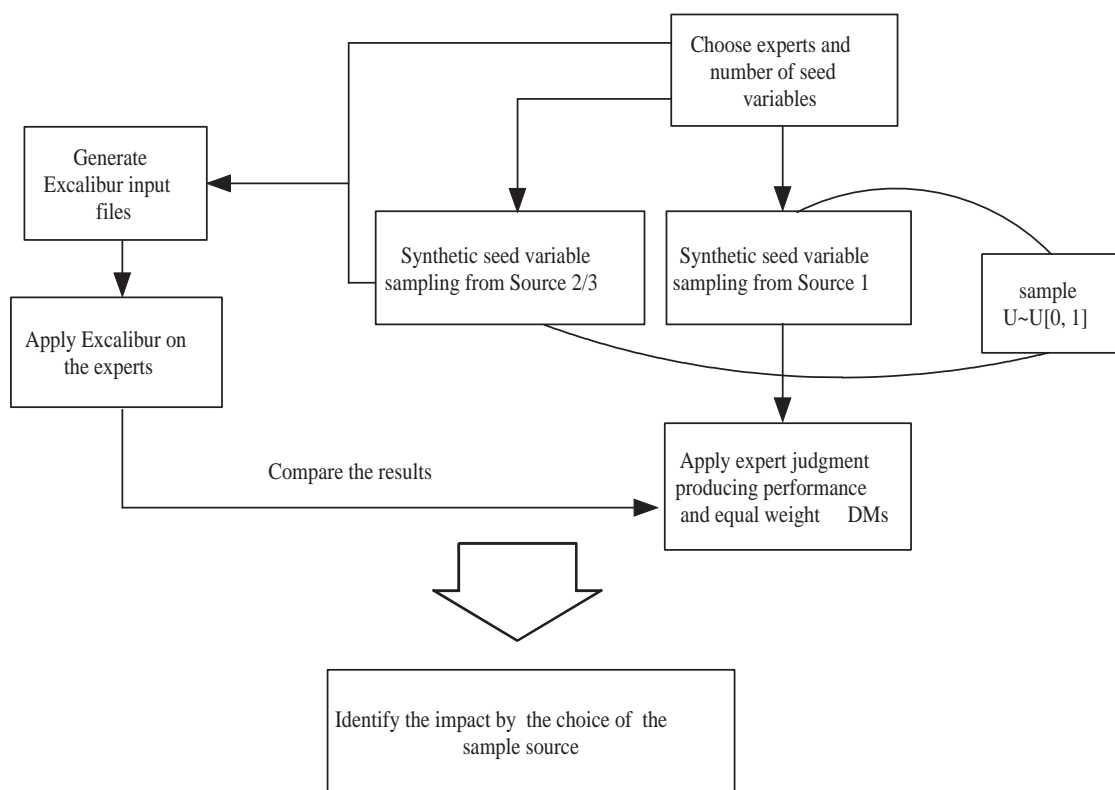


Figure 18: sample source comparison

seed ID	Source 1	Source 2	full name
seed1	0.982	0.19497	acc(g) rec=1E-006/year,freq=2.5HZ
seed2	1.8031	0.43157	acc(g) rec=1E-006/year,freq=4HZ
seed3	1.4377	0.32399	acc(g) rec=1E-006/year,freq=5HZ
seed4	1.7441	0.48517	acc(g) rec=1E-006/year,freq=6.6667HZ
seed5	1.3836	0.39345	acc(g) rec=1E-006/year,freq=10HZ
seed6	0.14322	0.01073	acc(g) rec=1E-007/year,freq=0.5HZ
seed7	0.49503	0.047621	acc(g) rec=1E-007/year,freq=1HZ
seed8	1.3031	0.19769	acc(g) rec=1E-007/year,freq=13.34HZ
seed9	1.3082	0.2319	acc(g) rec=1E-007/year,freq=25HZ
seed10	0.64452	0.077483	acc(g) rec=1E-007/year,freq=50HZ
seed11	0.50394	0.12275	acc(g) rec=1E-005/year,freq=2.5HZ
seed12	1.1727	0.44622	acc(g) rec=1E-005/year,freq=4HZ
seed13	0.93724	0.31217	acc(g) rec=1E-005/year,freq=5HZ
seed14	1.2481	0.60545	acc(g) rec=1E-005/year,freq=6.6667HZ
seed15	0.71127	0.25298	acc(g) rec=1E-005/year,freq=10HZ

Table 5: 2 seed sets sampled from Source 1, 2

	expert	Calibration	Information	weight		Id	Calibr.	Mean relative information	UnNormaliz weight	Normaliz.w without DM
	1	0.011022	0.42742	0		EX1	7.16E-18	0.4758	0	0
Matlab	2	0.00078483	0.13114	0	Excalibur	EX2	5.32E-16	0.17	9.042E-17	0.1874
Source 1	3	0.034163	1.0044	0.76187	Source 2	EX3	7.16E-18	1.055	0	0
	4	0.001136	0.79302	0		EX4	7.16E-18	0.8424	0	0
	5	0.015421	0.69547	0.23813		EX5	5.32E-16	0.7369	3.921E-16	0.8126
	DM	0.36631	0.77115	NaN		DM	5.32E-16	0.4981	2.65E-16	
	eq_DM	0.17562	0.35012	NaN		eq_DM	5.32E-16	0.3833	2.039E-16	

Table 6: trial results on Source 1, 2

8.2.2 Impact of the Joint Normal Dependence

Here we sample two sets of seed variables out of Source 1 and Source 2, see Table (5). From the table, we see that samples from Source 2 are much lower than corresponding samples from Source 1. This is because the optimistic dependence brings the hazard (acceleration) down. The score/weight tables from the Matlab code and Excalibur are shown as Table (6). From the table, the experts/DM are badly calibrated based on Source 2 seed variable set. For each expert the calibrations differ by scale of 10^{-4} over the two trials.

8.2.3 Impact of the Vine Dependence

Sample seed variable sets from Source 1 and Source 3 and compare the results as above. The seed sets and results are in Table (7) and Table (8). The results are similar with last section.

8.2.4 Comparison between Joint Normal and Vine

These two dependence structures share the same correlation matrix. Ideally, they have the same impact on the experts' performances and DM. Even though the experts/DM have bad performances, we can still compare these two dependence structures. Give the seed sets sampled from Source 2 and Source 3 to the Matlab code and Excalibur. The sample sets and the results are in Table (9),(10). The two seed sets sampled from Source 2 and Source 3 are very close as expected. The experts' performances and DM are the same within the trials. This result confirms the statement at beginning.

seed ID	Source 1	Source 3	full name
seed1	0.86417	0.13188	acc(g) rec=1E-006/year,freq=2.5HZ
seed2	1.8738	0.41936	acc(g) rec=1E-006/year,freq=4HZ
seed3	1.7287	0.46897	acc(g) rec=1E-006/year,freq=5HZ
seed4	1.9457	0.58605	acc(g) rec=1E-006/year,freq=6.6667HZ
seed5	1.1978	0.30079	acc(g) rec=1E-006/year,freq=10HZ
seed6	0.17466	0.014209	acc(g) rec=1E-007/year,freq=0.5HZ
seed7	0.33317	0.031623	acc(g) rec=1E-007/year,freq=1HZ
seed8	1.1506	0.179	acc(g) rec=1E-007/year,freq=13.34HZ
seed9	1.0945	0.16544	acc(g) rec=1E-007/year,freq=25HZ
seed10	0.59736	0.08968	acc(g) rec=1E-007/year,freq=50HZ
seed11	0.42304	0.1	acc(g) rec=1E-005/year,freq=2.5HZ
seed12	1.4294	0.55301	acc(g) rec=1E-005/year,freq=4HZ
seed13	1.313	0.59089	acc(g) rec=1E-005/year,freq=5HZ
seed14	1.0498	0.41953	acc(g) rec=1E-005/year,freq=6.6667HZ
seed15	0.79849	0.32743	acc(g) rec=1E-005/year,freq=10HZ

Table 7: 2 seed sets sampled from Source 1, 3

	expert	Calibration	lformation	weight		Id	Calibr.	Mean relative	UnNormaliz	Normaliz.w
								information	weight	without DM
	1	0.00043367	0.42742	0		EX1	7.16E-18	0.4736	0	0
Matlab	2	0.0006228	0.13114	0	Excalibur	EX2	9.59E-15	0.1687	1.618E-15	0.1865
Source 1	3	0.0013324	1.0044	0	Source 3	EX3	7.16E-18	1.052	0	0
	4	0.099605	0.79302	0.31199		EX4	7.16E-18	0.8403	0	0
	5	0.25047	0.69547	0.68801		EX5	9.59E-15	0.7356	7.054E-15	0.8135
	DM	0.49196	0.67061	NaN		DM	9.59E-15	0.4974	4.769E-15	
	eq_DM	0.11275	0.35012	NaN		eq_DM	5.32E-16	0.382	2.032E-16	

Table 8: trial results on Source 1, 3

seed ID	Source 2	Source 3	full name
seed1	0.19497	0.17837	acc(g) rec=1E-006/year,freq=2.5HZ
seed2	0.43157	0.39648	acc(g) rec=1E-006/year,freq=4HZ
seed3	0.32399	0.34977	acc(g) rec=1E-006/year,freq=5HZ
seed4	0.48517	0.46217	acc(g) rec=1E-006/year,freq=6.6667HZ
seed5	0.39345	0.37981	acc(g) rec=1E-006/year,freq=10HZ
seed6	0.01073	0.01073	acc(g) rec=1E-007/year,freq=0.5HZ
seed7	0.084684	0.079837	acc(g) rec=1E-007/year,freq=1.34HZ
seed8	0.19397	0.21245	acc(g) rec=1E-007/year,freq=20HZ
seed9	0.2319	0.22997	acc(g) rec=1E-007/year,freq=35HZ
seed10	0.077483	0.10332	acc(g) rec=1E-007/year,freq=50HZ
seed11	0.12275	0.12275	acc(g) rec=1E-005/year,freq=2.5HZ
seed12	0.44622	0.40608	acc(g) rec=1E-005/year,freq=4HZ
seed13	0.31217	0.34229	acc(g) rec=1E-005/year,freq=5HZ
seed14	0.60545	0.55563	acc(g) rec=1E-005/year,freq=6.6667HZ
seed15	0.25298	0.26773	acc(g) rec=1E-005/year,freq=10HZ

Table 9: 2 seed sets sampled from Source 2, 3

	expert	Calibration	Information	weight		Id	Calibr.	Mean relative information	UnNormaliz weight	Normaliz.w without DM
	1	0	0.4629	0		EX1	7.16E-18	0.4626	0	0
Matlab	2	5.5511e-016	0.16855	0.19151	Excalibur	EX2	5.32E-16	0.1681	8.943E-17	0.1912
Source 2	3	0	1.0164	0	Source 3	EX3	7.16E-18	1.016	0	0
	4	0	0.82064	0		EX4	7.16E-18	0.8203	0	0
	5	5.5511e-016	0.71158	0.80849		EX5	5.32E-16	0.7111	3.783E-16	0.8088
	DM	5.5511e-016	0.48513	NaN		DM	5.32E-16	0.4792	2.55E-16	
	eq_DM	5.5511e-016	0.38329	NaN		eq_DM	5.32E-16	0.3723	1.981E-16	

Table 10: trial results on Source 2, 3

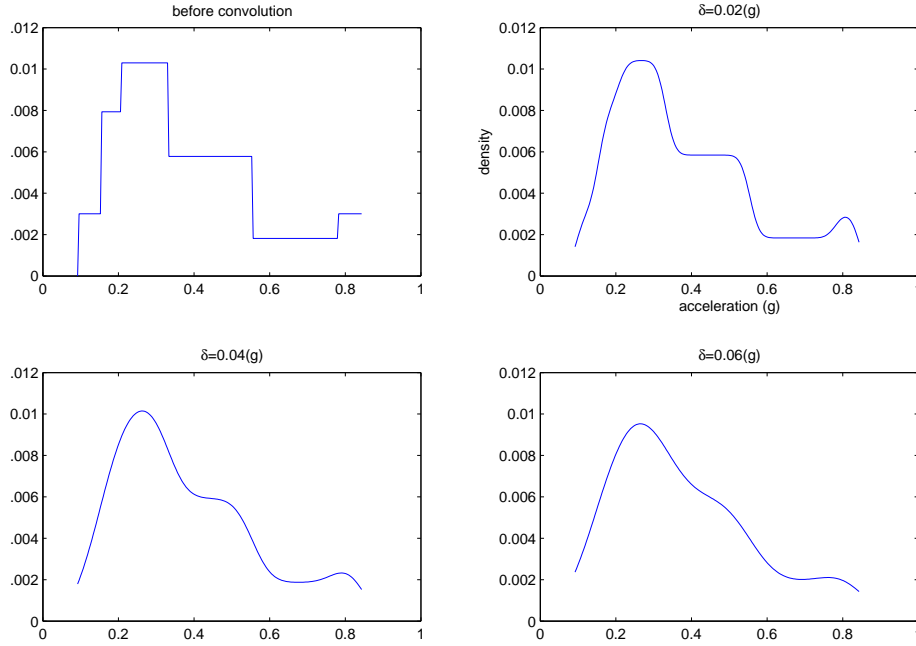


Figure 19: noise effect on an expert distribution

8.3 Noise Effect

8.3.1 Kraan 1

In many cases the observations contain some uncertainties (noise) which experts are not supposed to fold into their assessments. To investigate the noise effect, we have to know the noise term, recall the ϵ in Section (4.6). We assume ϵ is normally distributed with 0 mean and standard deviation δ .

With the noise term and experts' quantiles, we can apply the convolution, Formula (11) in Section (4.6.2), on the experts. Technically it folds the noise into the experts' distributions, Kraan 1 [5][Kraan].

Let us first see how an expert distribution looks before and after convolution. We choose the quantiles given by Expert 1 for the seed variable $acc|rec = 1E - 005/year, spec = 1HZ$. The quantiles are $Q_{.05} = 0.1546$, $Q_{.16} = 0.2068$, $Q_{.5} = 0.3311$, $Q_{.84} = 0.5527$, $Q_{.95} = 0.7808(g)$. We interpolate the expert distribution as piece-wise uniform within the quantile intervals, $\pm 10\%$ overshoot on both sides. Add one noise term a time with $\delta = [0.02, 0.04, 0.06](g)$. The results are in Figure (19) An easy way to see the noise effect is again redesign the data flow diagram as Figure

seed ID	value	full name
seed1	0.77236	acc(g) rec=1E-006/year,freq=2.5HZ
seed2	2.0852	acc(g) rec=1E-006/year,freq=4HZ
seed3	1.6659	acc(g) rec=1E-006/year,freq=5HZ
seed4	1.4669	acc(g) rec=1E-006/year,freq=6.6667HZ
seed5	0.9073	acc(g) rec=1E-006/year,freq=10HZ
seed6	0.23071	acc(g) rec=1E-007/year,freq=0.5HZ
seed7	0.80996	acc(g) rec=1E-007/year,freq=1.34HZ
seed8	0.99912	acc(g) rec=1E-007/year,freq=20HZ
seed9	1.219	acc(g) rec=1E-007/year,freq=35HZ
seed10	0.66888	acc(g) rec=1E-007/year,freq=50HZ
seed11	0.56281	acc(g) rec=1E-005/year,freq=2.5HZ
seed12	0.90828	acc(g) rec=1E-005/year,freq=4HZ
seed13	0.82379	acc(g) rec=1E-005/year,freq=5HZ
seed14	0.88561	acc(g) rec=1E-005/year,freq=6.6667HZ
seed15	0.63141	acc(g) rec=1E-005/year,freq=10HZ

	expert	Calibration	lformation	weight		Id	Calibr.	Mean relative information	UnNormaliz weight	Normaliz.w without DM
	1	0.00097647	0.41477	0		EX1	0.000131	0.4148	0	0
Matlab	2	0.0039772	0.12996	0	Excalibur	EX2	0.001593	0.13	0	0
noisy	3	0.00024994	0.96614	0	non-noisy	EX3	0.000534	0.9661	0	0
	4	0.0072291	0.77158	0		EX4	0.007225	0.7716	0	0
	5	0.40677	0.67067	1		EX5	0.2976	0.6707	0.1996	1
	DM	0.40677	0.66738	NaN		DM	0.2976	0.6707	0.1996	
	eq_DM	0.020309	0.33975	NaN		eq_DM	0.001181	0.3344	0.000395	

Table 11: noise effect, $\delta = 0.02(\text{g})$

(20). The sample source is chosen to be Groen without dependence. The seed set and scores are in Table (11). From the results, Expert 1, 2 and 5 are better calibrated if we add the noise to them before the calibration. The information scores are based on the noise-free experts, so they are identical within the trials. The DM performance is better in the noisy trial.

Now, add up the noise s.t. $\delta = 0.06(\text{g})$. Do the same trial as above. The result is in Table (12). This time, Expert 1, 4 and 5 show some improvement at their calibrations in the noisy trial. Again the DM has better performance in the noisy trial.

8.3.2 The Author's Approach

Let us recall Section (4.6.4). This approach requires a noisy set for each seed variable to interpolate the sample distribution, recall f_{Z^*} . This requirement could make the approach difficult to implement in many cases.

However in this study, the synthetic seed variables are actually sampled from given distributions as seed sources, [2][Groen]. We can treat the source distribution as f_{Z^*} and implement the filter-like algorithm, Formula (18). Using the outcome distribution f_{Y^*} as the noise-free source to re-sample the seed variable, Formula (21).

Let us first compare the f_{Z^*} and f_{Y^*} . Choose f_{Z^*} from Groen at $rec = 10^{-6}$ and $freq = 10(\text{HZ})$. Choose the noise term with $\delta = 0.02$. The densities are in Figure (21). The numerical optimization toolbox [10][Mosek] are used to implement Formula (18).

To compare the noise effect, we again redesign the data flow as Figure (22)

The noisy/noise-free seed sets and the corresponding scores are in Table (13). From the two seed variable sets, Z^* and Y^* are very close. Except Expert 3, the other experts/DM's scores are almost identical within the trials.

In the previous section, we tried some optimistic assumptions in Groen's model. The experts/DM have bad performances under the assumptions. We could treat the optimism as bias (noises with non-zero means), and implement the algorithms above. The performances could be improved.

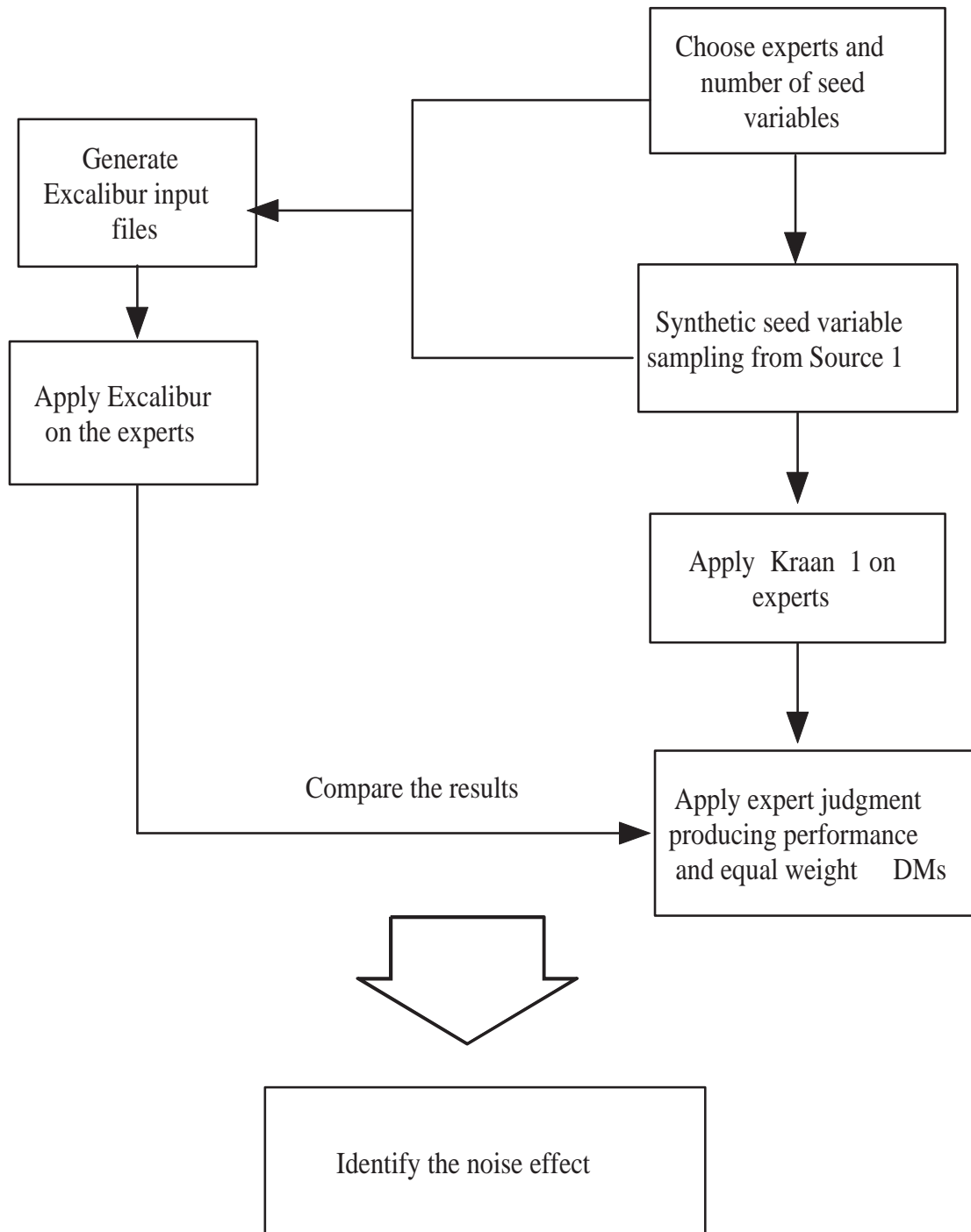


Figure 20: noise effect identification, Kraan 1

seed ID	value	full name
seed1	0.85595	acc(g) rec=1E-006/year,freq=2.5HZ
seed2	2.1554	acc(g) rec=1E-006/year,freq=4HZ
seed3	1.4837	acc(g) rec=1E-006/year,freq=5HZ
seed4	1.6704	acc(g) rec=1E-006/year,freq=6.6667HZ
seed5	0.83194	acc(g) rec=1E-006/year,freq=10HZ
seed6	0.18842	acc(g) rec=1E-007/year,freq=0.5HZ
seed7	0.35146	acc(g) rec=1E-007/year,freq=1HZ
seed8	1.2689	acc(g) rec=1E-007/year,freq=13.34HZ
seed9	1.2147	acc(g) rec=1E-007/year,freq=25HZ
seed10	0.73446	acc(g) rec=1E-007/year,freq=50HZ
seed11	0.61038	acc(g) rec=1E-005/year,freq=2.5HZ
seed12	1.0421	acc(g) rec=1E-005/year,freq=4HZ
seed13	1.0724	acc(g) rec=1E-005/year,freq=5HZ
seed14	0.79066	acc(g) rec=1E-005/year,freq=6.6667HZ
seed15	0.8033	acc(g) rec=1E-005/year,freq=10HZ

	expert	Calibration	lformation	weight		Id	Calibr.	Mean relative information	UnNormaliz weight	Normaliz.w without DM
	1	0.0022659	0.42742	0		EX1	0.000577	0.4274	0	0
Matlab	2	0.0007667	0.13114	0	Excalibur	EX2	0.000819	0.1311	0	0
noisy	3	0.002793	1.0044	0	non-noisy	EX3	0.002793	1.004	0	0
	4	0.018273	0.79302	0		EX4	0.01127	0.793	0	0
	5	0.33526	0.69547	1		EX5	0.1319	0.6955	0.09175	1
	DM	0.33526	0.69232	NaN		DM	0.1319	0.6955	0.09175	
	eq_DM	0.17562	0.35012	NaN		eq_DM	0.02653	0.3451	0.009154	

Table 12: noise effect, $\delta = 0.06(g)$

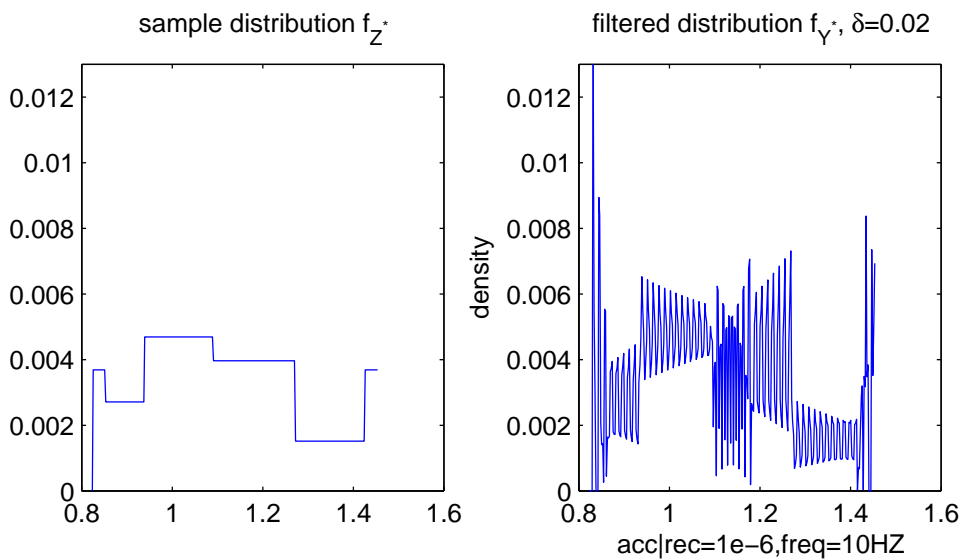


Figure 21: f_{Z^*} and f_{Y^*}

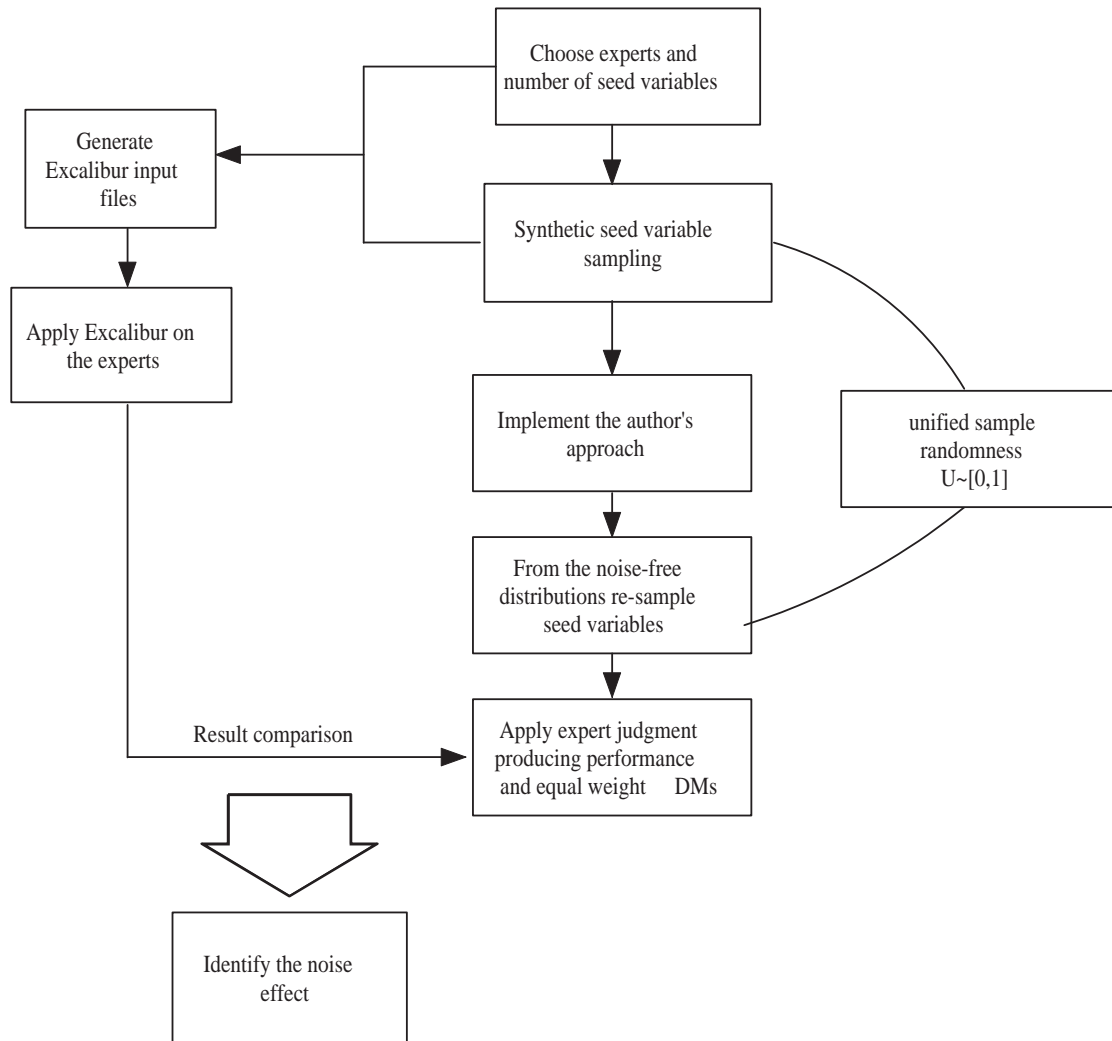


Figure 22: Noise effect identification, Meng

seed ID	Y*	Z*	full name
seed1	0.95189	0.95577	acc(g) rec=1E-006/year,freq=2.5HZ
seed2	1.8308	1.8421	acc(g) rec=1E-006/year,freq=4HZ
seed3	1.5668	1.5715	acc(g) rec=1E-006/year,freq=5HZ
seed4	1.3164	1.3197	acc(g) rec=1E-006/year,freq=6.6667HZ
seed5	1.0762	1.0778	acc(g) rec=1E-006/year,freq=10HZ
seed6	0.24888	0.24794	acc(g) rec=1E-007/year,freq=0.5HZ
seed7	0.85169	0.8579	acc(g) rec=1E-007/year,freq=1.34HZ
seed8	1.3293	1.3428	acc(g) rec=1E-007/year,freq=20HZ
seed9	1.1397	1.1379	acc(g) rec=1E-007/year,freq=35HZ
seed10	1.0517	1.0424	acc(g) rec=1E-007/year,freq=50HZ
seed11	0.53262	0.53381	acc(g) rec=1E-005/year,freq=2.5HZ
seed12	1.072	1.0685	acc(g) rec=1E-005/year,freq=4HZ
seed13	1.126	1.1211	acc(g) rec=1E-005/year,freq=5HZ
seed14	1.027	1.0301	acc(g) rec=1E-005/year,freq=6.6667HZ
seed15	0.69704	0.697	acc(g) rec=1E-005/year,freq=10HZ

	expert	Calibration	lformation	weight		Id	Calibr.	Mean relative information	UnNormaliz weight	Normaliz.w without DM
	1	0.066235	0.41477	0.26207		EX1	0.06621	0.4148	0.02746	0.262
Matlab	2	0.018634	0.12996	0	Excalibur	EX2	0.01864	0.13	0	0
Y*	3	0.00014907	0.96614	0	Z*	EX3	0.000202	0.9661	0	0
	4	0.034868	0.77158	0		EX4	0.03487	0.7716	0	0
	5	0.11534	0.67067	0.73793		EX5	0.1153	0.6707	0.07734	0.738
	DM	0.16631	0.47086	NaN		DM	0.1664	0.465	0.07736	
	eq_DM	0.049563	0.33975	NaN		eq_DM	0.05784	0.3344	0.01934	

Table 13: noise effect, $\delta = 0.02(g)$

8.4 Summary of the Matlab Code Functions

As we can see, the Matlab code has several extensions over Excalibur:

- It can read quantile curves from the ASCII data files.
- It can apply scoring rules on qauntile curves for given seed variables.
- It can sample synthetic seed variables from given distributions, when realizations are not available.
- It can produce DM quantile curves in the same format as the input experts'.
- It can handle noisy observations.

However, since it is written in Matlab, the speed is relatively slow.

9 DM Convergence

There is no theory tells how many seed variables/experts are needed in order to make the DM converge. In this study, there are 80 real experts, and we can sample the synthetic seed variables as many as we want. This fact provides us an opportunity to investigate the DM behavior w.r.t. the numbers of seed variables and experts.

9.1 Measure of Convergence

The convergence of the a DM serial is measured by the relative information:

$$I(t|s) = \text{Inf}(DM(t)|DM(s)); \quad (29)$$

where $s, t|0 < s < t$ are the numbers of seed variables/experts. If $I(t|s)$ is close to 0 for some value s , then it suggests the DM converges after s seed variables/experts.

seed ID	value	seed name		
seed1	0.645121	acc(g) rec=1E-006/year,spec=2.5HZ		
seed2	1.7831	acc(g) rec=1E-006/year,spec=4HZ		
seed3	1.3746	acc(g) rec=1E-006/year,spec=5HZ		
seed4	1.53036	acc(g) rec=1E-006/year,spec=6.6667HZ		
seed5	1.02794	acc(g) rec=1E-006/year,spec=10HZ		
seed6	0.197282	acc(g) rec=1E-007/year,spec=0.5HZ		
seed7	0.782908	acc(g) rec=1E-007/year,spec=1.34HZ		
seed8	1.60076	acc(g) rec=1E-007/year,spec=20HZ		
seed9	1.09137	acc(g) rec=1E-007/year,spec=35HZ		
seed10	0.802021	acc(g) rec=1E-007/year,spec=50HZ		
seed11	0.687735	acc(g) rec=1E-005/year,spec=2.5HZ		
seed12	0.911654	acc(g) rec=1E-005/year,spec=4HZ		
seed13	1.12569	acc(g) rec=1E-005/year,spec=5HZ		
seed14	1.03238	acc(g) rec=1E-005/year,spec=6.6667HZ		
seed15	0.679402	acc(g) rec=1E-005/year,spec=10HZ		

Expert	Calibr.	Mean relative	UnNormaliz	Normaliz.w
		information	weight	without DM
'a_SC_Studer'	0.2092	0.9241	0.1933	0.2499
'b_CO_Studer'	0.1961	0.6508	0.1276	0.165
'b_BO_Faeh '	0.009596	0.2088	0	0
'd_SC_Studer'	0.3834	0.8624	0.3306	0.4275
'd_CO_Pecker'	0.00074	0.8902	0	0
'a_BO_Bard '	0.007289	0.4179	0	0
'd_SA_Faeh '	0.05036	0.2623	0	0
'a_CO_Bard '	0.08828	0.3411	0	0
'a_BU_Studer'	0.004791	1	0	0
'd_BU_Bard '	0.2075	0.5872	0.1218	0.1575
DM(15)	0.7296	0.6342	0.4627	

Table 14: seed variables, s=15

9.2 DM w.r.t. Number of Seed Variables

Here we fix the effective number, the N value in Formula (4), of the seed variables to 15, while the real number is adding up.

Default recommendation for the number of experts is 5 to 10, see [1][Cooke]. Here we randomly choose 10 experts. The way to choose experts is

- Make an alphabetic list of the experts' names.
- Randomly permute the list.
- Pick the first 10 as out experts.

First sample 15 seed variables. The result is in Table (14)

Add the number of seed variables by 5 a time up to 50. The added seeds and $DM(50)$ is in Table (15).

Calculate $Inf(DM(15 + 5k)|DM(15 + 5i))$; where $k = 1, \dots, 7$; $i = 0, \dots, k - 1$. The results is in Table (16). After 15 or 20 seed variables, the DM is quite stable. In Figure (23), we can see the convergence of the DM w.r.t. the number of seed variables.

seed16	0.877141	acc(g) rec=1E-006/year,spec=2.5HZ
seed17	2.05037	acc(g) rec=1E-006/year,spec=4HZ
seed18	1.36971	acc(g) rec=1E-006/year,spec=5HZ
seed19	1.56811	acc(g) rec=1E-006/year,spec=6.6667HZ
seed20	1.37798	acc(g) rec=1E-006/year,spec=10HZ
seed21	0.172912	acc(g) rec=1E-007/year,spec=0.5HZ
seed22	0.458616	acc(g) rec=1E-007/year,spec=1.34HZ
seed23	1.60006	acc(g) rec=1E-007/year,spec=20HZ
seed24	0.912446	acc(g) rec=1E-007/year,spec=35HZ
seed25	0.881625	acc(g) rec=1E-007/year,spec=50HZ
seed26	0.421442	acc(g) rec=1E-005/year,spec=2.5HZ
seed27	1.18618	acc(g) rec=1E-005/year,spec=4HZ
seed28	0.937433	acc(g) rec=1E-005/year,spec=5HZ
seed29	1.14725	acc(g) rec=1E-005/year,spec=6.6667HZ
seed30	0.645828	acc(g) rec=1E-005/year,spec=10HZ
seed31	0.92682	acc(g) rec=1E-006/year,spec=2.5HZ
seed32	1.75332	acc(g) rec=1E-006/year,spec=4HZ
seed33	1.47735	acc(g) rec=1E-006/year,spec=5HZ
seed34	1.4951	acc(g) rec=1E-006/year,spec=6.6667HZ
seed35	1.02629	acc(g) rec=1E-006/year,spec=10HZ
seed36	0.207072	acc(g) rec=1E-007/year,spec=0.5HZ
seed37	0.803118	acc(g) rec=1E-007/year,spec=1.34HZ
seed38	1.35579	acc(g) rec=1E-007/year,spec=20HZ
seed39	1.26262	acc(g) rec=1E-007/year,spec=35HZ
seed40	0.596422	acc(g) rec=1E-007/year,spec=50HZ
seed41	0.536919	acc(g) rec=1E-005/year,spec=2.5HZ
seed42	1.15679	acc(g) rec=1E-005/year,spec=4HZ
seed43	0.916238	acc(g) rec=1E-005/year,spec=5HZ
seed44	0.872655	acc(g) rec=1E-005/year,spec=6.6667HZ
seed45	0.749826	acc(g) rec=1E-005/year,spec=10HZ
seed46	0.859441	acc(g) rec=1E-006/year,spec=2.5HZ
seed47	1.6384	acc(g) rec=1E-006/year,spec=4HZ
seed48	1.38028	acc(g) rec=1E-006/year,spec=5HZ
seed49	1.22838	acc(g) rec=1E-006/year,spec=6.6667HZ
seed50	0.859277	acc(g) rec=1E-006/year,spec=10HZ

Expert	Calibr.	Mean relative information	UnNormaliz weight	Normaliz.w without DM
'a_SC_Studer'	0.5975	0.9255	0.553	1
'b_CO_Studer'	0.2964	0.6651	0	0
'b_BO_Faeh '	0.004447	0.202	0	0
'd_SC_Studer'	0.1376	0.8646	0	0
'd_CO_Pecker'	0.003152	0.9221	0	0
'a_BO_Bard '	0.004939	0.4116	0	0
'd_SA_Faeh '	0.07366	0.253	0	0
'a_CO_Bard '	0.09501	0.3419	0	0
'a_BU_Studer'	0.1076	1.002	0	0
'd_BU_Bard '	0.1876	0.5825	0	0
DM(50)	0.5975	0.9255	0.553	

Table 15: Scores for 50 seed variables

Inf	s=15	s=20	s=25	s=30	s=35	s=40	s=45
t=15	0.05034						
t=25	0.1726	0.1546					
t=30	0.06242	0.06996	0.1634				
t=35	0.1726	0.1546	0	0.1221			
t=40	0.1726	0.1546	0	0.1221	0		
t=45	0.1726	0.1546	0	0.1221	0	0	
t=50	0.1726	0.1546	0	0.1221	0	0	0

Table 16: $Inf(DM(t)|DM(s))$, where $t = 15 + 5k$ and $s = 15 + 5i$ are the numbers of seed variables; $k = 1, \dots, 7$; $i = 0, \dots, k - 1$.

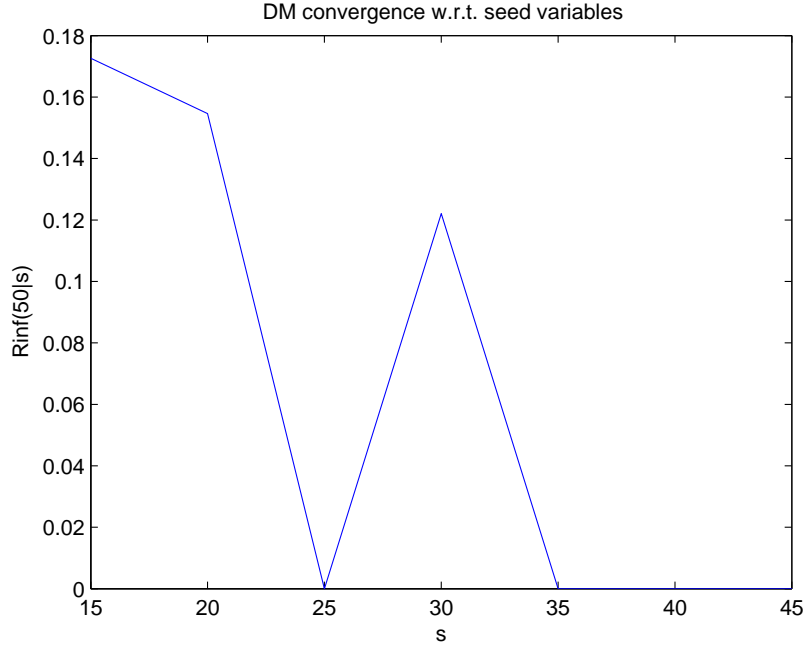


Figure 23: relative information of DM(50) w.r.t. DM(s)

9.3 DM w.r.t. Number of Seed Experts

9.3.1 Fix the Intrinsic Range

The quantile elicitations from the experts are far different. Thus, add up the number of experts, we also add up the intrinsic ranges mostly. The change of the intrinsic ranges may ultimately affect the DM. To investigate the DM convergence we need first to fix the intrinsic ranges. The method is to make up a badly calibrated expert, denoted as EX81, which has biggest quantile range $[Q_{0.05}, Q_{0.95}]$ for each seed variable. Make sure it has 0 weight. The only function of this EX81 is to make the intrinsic ranges unchanged when we add/remove experts.

Set the initial number of the experts as 10 aside EX81. Add the number up to 80 by 10 a time. Sample 35 seed variables. The scoring result see Table (17-19).

From Table(18,19), the information scores for the first 10 experts hold in both tables. This suggests that the intrinsic ranges are fixed by EX81. The weight plot of the DM(80) corresponding to Table (19) is shown in Figure(24).

Calculate $Inf(DM(10 + 10k)|DM(10 + 10i))$; where $k = 1, \dots, 7$; $i = 0, \dots, k - 1$. The results are in Table (20). From the table, the $Inf(DM(80)|DM(s))$ is step wise decreasing as $s \rightarrow 80$. This shows the sign of convergence. $Inf(DM(80)|DM(s)) < 0.007$ for $s \geq 50$, which suggests that the $DM(50)$ and $DM(80)$ are very close. Note, the expert list is randomly permuted to avoid the effect of the order. In general, after taking 10 experts, the DM is very stable. We can see the convergence behavior of the DM in Figure (25)

ID	Value	full name
seed1	0.647387	acc(g) rec=1E-006/year,spec=2.5HZ
seed2	1.63751	acc(g) rec=1E-006/year,spec=4HZ
seed3	1.46133	acc(g) rec=1E-006/year,spec=5HZ
seed4	1.29187	acc(g) rec=1E-006/year,spec=6.6667HZ
seed5	0.893568	acc(g) rec=1E-006/year,spec=10HZ
seed6	0.137586	acc(g) rec=1E-007/year,spec=0.5HZ
seed7	0.565418	acc(g) rec=1E-007/year,spec=1.34HZ
seed8	1.09186	acc(g) rec=1E-007/year,spec=20HZ
seed9	0.755491	acc(g) rec=1E-007/year,spec=35HZ
seed10	1.04162	acc(g) rec=1E-007/year,spec=50HZ
seed11	0.505681	acc(g) rec=1E-005/year,spec=2.5HZ
seed12	1.05407	acc(g) rec=1E-005/year,spec=4HZ
seed13	1.09546	acc(g) rec=1E-005/year,spec=5HZ
seed14	0.913722	acc(g) rec=1E-005/year,spec=6.6667HZ
seed15	0.861769	acc(g) rec=1E-005/year,spec=10HZ
seed16	0.628619	acc(g) rec=1E-006/year,spec=2.5HZ
seed17	1.82053	acc(g) rec=1E-006/year,spec=4HZ
seed18	1.53349	acc(g) rec=1E-006/year,spec=5HZ
seed19	1.29987	acc(g) rec=1E-006/year,spec=6.6667HZ
seed20	1.09381	acc(g) rec=1E-006/year,spec=10HZ
seed21	0.128415	acc(g) rec=1E-007/year,spec=0.5HZ
seed22	0.639504	acc(g) rec=1E-007/year,spec=1.34HZ
seed23	1.31266	acc(g) rec=1E-007/year,spec=20HZ
seed24	1.16032	acc(g) rec=1E-007/year,spec=35HZ
seed25	0.932957	acc(g) rec=1E-007/year,spec=50HZ
seed26	0.485637	acc(g) rec=1E-005/year,spec=2.5HZ
seed27	0.871146	acc(g) rec=1E-005/year,spec=4HZ
seed28	0.974502	acc(g) rec=1E-005/year,spec=5HZ
seed29	0.870007	acc(g) rec=1E-005/year,spec=6.6667HZ
seed30	0.797763	acc(g) rec=1E-005/year,spec=10HZ
seed31	0.861888	acc(g) rec=1E-006/year,spec=2.5HZ
seed32	2.06342	acc(g) rec=1E-006/year,spec=4HZ
seed33	1.41704	acc(g) rec=1E-006/year,spec=5HZ
seed34	1.73283	acc(g) rec=1E-006/year,spec=6.6667HZ
seed35	1.32139	acc(g) rec=1E-006/year,spec=10HZ

Table 17: 35 seed variables

Ex name	Calibr.	Mean relative information	UnNormaliz weight	Normaliz.w without DM
'd_SC_Bard '	5.8E-10	0.7833	0	0
'c_SA_Bard '	1.1E-05	0.7561	0	0
'c_BO_Faeh '	4.8E-08	0.5492	0	0
'a_CO_Pecker'	8.7E-07	1.268	0	0
'a_SA_Bard '	0.00044	0.822	0	0
'c_BU_Studer'	0.1698	1.29	0.219	1
'c_BU_Bard '	9.2E-06	0.9101	0	0
'a_BU_Studer'	0.00052	1.38	0	0
'c_BO_Studer'	6.2E-05	1.179	0	0
'c_BO_Bard '	4.1E-10	0.7177	0	0
Ex81	2.3E-05	0.6122	0	0
DM(10)	0.1698	1.29	0.219	
eq_DM(10)	3.9E-05	0.6222	2.397E-05	

Table 18: Scores and weights for the first 10 (plus 1) experts and DM

Ex name	Calibr.	Mean relative information	UnNormaliz weight	Normaliz.w without DM	Ex name	Calibr.	Mean relative information	UnNormaliz weight	Normaliz.w without DM
'd_SC_Bard '	5.79E-10	0.7834	0	0	'a_BO_Faeh '	1.12E-08	0.6139	0	0
'c_SA_Bard '	1.11E-05	0.7562	0	0	'd_SC_Pecker'	9.22E-15	1.557	0	0
'c_BO_Faeh '	4.78E-08	0.5493	0	0	'a_SA_Faeh '	0.000371	0.6822	0	0
'a_CO_Pecker'	8.73E-07	1.269	0	0	'a_SC_Studer'	0.3097	1.309	0.4053	0.4218
'a_SA_Bard '	0.000442	0.8221	0	0	'b_BU_Studer'	0.06126	1.306	0	0
'c_BU_Studer'	0.1698	1.29	0.2191	0.228	'c_CO_Faeh '	9.33E-06	0.4636	0	0
'c_BU_Bard '	9.21E-06	0.9102	0	0	'a_CO_Bard '	0.000103	0.6879	0	0
'a_BU_Studer'	0.000523	1.38	0	0	'd_BO_Pecker'	1.14E-15	1.566	0	0
'c_BO_Studer'	6.23E-05	1.18	0	0	'c_SC_Bard '	5.53E-11	0.754	0	0
'c_BO_Bard '	4.15E-10	0.7178	0	0	'c_SC_Pecker'	3.46E-15	1.596	0	0
'b_BU_Pecker'	2.4E-06	1.523	0	0	'd_SA_Pecker'	1.77E-07	1.345	0	0
'a_SC_Pecker'	0.000107	1.522	0	0	'b_SA_Faeh '	4.47E-05	0.5769	0	0
'b_BU_Faeh '	8.66E-05	0.7727	0	0	'd_CO_Bard '	0.000117	0.6528	0	0
'd_BU_Faeh '	0.001411	0.7751	0	0	'b_SC_Bard '	5.79E-10	0.7854	0	0
'c_SC_Faeh '	1.35E-09	0.599	0	0	'd_BO_Bard '	3.7E-11	0.7197	0	0
'd_SA_Studer'	0.04566	1.125	0	0	'a_BO_Studer'	0.000411	1.207	0	0
'c_BU_Pecker'	7.06E-10	1.548	0	0	'b_SC_Studer'	0.000708	1.247	0	0
'a_BO_Bard '	3.13E-09	0.7699	0	0	'd_CO_Faeh '	9.33E-06	0.4628	0	0
'd_BU_Studer'	0.1778	1.29	0.2294	0.2388	'a_BU_Bard '	0.009042	1.092	0	0
'a_BU_Pecker'	0.03771	1.418	0	0	'b_SC_Pecker'	9.22E-15	1.567	0	0
'a_SA_Pecker'	1.59E-06	1.31	0	0	'a_SA_Studer'	0.02481	1.157	0	0
'd_SC_Studer'	0.000424	1.25	0	0	'b_CO_Bard '	0.000117	0.6409	0	0
'b_SC_Faeh '	1.14E-09	0.6154	0	0	'c_CO_Bard '	0.000277	0.6517	0	0
'b_SA_Pecker'	4.81E-08	1.372	0	0	'd_CO_Pecker'	5.19E-09	1.308	0	0
'a_SC_Bard '	0.000606	0.9396	0	0	'd_SC_Faeh '	1.4E-08	0.6202	0	0
'c_SA_Studer'	0.004997	1.131	0	0	'd_BU_Bard '	0.002821	0.9553	0	0
'c_BO_Pecker'	1.14E-15	1.56	0	0	'd_BU_Pecker'	2.35E-05	1.457	0	0
'c_CO_Pecker'	1.05E-09	1.324	0	0	'c_BU_Faeh '	1.43E-05	0.7254	0	0
'b_CO_Studer'	0.05503	1.044	0	0	'd_BO_Faeh '	1.19E-07	0.5586	0	0
'c_SC_Studer'	9.6E-05	1.261	0	0	'c_SA_Pecker'	1.37E-08	1.406	0	0
'b_BO_Pecker'	1.14E-15	1.569	0	0	'b_SA_Bard '	9.02E-05	0.7646	0	0
'b_CO_Pecker'	4.12E-09	1.337	0	0	'a_CO_Studer'	0.07859	1.059	0	0
'c_CO_Studer'	0.043	1.041	0	0	'c_SA_Faeh '	2.2E-05	0.5673	0	0
'd_CO_Studer'	0.1035	1.033	0.107	0.1114	'b_BO_Bard '	3.7E-11	0.7012	0	0
'b_CO_Faeh '	1.24E-06	0.4531	0	0	'a_SC_Faeh '	0.000175	0.7834	0	0
'a_BO_Pecker'	1.14E-15	1.542	0	0	'b_BO_Faeh '	1.67E-08	0.5456	0	0
'b_BO_Studer'	9.12E-06	1.174	0	0	'a_BU_Faeh '	3.03E-05	0.9598	0	0
'd_SA_Bard '	0.00061	0.7659	0	0	'a_CO_Faeh '	2.43E-05	0.5377	0	0
'b_BU_Bard '	0.000211	0.9504	0	0	'b_SA_Studer'	0.01883	1.135	0	0
'd_SA_Faeh '	6.04E-05	0.5986	0	0	'd_BO_Studer'	9.12E-06	1.176	0	0
					EX 81	2.35E-05	0.6123	0	0
					DM(80)	0.3725	1.165	0.4341	
					eq_DM(80)	2E-05	0.6716	1.341E-05	

Table 19: Scores and weights for 80 (plus 1) experts and DM

Inf	s=10	s=20	s=30	s=40	s=50	s=60	s=70
t=20	0.127						
t=30	0.127	4.99E-07					
t=40	0.127	4.78E-07	3.06E-07				
t=50	0.1855	0.09457	0.09457	0.09457			
t=60	0.1855	0.09457	0.09457	0.09457	7.30E-07		
t=70	0.1855	0.09457	0.09457	0.09457	6.68E-07	6.45E-07	
t=80	0.1872	0.0985	0.0985	0.0985	0.006287	0.006287	0.006287

Table 20: $Inf(DM(t)|DM(s))$; where $t = 10 + 10k$, $s = 10 + 10i$ are the numbers of experts; $k = 1, \dots, 7$; $i = 0, \dots, k - 1$.

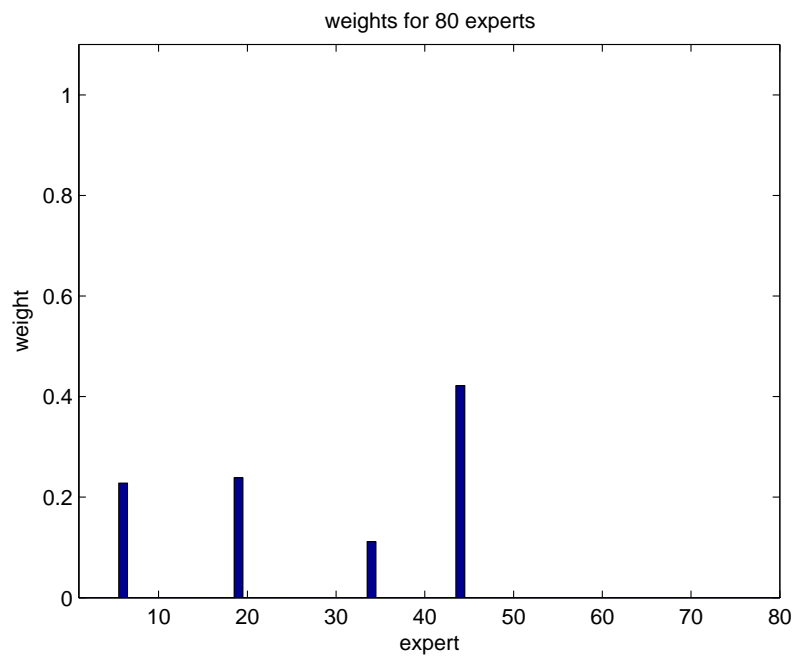


Figure 24: weight plot of DM(80)

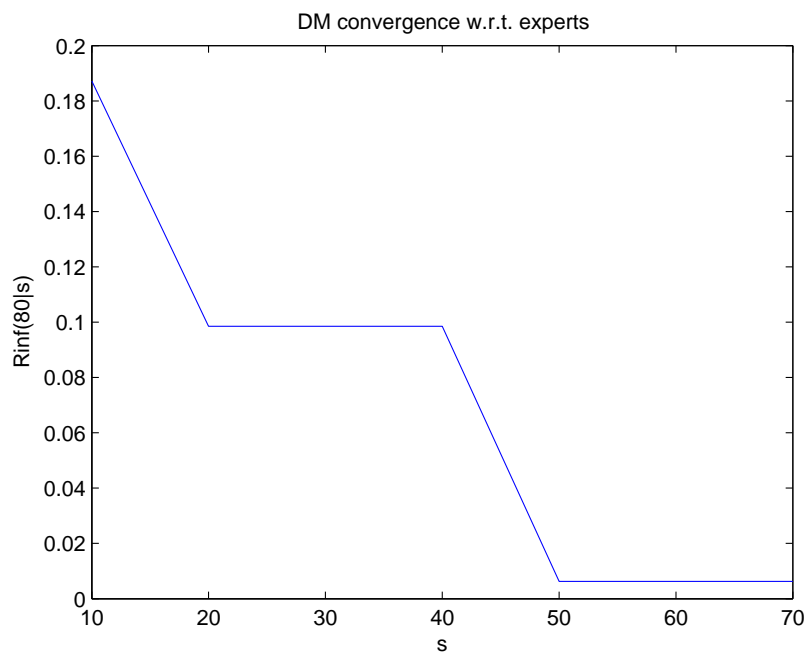


Figure 25: relative information of DM(80) w.r.t DM(s)

10 Conclusion

10.1 80 Experts' Performances

- From the trials, the performances of the 80 experts are significantly different, see Table (19).
- Experts' calibrations scores are very sensitive to different set of (synthetic) seed variables.
- These facts suggest that the methods to aggregate experts are important to the DM performance.

10.2 Different Decision Makers

- The performance DM and equal weight DM are compared. In many trials, they are significantly different.
- In some trials, the equal weight DM is well calibrated while in other trials it is not.
- The performance DM has good performance in most of the trials.
- In all the trials, the performance DM is more informative than equal weight DM.

10.3 Impact of the Optimistic Dependence Assumptions and Noise term

- Synthetic seed variables sampled from Groen plus optimistic dependence are lower than those sampled from original Groen.
- All the experts have bad performances based on the optimistic seed variables which means they are not optimistic.
- The samples from the two dependence structures, Vine and Joint Normal, are close.
- Taking the noise term into consideration could have positive effect on the experts/DM performances.

10.4 Performance DM Convergence

- Till now, there is no theory to prove the convergence of the DM.
- The trial results suggest that the DM is converging w.r.t. the growing numbers of seed variables.
- Convergence behavior of the DM can be seen as we add the number of experts up to 80.
- After we have 15 seed variables and 10 experts, the performance DM is fairly stable.

	Id	Calibr.	Mean relative information	UnNormaliz weight	Normaliz.w without DM		Id	Calibr.	Mean relative information	UnNormaliz weight	Normaliz.w without DM
	EX1	0.00043	0.4267	0	0		EX1	0.005374	0.4267	0	0
Excalibur trial1	EX2	0.00062	0.132	0	0	Excalibur trial2	EX2	0.00017	0.132	0	0
	EX3	0.00133	0.9831	0	0		EX3	5.23E-05	0.9831	0	0
	EX4	0.09965	0.7798	0.0777	0.3123		EX4	0.02264	0.7798	0	0
	EX5	0.2504	0.6835	0.1711	0.6877		EX5	0.1971	0.6835	0.1347	1
	DM	0.492	0.6451	0.3174			DM	0.1971	0.6835	0.1347	
	eq_DM	0.1757	0.3396	0.05967			eq_DM	0.06398	0.3396	0.02173	
	Id	Calibr.	Mean relative information	UnNormaliz weight	Normaliz.w without DM		Id	Calibr.	Mean relative information	UnNormaliz weight	Normaliz.w without DM
	EX1	0.00222	0.4268	0	0		EX1	0.006694	0.4267	0	0
Excalibur trial3	EX2	0.00082	0.1321	0	0	Excalibur trial4	EX2	0.003978	0.132	0	0
	EX3	0.00279	0.9832	0.002746	0.2266		EX3	8.58E-05	0.9831	0	0
	EX4	0.00503	0.7799	0.00392	0.3235		EX4	0.03487	0.7798	0	0
	EX5	0.00797	0.6836	0.005451	0.4498		EX5	0.1868	0.6835	0.1277	1
	DM	0.4779	0.6107	0.2918			DM	0.1868	0.6835	0.1277	
	eq_DM	0.1757	0.3397	0.05969			eq_DM	0.03046	0.3396	0.01035	
	Id	Calibr.	Mean relative information	UnNormaliz weight	Normaliz.w without DM		Id	Calibr.	Mean relative information	UnNormaliz weight	Normaliz.w without DM
	EX1	0.00375	0.4267	0.001601	0.09032		EX1	0.00084	0.4267	0	0
Excalibur trial5	EX2	0.00159	0.132	0	0	Excalibur trial6	EX2	0.000819	0.132	0	0
	EX3	4.2E-05	0.9831	0	0		EX3	3.89E-06	0.9831	0	0
	EX4	0.00318	0.7798	0.002477	0.1397		EX4	0.00194	0.7798	0	0
	EX5	0.01997	0.6835	0.01365	0.7699		EX5	0.3664	0.6835	0.2504	1
	DM	0.04926	0.5613	0.02765			DM	0.3664	0.6835	0.2504	
	eq_DM	0.02031	0.3396	0.006898			eq_DM	0.04909	0.3397	0.01667	

Table 21: score table for the 6 trials from Excalibur

11 Appendix A

11.1 Figure (26 to 31) are rang graphs corresponding to Figure (15)

11.2 Table (21) is the Excalibur counterpart of Table (4) from Excalibur

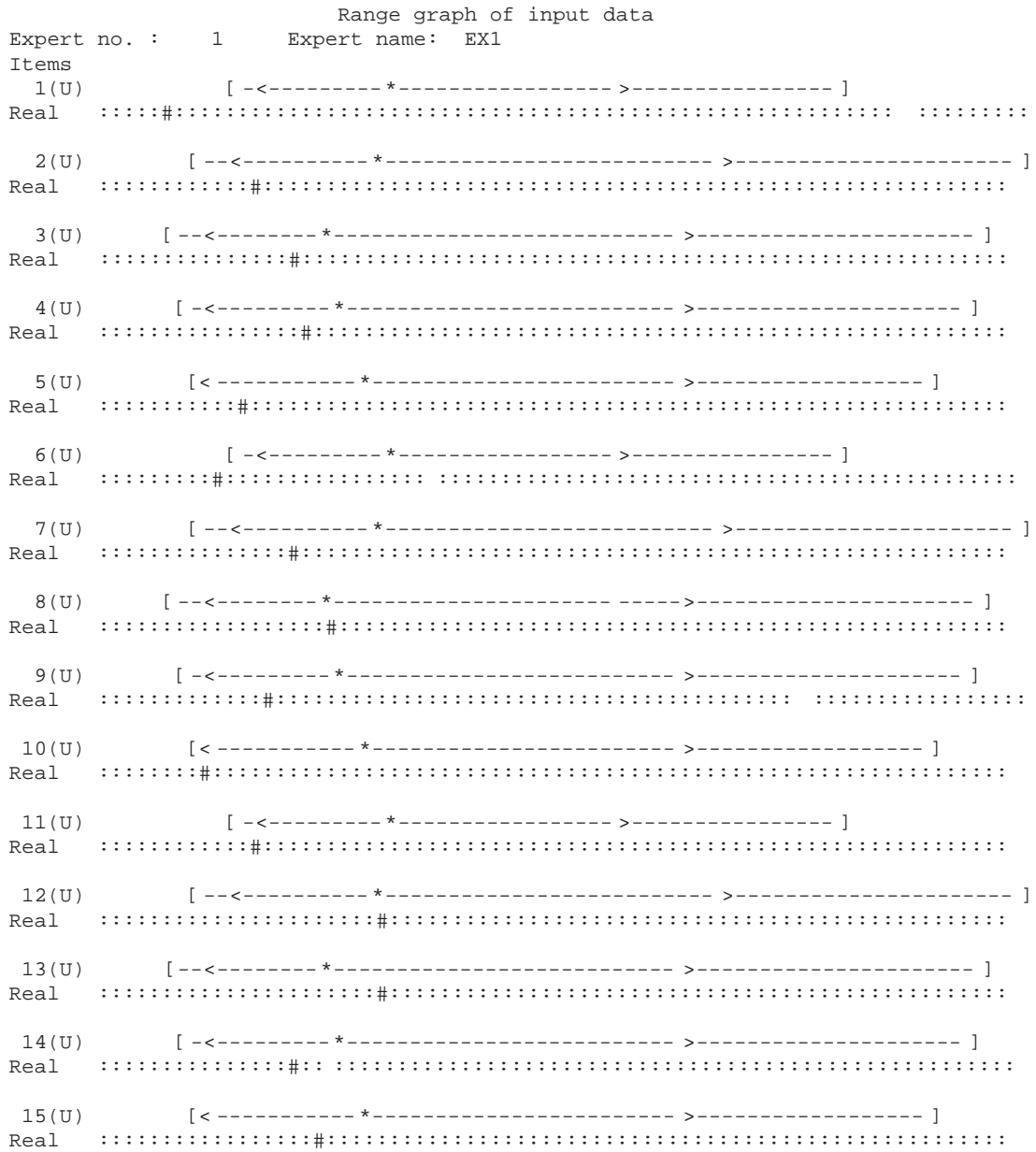


Figure 26: uniform background range graph for EX1

```

Expert no. :    2      Expert name:  EX2
Items
  1(U)  [ -----<-----*----->----- ]
Real   ::::#::::::
      2(U) [ -----<-----*----->----- ]
Real   ::::#::::::
      3(U) [ -----<-----*----->----- ]
Real   ::::#::::::
      4(U) [ -----<-----*----->----- ]
Real   ::::#::::::
      5(U) [ -----<-----*----->----- ]
Real   ::::#::::::
      6(U)  [ -----<-----*----->----- ]
Real   ::::#::::::
      7(U) [ -----<-----*----->----- ]
Real   ::::#::::::
      8(U) [ -----<-----*----->----- ]
Real   ::::#::::::
      9(U) [ -----<-----*----->----- ]
Real   ::::#::::::
     10(U) [ -----<-----*----->----- ]
Real   ::::#::::::
     11(U)  [ -----<-----*----->----- ]
Real   ::::#::::::
     12(U) [ -----<-----*----->----- ]
Real   ::::#::::::
     13(U) [ -----<-----*----->----- ]
Real   ::::#::::::
     14(U) [ -----<-----*----->----- ]
Real   ::::#::::::
     15(U) [ -----<-----*----->----- ]
Real   ::::#::::::

```

Figure 27: uniform background range graph for EX2

```

Expert no. : 3      Expert name: EX3
Items
  1(U)      [ -----<-----*----->----- ]
Real  ::::: #: :::::
  2(U)      [ ---<---*--->--- ]
Real  ::::: #: :::::
  3(U)      [ --<--*-->-- ]
Real  ::::: #: :::::
  4(U)      [ --<--*-->-- ]
Real  ::::: #: :::::
  5(U)      [ -<--*-->- ]
Real  ::::: #: :::::
  6(U)      [ -----<-----*----->----- ]
Real  ::::: #: :::::
  7(U)      [ ---<---*--->--- ]
Real  ::::: #: :::::
  8(U)      [ --<--*-->-- ]
Real  ::::: #: :::::
  9(U)      [ --<--*-->-- ]
Real  ::::: #: :::::
  10(U)     [ -<--*-->- ]
Real  ::::: #: :::::
  11(U)     [ -----<-----*----->----- ]
Real  ::::: #: :::::
  12(U)     [ ---<---*--->--- ]
Real  ::::: #: :::::
  13(U)     [ --<--*-->-- ]
Real  ::::: #: :::::
  14(U)     [ --<--*-->-- ]
Real  ::::: #: :::::
  15(U)     [ -<--*-->- ]
Real  ::::: #: :::::

```

Figure 28: uniform background range graph for EX3

```

Expert no. :    4      Expert name:  EX4
Items
  1(U)          [ --<-----*----->----- ]
Real  :::::###:::::
      2(U)      [ -<-----*----->----- ]
Real  :::::###:::::
      3(U)      [ <-----*----->----- ]
Real  :::::###:::::
      4(U)      [ <-----*----->----- ]
Real  :::::###:::::
      5(U)      [ <-----*----->----- ]
Real  :::::###:::::
      6(U)          [ --<-----*----->----- ]
Real  :::::###:::::
      7(U)      [ -<-----*----->----- ]
Real  :::::###:::::
      8(U)      [ <-----*----->----- ]
Real  :::::###:::::
      9(U)      [ <-----*----->----- ]
Real  :::::###:::::
     10(U)      [ <-----*----->----- ]
Real  :::::###:::::
     11(U)          [ --<-----*----->----- ]
Real  :::::###:::::
     12(U)      [ -<-----*----->----- ]
Real  :::::###:::::
     13(U)      [ <-----*----->----- ]
Real  :::::###:::::
     14(U)      [ <-----*----->----- ]
Real  :::::###:::::
     15(U)      [ <-----*----->----- ]
Real  :::::###:::::

```

Figure 29: uniform background range graph for EX4


```

Expert no. :    5      Expert name:  EX5
Items
  1(U) [ -----<-----*----->----- ]
Real   :::::#::::::
      2(U) [ ---<-----*----->----- ]
Real   :::::#::::::
      3(U) [ --<-----*----->----- ]
Real   :::::#::::::
      4(U) [ --<-----*----->----- ]
Real   :::::#::::::
      5(U) [ --<-----*----->----- ]
Real   :::::#::::::
      6(U) [ -----<-----*----->----- ]
Real   :::::#::::::
      7(U) [ ---<-----*----->----- ]
Real   :::::#::::::
      8(U) [ --<-----*----->----- ]
Real   :::::#::::::
      9(U) [ --<-----*----->----- ]
Real   :::::#::::::
     10(U) [ --<-----*----->----- ]
Real   :::::#::::::
     11(U) [ -----<-----*----->----- ]
Real   :::::#::::::
     12(U) [ ---<-----*----->----- ]
Real   :::::#::::::
     13(U) [ --<-----*----->----- ]
Real   :::::#::::::
     14(U) [ --<-----*----->----- ]
Real   :::::#::::::
     15(U) [ --<-----*----->----- ]
Real   :::::#::::::

```

Figure 30: uniform background range graph for EX5

```

Expert no. : DM      Expert name: DM
Items
  1(U)      [=====<=====*=====>=====]
Real       :::::#:::::
:::::

  2(U)      [=====<=====*=====>===== ]
Real       :::::#:::::
:::::

  3(U)      [=====<=====*=====>===== ]
Real       :::::#:::::
:::::

  4(U)      [=====<=====*=====>===== ]
Real       :::::#:::::
:::::

  5(U)      [=====<=====*=====>===== ]
Real       :::::#:::::
:::::

  6(U)      [=====<=====*=====>===== ]
Real       :::::#:::::
:::::

  7(U)      [=====<=====*=====>===== ]
Real       :::::#:::::
:::::

  8(U)      [=====<=====*=====>===== ]
Real       :::::#:::::
:::::

  9(U)      [=====<=====*=====>===== ]
Real       :::::#:::::
:::::

  10(U)     [=====<=====*=====>===== ]
Real       :::::#:::::
:::::

  11(U)     [=====<=====*=====>===== ]
Real       :::::#:::::
:::::

  12(U)     [=====<=====*=====>===== ]
Real       :::::#:::::
:::::

  13(U)     [=====<=====*=====>===== ]
Real       :::::#:::::
:::::

  14(U)     [=====<=====*=====>===== ]
Real       :::::#:::::
:::::

```

Figure 31: uniform background range graph for DM

12 Appendix B

Matlab Code and Description

seismic_case :

Chooses *rec* and *freq* values

seismic_data_load.m :

Loads data from ASCII files.

C_score.m :

Returns the calibration scores of experts.

I_score.m :

Returns the information scores of experts.

no_eq_DM.m :

Multiple events decision making.

kraan1.m :

Implements Kraan 1 on experts' elicitations.

seed_filter :

Filters the noise out of the seed variable set.

Other Matlab files to implement Groen with/without the dependence structures.

References

- [1] Roger M. Cooke, 1991, Experts in Uncertainty.
- [2] Frank Groen, May 2005, Assessment of Distribution Shape and Distribution parameters for Key Ground Motion Parameters.
- [3] Kurowicka and Cooke, Uncertainty Analysis Mathematical Foundation and Application.
- [4] Wettingen, 31 July 2004, Probabilistic Seismic Hazard Analysis for Swiss Nuclear Power Plant Sites, Volume 1-6.
- [5] Bernd Kraan, 2002, Probabilistic Inversion in Uncertainty Analysis.
- [6] R.J. Budnitz, 1997, Recommendations for Probabilistic Seismic Hazard Analysis.
- [7] Dateinamen der ASCII Dateien mit Soil-Hazard Resultaten setzen sich wie folgt zusammen.
- [8] Chunfang Meng, 2005, Implementation of Dependence Structure in Monte Carlo Simulation.
- [9] R. Cooke and V. Kitchallo, 2000, Excalibur v1.0 pro.
- [10] Mosek, www.mosek.com, Matlab toolbox for optimization.