

A method to calibrate assessors

R.S.MacKay

Mathematics Institute and Centre for Complexity Science
University of Warwick, Coventry CV4 7AL, U.K.

January 27, 2009

Abstract

1 Introduction

It frequently happens that a set of objects has to be evaluated by a panel of assessors but each assessor evaluates only a subset of the objects. There may be too many objects or the range of expertise of an assessor might not cover the range of the objects.

Examples are:

- evaluation of grant proposals submitted to a panel
- evaluation of candidates for appointment to a lectureship
- evaluation of students' performance in an options system where they can choose to take examinations from a list of options
- evaluation of the quality of submitted research publications in the UK research assessment exercise

One might propose just to assign to each object the average of the scores assigned by those assessors who assessed the object, but this would ignore the likely possibility that assessors have different standards. One could shift the scores for each assessor to make his average take a standard value, but this would ignore the possibility that the set of objects assigned to one assessor may be of a genuinely different standard from that assigned to another.

For any experimental scientist, the issue is obvious: "calibration".

One option is to calibrate the assessors beforehand on a common subset of objects, perhaps disjoint from the set to be evaluated, meaning that they each evaluate all the objects in the subset and then some rescaling is agreed to bring the assessors into line as far as possible. This would not work well, however, in a situation where the range of objects is broader than the expertise of an assessor.

If two assessors' expertises overlap on some subject, however, any discrepancy between their evaluations can be used to infer something about the assessors' relative standards. Thus if the graph on the set of assessors formed by linking two whenever they can assess a common object is sufficiently well connected (to be made precise), one can expect to be able to infer a robust calibration of the assessors.

In this paper I propose a method to achieve robust calibration of assessors and hence robust scores for the outputs.

2 The basic model

Let us suppose that each assessor is assigned a subset of the objects to evaluate. Denote the resulting set of (assessor, object) pairs by E .

Let us suppose that the score s_{ao} that assessor a assigns to object o is a real number related to a "true" value v_o for the object by

$$s_{ao} = v_o + b_a + \varepsilon_{ao}, \quad (1)$$

where b_a can be called the *bias* of assessor a and ε_{ao} are independent zero-mean random variables.

To take into account the varying expertise of the assessors, each assessor is asked in addition to the score s_{ao} to specify a level of confidence $c_{ao} \in \mathbb{R}^+$ for that evaluation. Let us suppose that

$$\varepsilon_{ao} = \eta_{ao}/c_{ao}, \quad (2)$$

with η_{ao} random variables of unit variance.

Thus our basic model is

$$s_{ao} = v_o + b_a + \frac{\eta_{ao}}{c_{ao}}, \quad (3)$$

with η_{ao} independent zero-mean, unit variance random variables. Shortly, we will implicitly take them to be normally distributed.

This model is likely to be too simple. For example, assessor bias might not be a purely additive effect (e.g. an assessor may have a bias for or against topics in which they have lower confidence), assessors may like to give round-number scores, and they may have different scales for confidence. Nevertheless, let us see how far we can proceed with it.

3 Solution of the basic model

Given the data $\{(s_{ao}, c_{ao}) : (a, o) \in E\}$ for all assigned assessor-object pairs, we wish to extract the true values v_o and assessor biases b_a . The simplest procedure is to minimise the sum of squares $\sum_{(a,o) \in E} \eta_{ao}^2$. This can be justified as giving the maximum-likelihood values for v_o and b_a if the η_{ao} are assumed to be normally distributed.

Now

$$\sum_{(a,o) \in E} \eta_{ao}^2 = \sum_{(a,o) \in E} c_{ao}(s_{ao} - v_o - b_a)^2 \quad (4)$$

is minimised with respect to v_o iff

$$\sum_{a:(a,o) \in E} c_{ao}(s_{ao} - v_o - b_a) = 0, \quad (5)$$

and with respect to b_a iff

$$\sum_{o:(a,o) \in E} c_{ao}(s_{ao} - v_o - b_a) = 0. \quad (6)$$

It is notationally convenient to extend the sums to all assessors (respectively objects) by assigning the value $c_{ao} = 0$ to any assessor-object pair that is not in E (i.e. for which a score was not returned). Then the above conditions can be written as

$$\begin{aligned} \left(\sum_a c_{ao}\right)v_o + \sum_a b_a c_{ao} &= V_o := \sum_a c_{ao}s_{ao} \\ \sum_o c_{ao}v_o + \left(\sum_o c_{ao}\right)b_a &= B_a := \sum_o c_{ao}s_{ao}. \end{aligned} \quad (7)$$

This is a linear system of equations for the v_o and b_a . It has an obvious degeneracy, namely that one could add a constant k to all the v_o and subtract k from all the b_a and obtain another solution. Thus we should fix this degeneracy by for example imposing $\sum_a b_a = 0$ (or perhaps better $\sum_{ao} b_a c_{ao} = 0$, but I'll stick to the first choice).

Does it have a solution and a unique one, after this fix? That depends on the connectivity of the graph Γ of assessors and objects, where an assessor a and object o are linked iff $(a, o) \in E$ (I will soon put weights on the edges of this graph, but they are not required yet). The linear operator defined by writing (7) as

$$L \begin{bmatrix} v \\ b \end{bmatrix} = \begin{bmatrix} V \\ B \end{bmatrix}, \quad (8)$$

where v, b, V and B denote the column vectors formed by the v_o, b_a, V_o and B_a respectively, has null space of dimension equal to the number of connected components of Γ (this follows from Perron-Frobenius theory). Thus if Γ is connected then the null space of L has dimension one, so corresponds precisely to the null vectors $v_o = k\forall o, b_a = -k\forall a$, we already noticed and dealt with.

The only thing remaining to check is that the right hand side (RHS) of (8) lies in the range of L , thus ensuring that a solution exists. This is true if all null forms of the adjoint operator L^\dagger send the RHS to zero. The null space of L^\dagger has the same dimension as that of L , because square, and an obvious non-zero null form is α defined by

$$\alpha(v, b) = \sum_o v_o - \sum_a b_a, \quad (9)$$

so it suffices to check that $\alpha(V, B) = 0$, which is true.

Thus under the assumption that the assessor-object graph Γ is connected, (7) has a unique solution (v, b) satisfying $\sum_a b_a = 0$. Note that it is clear that connectedness

of Γ is necessary for uniqueness, else one could play the same sort of game adding and subtracting constants independently in each connected component of Γ and thereby produce more solutions.

What is a good algorithm to find this unique solution? Numerical linear algebra is a highly developed subject, so I hesitate to suggest an algorithm myself, but the equations (7) have a special structure that is probably worth exploiting. For example, the first equation can be written as

$$v_o = \frac{V_o - \sum_a b_a c_{ao}}{C_o}, \text{ where } C_o = \sum_a c_{ao}. \quad (10)$$

This can be substituted into the second equation to obtain

$$\sum_o \frac{c_{ao}(V_o - \sum_{a'} b_{a'} c_{a'o})}{C_o} + (\sum_o c_{ao})b_a = B_a, \quad (11)$$

i.e.

$$\sum_{a'} \sum_o \frac{c_{ao} c_{a'o}}{C_o} b_{a'} - (\sum_o c_{ao})b_a = \sum_o \frac{c_{ao} V_o}{C_o}. \quad (12)$$

This system (12) has dimension the number of assessors (rather than the sum of the numbers of assessors and objects), which is probably relatively small in applications. Replacing one of the equations by $\sum_a b_a = 0$ gives a system with a unique solution that can be solved for b by any method, e.g. LUP decomposition [?]. Then v can be recovered from (10).

4 Robustness

A key question with any black-box solution like this is how robust is the outcome to mistakes or odd judgements. That is mainly a question of the size of L^{-1} , regarded as an operator from the subspace where $\alpha(v, b) = 0$ to that where $\sum b_a = 0$.

We have to decide which norm is most appropriate to measure this. It seems to me that supremum norm is the relevant one on (v, b) , because we want to know what is the maximum effect of a change to the s_{ao} on all the v_o and b_a . One could argue for either the supremum or ℓ_1 -norm on $s = (s_{ao})_{(a,o) \in E}$. Then one would quantify the size of L^{-1} by its operator norm $\|L^{-1}\|$. Also we might want to use a semi-norm that quotients out the null space of L .

In any case, $\|L^{-1}\|$ is a measure of how well connected Γ is, taking into account weights on its edges given by the expressed confidences c_{ao} . If one has to take a chain with a low product of confidences to connect some pair of vertices in Γ , then $\|L^{-1}\|$ will be small, whereas if each pair is connected by a path with large product of confidences then it is large.

In ℓ_2 -norm, $\|L^{-1}\|^{-1}$ is just the second eigenvalue of L (the first being 0 on account of the degeneracy), known as the *algebraic connectivity* of Γ in the case that the edges all have weight one [Fi].

For $s = (s_{ao})_{(a,o) \in E}$, define operator K by

$$Ks = \begin{bmatrix} V \\ B \end{bmatrix}, \quad (13)$$

as a shorthand for the definitions in (7), so equation (7) can be written as

$$L \begin{bmatrix} v \\ b \end{bmatrix} = Ks. \quad (14)$$

Whatever norms are chosen, if a change δs is made to the scores, we obtain changes δv , δb of size

$$\left\| \begin{bmatrix} \delta v \\ \delta b \end{bmatrix} \right\| \leq \|L^{-1}K\| \|\delta s\| \leq \|L^{-1}\| \|K\| \|\delta s\|, \quad (15)$$

where the appropriate operator norms are taken for L^{-1} and K . Thus the task for the designer of E is to make $\|L^{-1}\| \|K\|$ small (or better, $\|L^{-1}K\|$ if it is easy to estimate).

Another point of view on robustness is the Bayesian one. From a prior probability on (v, b) and a model for the η_{ao} one would infer a posterior probability for (v, b) , whose width would tell one how robust the inference was. In the particular case of flat priors and Gaussian noise, the posterior is Gaussian with mean at the value solving (7) and with covariance matrix related to L^{-1} .

One has also to consider robustness with respect to changes in the confidences c_{ao} . If an assessor declares extra high confidence for an evaluation, that can skew the resulting v and b a lot. The analysis is more subtle, however, because of how c_{ao} appears in the equations.

5 Refinements

More complicated models for the scoring process can be proposed. In particular, assessors might have not only an additive bias but also different scales, so for example

$$s_{ao} = \lambda_a v_o + b_a + \eta_{ao}/c_{ao}. \quad (16)$$

Fitting λ, v, b is more complicated than just v, b .

One problem is that often assessors are asked to assign scores in a fixed range, e.g. 1–10, and then any model for bias really ought to be nonlinear to respect the endpoints. On the other hand, I feel it can be a mistake to specify a fixed range because it requires an assessor to have a feel for the range of the objects before starting scoring. Thus I would propose asking assessors to use any real numbers and then use (16) to extract true values v .

A simpler strategy that might work nearly as well is to allow assessors to use any positive numbers but then to take logarithms and fit (3) to the log-scores. The assessor biases would then be like logarithms of exchange rates. The confidences would need translating appropriately too.

One might need a more subtle model if some assessors value objects nearer to their expertise more highly than those further away.

6 Comments

An advantage of this method is that it does not produce the artificial discontinuities across field boundaries that tend to arise if the domain is partitioned into fields and evaluation in each field carried out separately.

7 Example

References

[Fi] Fiedler M, Algebraic connectivity of graphs, Czech Math J 23 (1973) 298–305.