**SOM B. Performance Measures and Combination: The classical model**

There are two generic, quantitative measures of expert performance, *calibration* and *information*. Loosely, calibration measures the statistical likelihood that a set of experimental results correspond, in a statistical sense, with an expert’s assessments. More precisely, under the Classical Model, calibration is scored as the p-value at which we would falsely reject the hypothesis that an expert's probability statements were accurate. In this study the 5, 50 and 95 percentiles, or quantiles, were elicited from each expert for each of the continuous variables. Hence, each expert divides the range of possible outcomes of each variable into 4 intervals: less than or equal to the 5% value, greater than the 5% value and less than or equal to the 50% value, etc. The probabilities for these intervals are expressed as a vector

*p=* *(p1,p2, p3, p4)*  = *(0.05, 0.45, 0.45, 0.05).*

*Calibration*

If N quantities are assessed, each expert may be regarded as a statistical hypothesis, namely that each realization falls in one of the four inter-quantile intervals with probability vector *p*. Suppose we have realizations *x1,…xN* of these quantities. We may then form the sample distribution of the expert's inter quantile intervals as:

*s1(e) = #{ i | xi ≤ 5% quantile}/N*

*s2(e) = #{ i | 5% quantile < xi ≤ 50% quantile}/N*

*s3(e) = #{ i | 50% quantile < xi ≤ 95% quantile}/N*

*s4(e) = #{ i | 95% quantile < xi }/N*

*s(e) = (s1,…s4)*

Note that the sample distribution depends on the expert *e*. If the realizations are indeed drawn independently from a distribution with quantiles as stated by the expert then the quantity

*2NI(s(e) | p) = 2N ∑i=1..4 s­i­ ln(si / pi) (1)*

is asymptotically distributed as a chi-square variable with 3 degrees of freedom. This is the likelihood ratio statistic, and I(s | p) is the relative information of distribution *s* with respect to *p*. Extracting the leading term of the logarithm yields the familiar chi-square test statistic for goodness of fit.

If after a few realizations the expert were to see that all realization fell outside his 90% central confidence intervals, (s)he might conclude that these intervals were too narrow and might broaden them on subsequent assessments. This means that for this expert the uncertainty distributions are *not* independent, and (s)he learns from the realizations. Expert learning is not a goal of an expert judgment study. Rather, the problem owner wants experts who do not need to learn from the elicitation. Independence is not an assumption about the expert's distribution but a desideratum of the problem owner. Hence the decision maker (see below) scores expert *e* as the statistical likelihood of the hypothesis

*He: "the inter quantile interval containing the true value for each variable is drawn independently from probability vector p."*

A simple test for this hypothesis uses the test statistic (1), and the likelihood, or p-value, or **calibration score** of this hypothesis, is:

*Cal(e)* = *p-value(e) = Prob{2NI(s(e) p)≥ r | He}*

where *r* is the value of (1) based on the observed values *x1,…xN*. It is the probability under hypothesis He that a deviation at least as great as *r* should be observed on *N* realizations if *He* were true. Calibration scores are absolute and can be compared across studies. However it is appropriate to compare different hypothesis tests (across different studies, for example) by equalizing the effective number of realizations. To compare scores on two data sets with *N* and *N’ realizations*, we simply use the minimum of *N* and *N'* in (1), without changing the sample distribution *s.*

Although the calibration score uses the language of simple hypothesis testing, it must be emphasized that we are not rejecting expert-hypotheses; rather we are using this language to measure the degree to which the data supports the hypothesis that the expert's probabilities are accurate. Low scores, near zero, mean that it is unlikely that the expert’s probabilities are correct. High scores, near 1, indicate good support.

*Information*

The second scoring variable is information. Loosely, the information in a distribution is the degree to which the distribution is concentrated. Information cannot be measured absolutely, but only with respect to a background measure. Being concentrated or "spread out" is measured relative to some other distribution. Commonly, the uniform and log-uniform background measures are used.

Measuring information requires associating a density with each assessment of each expert. To do this, we use the unique density that complies with the experts' quantiles and is minimally informative with respect to the background measure. This density can easily be found with the method of Lagrange multipliers. For a uniform background measure, the density is constant between the assessed quantiles. The background measure is not elicited from experts as indeed it must be the same for all experts; instead it is chosen by the analyst.

The uniform and log-uniform background measures require an *intrinsic range* on which these measures are concentrated. The classical model implements the so-called *k*% overshoot rule: for each item we consider the smallest interval *I = [L, U]* containing all the assessed quantiles of all experts and the realization, if known. This interval is extended to

*I\* = [L\*, U\*]; L\* = L – k(U-L)/100; U\* = U + k(U-L)/100.*

The value of k is chosen by the analyst. A large value of *k* tends to make all experts look quite informative, and tends to suppress the relative differences in information scores. *The* **information score** of expert *e* on assessments for uncertain quantities 1…N is

*Inf (e) =Average Relative information w.r.t. Background = (1/N) ∑i = 1..N I(fe,i | gi)*

where *gi* is the background density for variable *i* and *fe,i* is expert *e*'s density for item i. This is proportional to the relative information of the expert's joint distribution given the background, under the assumption that the variables are independent. As with calibration, the assumption of independence here reflects a desideratum of the decision maker and not an elicited feature of the expert's joint distribution. The information score does not depend on the realizations. An expert can give her/himself a high information score by choosing his quantiles very close together. The information score of *e* depends on the intrinsic range and on the assessments of the other experts. Hence, information scores cannot be compared across studies.

The above information score is chosen because it is familiar, tail insensitive, scale invariant and "slow". The latter property means that relative information is a slow function; large changes in the expert assessments produce only modest changes in the information score. This contrasts with the likelihood function in the calibration score, which is a very "fast" function. This causes the product of calibration and information to be driven by the calibration score.

*Combination: Decision Maker*

The **combined score** of expert *e* will serve as an (unnormalized) weight for *e*:

*wα(e) = Cal (e) × Inf (e) × 𝟙α(Cal(e) ≥ α), (2)*

where *𝟙α(Cal(e)α) = 1* if *Cal(e) ≥ α,* and is zero otherwise. The combined score thus depends on α; if *Cal(e)*  falls below cut-off level α, expert *e* is unweighted. The presence of a cut-off level is imposed by the requirement that the combined score be an asymptotically strictly proper scoring rule. That is, an expert maximizes his/her long run expected score by and only by ensuring that his probabilities *p= (0.05, 0.45, 0.45, 0.05)* correspond to his true beliefs (Cooke, 1991). α is similar to a significance level in simple hypothesis testing, but its origin is to measure ‘goodness’ and not to reject hypotheses.

A combination of expert assessments is called a "decision maker" (DM). All decision makers discussed here are examples of linear pooling; the classical model is essentially a method for deriving weights in a linear pool. "Good expertise" corresponds to good calibration (high statistical likelihood, high p-value) and high information. Weights that reward good expertise and pass these virtues on to the decision maker are desired.

The reward aspect of weights is very important. We could simply solve the following optimization problem: find a set of weights such that the linear pool under these weights maximizes the product of calibration and information. Solving this problem on real data, one finds that the weights do not generally reflect the performance of the individual experts. As an expert's influence on the decision maker should not appear haphazard, and "gaming" the system with assessments tilted to achieve a desired outcome should be discouraged, we must impose a strictly scoring rule constraint on the weighting scheme.

The scoring rule constraint requires the term *α(Cal(e) ≥ α)* in eq (2)*,* but does not indicate what value of α we should choose. Therefore, we choose α to maximize the combined score of the resulting decision maker. Let *DMα(i)* be the result of linear pooling for any item *i* with weights proportional to (2):

*DMα(i) = ∑e=1,..E wα(e) fe,i / ∑e=1,..E wα(e)* (3)

The *optimized global weight DM* is *DMα\** where α\* maximizes

*calibration score(DMa\*) × information score(DMα\*).* (4)

This weight is termed global as the information score is based on all the assessed calibration items.

A variation on this scheme allows a different set of weights to be used for each item. This is accomplished by using information scores for each item rather than the average information score:

*wα (e,i) = 1α(calibration score)×calibration score(e) × I(fe,i | gi)* (5)

For each α we define the *Item weight DMα* for item *i* as

*IDMα(i) = ∑e=1,..E wα(e,i) fe,i / ∑e=1,..E wα(e,i)* (6)

The *optimized item weight DM* is *IDMα\** where α\* maximizes

*calibration score(IDMa) × information score(IDMα).* (7)

The non-optimized versions of the global and item weight DM’s are obtained by setting α = 0.

Item weights are potentially more attractive as they allow an expert to up- or down- weight her/himself for individual items according to how much (s)he feels (s)he knows about that item. "Knowing less" means choosing quantiles farther apart and lowering the information score for that item. Of course, good performance of item weights requires that experts successfully perform this up-down weighting. Anecdotal evidence suggests that item weights improve over global weights as the experts receive more training in probabilistic assessment. Both item and global weights can be described as optimal weights under a strictly proper scoring rule constraint. With both global and item weights, calibration strongly dominates over information, and information serves to modulate between more or less equally well calibrated experts.

Since any combination of expert distributions yields assessments for the calibration variables, any combination can be evaluated on the calibration variables. In particular, we can compute the calibration and the information of any proposed decision maker. We should hope that the decision maker would perform better than the result of simple averaging, called the *equal weight DM* (EQ), and we should also hope that the proposed DM is not worse than the best expert in the panel. The global and item weight DM’s discussed above (optimized or not) are *Performance based DM’s* (PB)*.* In general the optimized global weight DM is used, unless the optimized item weight DM is markedly superior.

For additional information and detailed tutorial on the implementation of this procedure please refer to:

Aspinall, W. (2011) Expert elicitation and judgment pooling using the Cooke Classical Model and EXCALIBUR. Available at: http://risk2.ewi.tudelft.nl/ejwshop\_2011\_files/Aspinall\_Briefing\_Note\_on\_Classical\_Model\_and\_EXCALIBUR\_Jul\_2011.pdf