

Technical Report IDSIA-25-98

May 16, 1998

Scoring Conference Submissions by Alternate Maximization of Likelihood

Nicol N. Schraudolph

`nic@idsia.ch`

IDSIA, Corso Elvezia 36

6900 Lugano, Switzerland

<http://www.idsia.ch/>

Abstract

We address the problem of combining the subjective, confidence-tagged opinions of experts that independently review a set of like items, such as submissions to a scientific conference. The conventional approach of confidence-weighted averaging is improved upon by augmenting a probabilistic error model with *bias* and *trust* parameters that characterize the subjectivity of the referees' quality and confidence judgments, respectively. The likelihood of the review data under this model is then optimized by *alternate maximization* (AM) with respect to item scores and referee parameters. Since conditionally optimal trust parameters cannot be calculated explicitly, we provide two iterative schemes for this purpose. In preliminary experiments the resulting *generalized* AM algorithm was found to be robust, efficient and effective. We are set to field-test it in the peer review process of the NIPS*98 conference.

1 Confidence-Weighted Averaging

Consider a set of experts who independently review a set of like items by scoring the quality of each given item, as well as the confidence they have in their own judgment. In order to eliminate trivial subcases, we assume that each of the experts reviews more than one item, that items typically have more than one referee, and that not every referee reviews the same items. This is the typical situation, for instance, in the peer review process of a scientific conference.

How are such fragmentary, confidence-tagged individual opinions best combined into a consensus score for each item? One principled approach is to set up a probabilistic model of the errors made by the referees, then find the consensus scores that maximize the likelihood of the set of items. Let us assume that each

referee r misjudges each item i 's true value s_i^* by a Gaussian error with zero mean and variance w_{ir}^{-1} . Since referees are deemed to act independently, the likelihood of a consensus score s_i is described by the product distribution

$$s_i \sim \prod_r N\left(s_{ir}, w_{ir}^{-\frac{1}{2}}\right)_{s_i} \quad (1)$$

where $N(\mu, \sigma)_x$ designates the normal density with mean μ and standard deviation σ evaluated at x , and s_{ir} is the score assigned to item i by referee r . (For notational convenience, we shall imply that an index r always ranges over the set of experts that have reviewed a given item i , and *vice versa*.) The maximum likelihood estimate \hat{s}_i for s_i^* is then found by setting the derivative of the logarithm of (1) to zero:

$$\begin{aligned} \frac{\partial}{\partial s_i} \ln \prod_r N\left(s_{ir}, w_{ir}^{-\frac{1}{2}}\right)_{s_i} &= 0 \\ \frac{\partial}{\partial s_i} \frac{1}{2} \left[\sum_r \ln \frac{w_{ir}}{2\pi} - \sum_r w_{ir} (s_i - s_{ir})^2 \right] &= 0 \\ \sum_r w_{ir} (s_i - s_{ir}) &= 0 \\ \Rightarrow \hat{s}_i &= \sum_r w_{ir} s_{ir} / \sum_r w_{ir} . \end{aligned} \quad (2)$$

The maximum likelihood consensus scores \hat{s}_i are thus weighted averages of the individual referees' scores s_{ir} . A popular way to obtain appropriate weights is to ask referees for the confidence $c_{ir} > 0$ they have in each of their judgments. The weight w_{ir} is then set as a function of c_{ir} ; in the simplest case — $w_{ir} \propto c_{ir}$ — we obtain the widely-used method of confidence-weighted averaging.

2 Referee Bias

Although the referees may receive *a priori* instruction on how to assign scores to items, they must not be allowed to interact with each other, lest this compromise their independence — and with it our probabilistic model. Experts would typically use mutual interaction to jointly calibrate the scale on which they assess items; the absence of such a common baseline makes it likely that individual referees make systematic errors, which are not well-accounted for by the zero-mean error model above. We therefore augment our probabilistic model by giving each referee a *bias* β_r by which their scores are shifted:

$$\Pr(s_i) = \prod_r N\left(s_{ir} - \beta_r, w_{ir}^{-\frac{1}{2}}\right)_{s_i} \quad (3)$$

The maximum likelihood estimate \hat{s}_i of the true value s_i^* under model (3) is then found — analogously to (2) — to be

$$\hat{s}_i = \sum_r w_{ir}(s_{ir} - \beta_r) \bigg/ \sum_r w_{ir} . \quad (4)$$

The problem here is that we do not know the true biases β_r^* of the referees. Given that multiple referees score overlapping sets of items, however, we may use a referee r 's deviation of opinion from her peers to infer a plausible value of β_r . One possibility is to use the weighted average score given by each expert to the items she reviewed:

$$\beta_r = \sum_i w_{ir}s_{ir} \bigg/ \sum_i w_{ir} . \quad (5)$$

This normalizes the scores of each referee to a weighted average of zero. While this serves to align the experts' subjective scales, it creates a new problem: referees that happen to have evaluated predominantly above- or below-average items now have no more means to communicate this. In the realistic case that each expert reviews only a few items, we have thus merely traded subjectivity for sampling error. A better alternative is to use our probabilistic model (3) to determine the maximum likelihood biases $\hat{\beta}_r$ from given scores s_i . Using the Gaussian prior $\beta_r \sim N(0, \lambda^{-\frac{1}{2}})$ for the biases we obtain

$$\begin{aligned} \frac{\partial}{\partial \beta_r} \ln \left[N\left(0, \lambda^{-\frac{1}{2}}\right)_{\beta_r} \prod_i N\left(s_{ir} - \beta_r, w_{ir}^{-\frac{1}{2}}\right)_{s_i} \right] &= 0 \\ \frac{\partial}{\partial \beta_r} \frac{1}{2} \left[\ln \frac{\lambda}{2\pi} - \lambda \beta_r^2 + \sum_i \ln \frac{w_{ir}}{2\pi} - \sum_i w_{ir}(s_i - s_{ir} + \beta_r)^2 \right] &= 0 \\ \lambda \beta_r + \sum_i w_{ir}(s_i - s_{ir} + \beta_r) &= 0 \\ \Rightarrow \hat{\beta}_r = \sum_i w_{ir}(s_{ir} - s_i) \bigg/ \left(\lambda + \sum_i w_{ir} \right) . & \quad (6) \end{aligned}$$

So given a set of biases β_r , we can now compute a maximum likelihood estimate of item scores s_i via (4), and *vice versa* via (6). By alternating between these two equations (always using most recent estimates on the respective right-hand sides), we can iteratively improve both estimates in likelihood. We call this approach *alternate maximization* (AM), as it bears some resemblance to the *expectation-maximization* (EM) algorithm [1, 2].

3 Referee Trust

Our probabilistic model is still incomplete, since we have not yet taken into account that the experts have no common basis for their confidence assessments

either. Some referees may well be more reliable than others, even if they express similar confidence c_{ir} in their judgment. To account for this, we adopt an inherent *trust* γ_r in each expert, and posit that their reviews suffer from an additional, independent source of error with variance $e^{-\gamma_r}$. That is, the total variance of their score is now given by

$$w_{ir}^{-1} = e^{-\gamma_r} + c_{ir}^{-1}. \quad (7)$$

Note that the effect of γ_r is nonlinear: low trust affects high-confidence judgments more severely than low-confidence ones, while high trust implies that the referee's confidence assessment (whether high or low) is taken at face value. The maximum likelihood trust parameters $\hat{\gamma}_r$ (given s_i and β_r) now satisfy

$$\begin{aligned} \frac{\partial}{\partial \gamma_r} \ln \prod_i N\left(s_{ir} - \beta_r, w_{ir}^{-\frac{1}{2}}\right)_{s_i} &= 0 \\ \frac{\partial}{\partial \gamma_r} \frac{1}{2} \left[\sum_i \ln \frac{w_{ir}}{2\pi} - \sum_i w_{ir} x_{ir} \right] &= 0 \\ \sum_i w_{ir}^2 (e^{-\gamma_r} + c_{ir}^{-1} - x_{ir}) &= 0, \\ \text{where } x_{ir} &\equiv (s_i - s_{ir} + \beta_r)^2. \end{aligned} \quad (8)$$

Unfortunately this does not provide a closed-form solution for $\hat{\gamma}_r$, due to the nonlinear — though in our eyes most appropriate — fashion in which the trust parameters modulate score weighting according to (7). Instead we can employ modified Newton-Raphson iterations to locally maximize the likelihood as a function of γ_r :

$$\gamma_r \leftarrow \gamma_r - \frac{L'_r(\gamma_r)}{\min[L''_r(\gamma_r), -\eta |L'_r(\gamma_r)|]}, \quad \text{where} \quad (9)$$

$$\begin{aligned} L_r(\gamma_r) &\equiv \ln \prod_i N\left(s_{ir} - \beta_r, w_{ir}^{-\frac{1}{2}}\right)_{s_i}, \\ L'_r(\gamma_r) &\equiv \frac{\partial L_r(\gamma_r)}{\partial \gamma_r} = \frac{1}{2e^{\gamma_r}} \sum_i w_{ir} [1 - w_{ir} x_{ir}], \quad \text{and} \\ L''_r(\gamma_r) &\equiv \frac{\partial^2 L_r(\gamma_r)}{(\partial \gamma_r)^2} = \\ &= \frac{1}{2e^{2\gamma_r}} \sum_i [w_{ir}(1 - w_{ir} x_{ir})(w_{ir} e^{-\gamma_r} - 1) - w_{ir}^3 x_{ir} e^{-\gamma_r}]. \end{aligned}$$

The modification in the denominator of (9) ensures that $L_r(\gamma_r)$ is maximized with steps limited in size to $|\Delta \gamma_r| \leq \eta^{-1}$. The choice of η is not critical; in our experiments (see Section 5) we have found $\eta = 2$ to afford reasonably fast yet

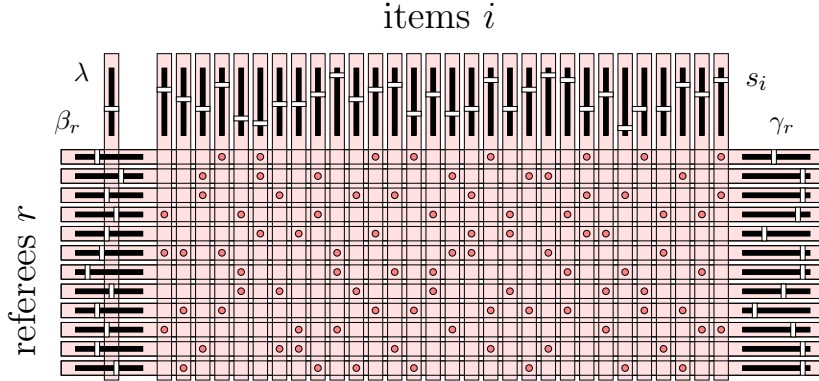


Figure 1: A graphical depiction of our probabilistic model. Both referees (rows) and items (columns) have adjustable parameters (sliders) whose settings affect the likelihood of a corresponding subset of reviews (disks). Our algorithm alternately optimizes each of the three banks of sliders, representing item score, referee bias, and trust parameters, respectively.

reliable convergence. Alternatively, we have found that the iteration

$$e^{-\gamma_r} \leftarrow \left[\sum_i w_{ir}^2 (x_{ir} - c_{ir}^{-1}) \right]^+ / \sum_i w_{ir}^2, \quad (10)$$

— which follows from (8) under addition of a positive-bounding operation — generally converges faster than the modified Newton-Raphson approach, and just as reliably. Iterating over $e^{-\gamma_r}$ rather than γ_r has the additional advantages that the important case of infinite trust ($e^{-\gamma_r} = 0$) is handled well numerically, and that repeated — and costly [3] — exponentiation operations are avoided, since the trust parameters can now always be kept in exponentiated form.

4 Optimization Algorithm

We wish to optimize the empirical log-likelihood

$$L(s_i, \beta_r, \gamma_r) = \sum_{i,r} \ln N\left(s_{ir} - \beta_r, w_{ir}^{-\frac{1}{2}}\right)_{s_i} \quad (11)$$

of the review data with respect to three sets of parameters: the (estimated) consensus scores s_i of the items, and bias β_r and trust γ_r parameters of the referees (see Figure 4 for a graphical depiction). In the preceding two sections we have derived formulæ for optimizing each set of parameters, given the other two. We can thus locally maximize the likelihood by applying each formula in turn, always using the most recent parameter estimates on the right-hand side.

This approach — which we call *alternate maximization* (AM) — is reminiscent of the well-known expectation-maximization (EM) algorithm [1, 2]. Where EM interleaves estimation (“E”) and maximization (“M”) steps, however, AM alternates between multiple, complementary maximization steps.

Since we have no closed-form maximum likelihood estimate for the γ_r , we iteratively optimize them in the innermost loop; by analogy with EM nomenclature we call this a *generalized* AM (GAM) method. Our GAM algorithm proceeds as follows:

1. \forall referees r : initialize $\beta_r = \gamma_r = 0$.
2. \forall items i : compute s_i according to (4).
3. \forall referees r : compute β_r according to (6).
4. \forall referees r : optimize γ_r according to (9) or (10); if *improved* go to 4.
5. \forall referees r : recompute β_r according to (6); if *improved* go to 4.
6. \forall items i : recompute s_i according to (4); if *improved* go to 4.

The predicate *improved* holds whenever the empirical log-likelihood (11) has been increased appreciably (*i.e.*, by more than a small constant ϵ) by the preceding optimization step. We used $\epsilon = 10^{-9}$ in the experiments reported below.

5 Empirical Results

We tested the above algorithm, along with a number of simpler controls, on a large synthetic data set intended to mimic the peer review process of a scientific conference: 1500 items with random scores $s_i^* \sim N(0, 3)$ were each reviewed by three independent referees, chosen at random from a pool of 500. Each referee had random bias $\beta_r^* \sim N(0, 1)$ and trust $\gamma_r^* \sim N(0, 1)$; each examined between 1 and 20 items (9 on average). For each review, the score s_{ir} and confidence c_{ir} were given by

$$s_{ir} \equiv s_i^* + \beta_r^* + \mu_{ir} + \nu_{ir} \quad \text{and} \quad c_{ir} \equiv (\nu_{ir}^2 + 0.1)^{-1}, \quad (12)$$

where $\mu_{ir} \sim N(0, e^{-0.5\gamma_r^*})$ and $\nu_{ir} \sim N(0, 0.5)$ are independent noise terms. Thus while the referees reported on one source of noise (ν_{ir}) via their confidence self-assessment, the other noise term (μ_{ir}) — which depends on the hidden parameter γ_r^* — remained invisible, and had to be estimated by the algorithm, along with the biases β_r and item scores s_i .

The 4500 tuples (i, r, s_{ir}, c_{ir}) generated according to (12) were given to the algorithm described in Section 4. As controls, we selectively disabled the optimization of β_r and/or γ_r , and also implemented an option to normalize biases according to (5). The weights in the model were always given by (7); where trust was not optimized it was assumed to be infinite ($\forall r : e^{-\gamma_r} \equiv 0$). The algorithm is not very sensitive with respect to its free parameters, which we set to $\eta = 2$ and $\epsilon = 10^{-9}$.

Table 5 shows the results of running this algorithm in four different configurations, with a uniform prior on biases ($\lambda = 0$). The upper half names each

Experiment:	a)	b)	c)	d)
normalization: bias (5)		✓		
optimization: bias (6)			✓	✓
trust (9)				✓
average log-likelihood:	-5.425	-5.310	-3.329	-2.025
item rank differences:				
average absolute	108.1	118.6	91.2	82.9
root-mean-square	147.4	163.1	129.3	113.6
maximum absolute	735	848	851	608

Table 1: Benchmark results obtained on a synthetic data set.

experiment, and shows which options (with equation numbers shown in parentheses) were used in it. The bottom half shows the average log-likelihood per review achieved by the algorithm, and the average, RMS, and maximum absolute rank difference between the items ranked according to their true score s_i^* *vs.* the score s_i computed by the algorithm from the review data. We observe that

1. although the bias normalization in b) slightly raised the log-likelihood of the model, it actually worsened performance in terms of item rank differences by about 10% as compared to a). On the other hand,
2. the maximum likelihood approach of c) and d) not only (naturally) increased the likelihood, but also reduced the average and root-mean-square item rank differences *vs.* the confidence-weighted averaging of a).
3. The best results were achieved by optimizing both bias and trust parameters in d), thus validating the utility of our probabilistic model.

We also explored the behavior of these four algorithms on a set of real data: in a blind experiment, we computed consensus scores for the 79 papers submitted to the theory section of the NIPS*96 conference, given only the quality and confidence judgments of the 25 referees.

Since an objective measurement of item value is naturally not available in this context, however, we found it rather difficult to interpret the results. Confidence-weighted averaging correlated best with the program committee’s acceptance/rejection decisions, but this is not surprising given that the committee used confidence-weighted average scores in their decision-making process.

An important outcome of this test, however, was that our approach is capable of providing a wealth of useful information *in addition to* the consensus scores. For 6 of the 7 referees flagged by the GAM algorithm as “weak” (low trust), for instance, the area chair concurred with this assessment [4]. Similarly, the average log-likelihood of reviews for each paper promises to be very useful in guiding the allocation of additional reviews to those papers that need it most.

6 Conclusion

We have proposed a probabilistic model for subjective, confidence-tagged expert opinions, and described alternate maximization, a maximum likelihood technique for combining them into consensus scores. On a large synthetic data set our model outperforms conventional confidence-weighted averaging. A blind test on past conference submissions showed that in addition to the consensus scores, our method furnishes useful complementary information about individual referees and items.

We envision the use of our algorithm as an interactive tool to assist in decision-making (*e.g.*, which papers to accept at a conference): by modeling the complex web of interactions between referees and items, the expected and unexpected consequences of interventions in this system are made apparent (“what-if” scenarios). We intend to explore this approach in the peer review process of NIPS*98.

Acknowledgments

We would like to thank Sara Solla for her invaluable advice and support. This work was supported by the Swiss National Science Foundation under grant numbers 2100-045700.95/1 and 2000-052678.97/1.

References

- [1] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm (with discussion)”, *Journal of the Royal Statistical Society series B*, **39**:1–38, 1977.
- [2] M. I. Jordan and R. A. Jacobs, “Hierarchical mixtures of experts and the EM algorithm”, *Neural Computation*, **6**(2):181–214, 1994.
- [3] N. N. Schraudolph, “A fast, compact approximation of the exponential function”, Technical Report IDSIA-07-98, Istituto Dalle Molle di Studi sull’Intelligenza Artificiale, Corso Elvezia 36, 6900 Lugano, Switzerland, 1998. <ftp://ftp.idsia.ch/pub/nic/exp.ps.gz>
- [4] S. A. Solla, personal communication, 1998.