GLOSSARY

EXCALIBUR

Items

quantile

If X is a continuous valued random variable, the r% quantile of the distribution of X is the smallest number xr such that

Probability{$X \leq x_r$} = r/100.

quantile point

r/100 is the quantile point corresponding to the r% quantile x_r.

scale (log, uni)

Variables measured on a uniform scale (uni) are uniformly distributed between quantiles, and relative information is measured with respect to the uniform distribution (suitably truncated).

Variables measured on a loguniform scale (log) are loguniformly distributed between quantiles, and relative information is measured with respect to the loguniform distribution (suitably truncated).

Thumb rule: if you think in terms of decades, use log scale

weights

Weights are used to determine the Decision Maker's distribution, as a normalized weighted linear combination of the experts' distributions (see DM).
In EXCALIBUR the user may choose one of four weighting schemes, namely global weights, equal weights, item weights and user weights.

Calculate

global weights

Global weights are determined by global measures of performance on seed variables, namely, calibration and average (over seed variables) relative information. When the significance level is set equal to zero, an expert's global weight is proportional to the product of calibration and average relative information over seed variables. For each expert, global weights are the same for all items.

item weights

Item weights are determined for each item separately, using (the global measure) calibration and the relative information for each item.

user weights

Weights supplied by the user.

significance level

The significance level determines the calibration threshold value (see Calibration). Calibration scores greater or equal to the significance level correspond to non-rejected statistical hypotheses. The unnormalized global and item weights are asymptotically strictly proper scoring rules (see proper scoring rule or Experts in Uncertainty, chapter 9) only if combined with significance testing at a non-zero significance level. The significance testing entails that the weights become zero whenever the calibration score is strictly less than the significance level. The theory of strictly proper scoring rules does not determine what the significance level should be, this is determined by optimization.

DM

Decision Maker. The DM's distributions are determined as the weighted combinations of the experts' assessments. If F_i is the cumulative distribution function of expert i for a given item, and NE is the number of experts, then:

$$F_i = \frac{\Sigma_{i=1..NE} [w_i * F_i]}{\Sigma_{i=1..NE} [w_i]} .$$

where w_i are the global, item weights or user weights.

Optimized DM

The optimized DM results by choosing that significance level for which the global unnormalized weight of the DM is maximal. Optimization is meaningful for global and item weights. It is not meaningful for equal or user weights.

Calibration power

The power of a statistical test is its ability to distinguish between rival hypotheses, and increases with the number of independent samples. Calibration power may be chosen from the interval [0.1, 1.0], and determines the effective number of samples. Choosing 50% power means reducing the resolution of the significance test to that of a test with half the number of samples.

Instead of calculating experts calibration with the formula (see Calibration)

$C(e) = 1 - \text{Chisq}_R(2*M*I(s(e),p))$

the following formula is used in calculations:

$C(e) = 1 - \text{Chisq}_R(2*M*I(s(e),p)*\text{Power})$

where Power $\in$ [0.1, 1.0].

Intrinsic Range

The expert's quantile assessments are used to fit a minimally informative distribution, relative to the background measure (see scale). For this minimization it is necessary that the set of possible values be restricted to a bounded interval. By default CLASS uses the "10% overshoot" rule: The smallest interval containing all assessments for a given item (plus the realization, if available) is overshot by 10% above and below. The expert's information scores are affected by the choice of the overshoot; making this overshoot very large tends to suppress differences in the experts' information scores, however, the effect is very slow.

Bayesian updates

Bayesian updating in the present implementation is accomplished using a non-informative prior and a multinomial likelihood function. The number of possible outcomes is one plus the number of quantiles. If the n quantile points are arranged:

$0 = r_0 < r_1 < r_2 < ... r_n < r_{n+1} = 1$

then the j-th outcome of the multinomial distribution corresponds to the realization falling between the r_j-th and r_{j-1}-th quantile. By definition, an expert assigns the j-th outcome the probability $r_j - r_{j-1}$.

In effect, the Bayesian updating recalibrates the experts' assessments (conditional on a choice of scale and cut-off points for each variable). Updated 5%, 50% and 95% quantiles are computed using a minimum information fit to the updated quantiles.

Discrepancy:

To perform discrepancy analysis, the relative information of each expert's assessment, per item, is compared with the DM's assessment for that item, and the relative information of the expert with respect to the DM is computed. These scores are averaged over all items. The average scores (which are proportional to the relative information of the respective joint distributions if all items are independent) are output. This enables the user to see which experts agree or disagree most with the DM. (Dis)agreement is not well predicted by an experts unnormalized weight.

In addition, the ratio of the largest/smallest relative information score per item is output. This enables the user to flag those items for which the experts assessment of uncertainty differs most.

Robustness (items):

Seed items are excluded from the analysis, one at a time and the resulting DMs are computed and compared, using the current parameter values under "RUN". The total relative information with respect to the background measure, the calibration and the total relative information with respect to the original DM are tallied.

idem (experts)

This feature is similar to the previous feature, except that experts are excluded form the analysis one at a time.

Display Results

Calibration

Calibration measures the statistical likelihood of the hypothesis that the realizations are sampled independently from distributions agreeing with the expert's assessments. If s and p are the sample and theoretical distributions respectively for the the "inter-quantile interval" multinomial variable (see Bayesian Updates), and if I(s,p) denotes the relative information of s with respect to p), then calibration corresponds to the probability of seeing a deviation between s and p at least as great as I(s,p) under the above mentioned hypothesis. The larger this probability, the better the calibration. p is the same for all experts, as all experts assess the same quantiles, and s depends on the expert assessments. To indicate the dependence on expert e, we write s(e). The asymptotic value of this probability is

$$C(e) = 1 - \mathrm{Chisq}_R\ (2\times M\times I(s(e),p))$$

Where Chisq_R is the cumulative Chi square distribution function with R degrees of freedom; R = the number of quantiles, M is the number of seed variables and e is the expert in question.

Relative information :

The relative information of probability vector $s = (s_1,...s_n)$ with respect to $p = (p_1,...p_n)$ is:

$$I(s|p) = \Sigma_{(i=1..n)}[s_i\times \ln(s_i/p_i)]$$

If s is a sample distribution gotten from M independent samples from p, then 2MI(s,p) is asymptotically Chi square distributed with n-1 degrees of freedom.

Relative information is used as a Chi square test statistic in the measure of calibration, but this quantity is not displayed (one can recover this value by inverting the Chi square cumulative distribution function, and dividing by (2×M×power).

The relative information score displayed is the average over i of I(f(i,e),g(i)) where f(i,e) is the minimal information density function fitted to expert e's quantiles for item i, and g(i) is either the uniform or loguniform density function, depending on the scale of the item. Two scores are displayed, namely the average over all items and the average over the seed items.

Unnormalized weight

Letting a denote the significance level, M the number of seed variables, and 1_A the indicator function for event A, the unnormalized (global) weight for assessor j (either an expert or the DM) is computed as:

$$w_j = C(j)\times(1/M)\times\Sigma_{(I=1..M)}[I(f(i,j),g(I))\times 1_{\{C(j) > a\}}]$$

In other words, the unnormalized weight displayed is always the global weight.

When using equal weights or user weights, a in the above expression is set equal to zero.

When using item weights, a is chosen to optimize the DM's global unnormalized weight. However, the weights used to calculate the DM's assessments are not the unnormalized weights shown for the experts. Expert j's weight for item i is:

$$w_j(i) = C(j)\times I(f(i,j),g(i))\times 1_{\{C(j) > a\}}$$

normalized weight NO DM :

These are the weights used in determining the DM. If equal or user weights are used, then these are shown, if global weights are used, then the normalized global weights are shown. If item weights are used, then this column is blank, since the weights used for the DM vary from item to item.

normalized weight with DM:

These weights result from normalizing the experts and the DM's unnormalized weights.

Strictly proper scoring rules:

Suppose an expert assesses that an uncertain quantity with outcomes 1,...n has distribution $p = p_1,..p_n$. A scoring rule R assigns a score to this distribution on the basis of the realization, say i, R(p,i). If the expert "really believes" that the uncertain quantity follows a distribution q, then his expected score on stating p is $E_q(R(p,i))$. R is a strictly proper scoring rule for a single uncertainty quantity if

$$\underset{p}{\text{argmax}}\ E_q(R(p,i)) = q$$

In other words, the expert maximizes his expected score under R by, and only by, stating his true belief. The classical model is based on a generalization of this notion, whereby a score is associated with a SET of assessments and realizations. The variables in question are assumed for convenience to have the same range (1,...n). Let the set of assessments consists of M variables with assessed distribution p (see Calibration). The set of sample realizations may be represented with the sample distribution s (see Calibration), and let R(p,M,s) assign a score on the basis of this information. If the expert really believes that the uncertain quantities have (joint) distribution Q, then $E_Q(R(p,M,s))$ is his expected score. Consider the M marginal distributions gotten from Q for each uncertain quantity, and let $q = q_1,...q_n$ denote the arithmetical average of these M marginal distributions. It is not difficult to show that q is also the expected relative frequency distribution under Q for the uncertain quantities. That is, qi is the expected relative frequency of outcome i under Q. Rule R(p,M,s) is a strictly proper scoring rule for average probabilities (or equivalently for expected relative frequencies) if

$$\underset{p}{\text{argmax}}\ E_Q(R(p,M,s)) = q$$

In (R. Cooke, Experts in Uncertainty, Oxford U. 1991) a characterization of rules with this property is given. The score 2*M*I(s,p) is strictly proper in this sense.

The global and item weights are "asymptotically strictly proper" in the following sense: Under suitable assumptions, for all p, if $p \neq q$ then for sufficiently large M (depending on p)

$$E_Q(R(p,M,s)) > E_Q(R(p,M,s))$$