Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence

How to Calibrate the Scores of Biased Reviewers by Quadratic Programming

Magnus Roos
Institut für Informatik
Heinrich-Heine-Univ. Düsseldorf
40225 Düsseldorf, Germany

Jörg Rothe
Institut für Informatik
Heinrich-Heine-Univ. Düsseldorf
40225 Düsseldorf, Germany

Björn Scheuermann
Institut für Informatik
Julius-Maximilians-Univ. Würzburg
97074 Würzburg, Germany

Abstract

Peer reviewing is the key ingredient of evaluating the quality of scientific work. Based on the review scores assigned by the individual reviewers to the submissions, program committees of conferences and journal editors decide which papers to accept for publication and which to reject. However, some reviewers may be more rigorous than others, they may be biased one way or the other, and they often have highly subjective preferences over the papers they review. Moreover, each reviewer usually has only a very local view, as he or she evaluates only a small fraction of the submissions. Despite all these shortcomings, the review scores obtained need to be aggregrated in order to globally rank all submissions and to make the acceptance/rejection decision. A common method is to simply take the average of each submission's review scores, possibly weighted by the reviewers' confidence levels. Unfortunately, the global ranking thus produced often suffers from a certain unfairness, as the reviewers' biases and limitations are not taken into account.

We propose a method for calibrating the scores of reviewers that are potentially biased and blindfolded by having only partial information. Our method uses a maximum likelihood estimator, which estimates both the bias of each individual reviewer and the unknown "ideal" score of each submission. This yields a quadratic program whose solution transforms the individual review scores into calibrated, globally comparable scores. We argue why our method results in a fairer and more reasonable global ranking than simply taking the average of scores. To show its usefulness, we test our method empirically using real-world data.

Introduction

Have you ever wondered why your paper—the one you've been so proud of—was rejected at some conference? Have you ever wondered why your other paper—the one that you thought is OK but not great—was accepted at the same conference? Many authors have experienced situations like those, and the reason is simple: The reviewing process for conferences (and, though perhaps to a lesser extent, also for scientific journals) is based on the reviewers' highly subjective preferences. Some reviewers may be more rigorous than others in evaluating submissions; some reviewers may like a particular field (or topic or approach or technique etc.) more than others; and so on.

Moreover, every reviewer usually evaluates only a small number of submissions, so his or her scores are based on partial information only. This may lead to a certain unfairness when looking from a global perspective at the scores for all submissions. For example, a reviewer evaluating three rather good and one excellent submission may tend to downgrade the good submissions in comparison with the excellent submission, whereas another reviewer who evaluates three rather good and one really bad submission may tend to upgrade the good submissions in comparison with the bad one. Nonetheless, the program committee members (or chairs) of the conference eventually have to reach a consensus as to which paper to accept and which to reject. That is, they need to aggregrate the scores of all submissions, which are given depending on the reviewers' subjective, partial-information preferences, in a way as equally and fair as possible.

However, this task is in fact not easy, and the outcome is not always as one might expect. Let us illustrate this using a concrete example from the (purely fictional) "*Third International Three Papers Get Accepted Conference*."

Example 1 *Consider a scenario with nine submissions, distributed among five reviewers. Let us assume the submissions are ordered according to their (unknown, objective, absolute) quality:*

$$S_{i+1} \succeq S_i, \quad i \in \{1,2,\ldots,8\}, \tag{1}$$

where $A \succeq B$ means A's quality is no worse than B's.

Each submission is assigned to exactly three reviewers. Each reviewer gives a score between 0 *and* 1 *to each assigned submission, see Table 1 for their scores in this example. Note that every reviewer orders the submissions he or she evaluates according to (1), by giving a higher score to a submission with a higher number. It would therefore be rational to have a final arrangement of the submissions according to this condition.*

However, observe that if we simply compute the arithmetic mean of the scores for each paper and order them accordingly, we would get:

$$S_9 \succeq S_8 \succeq S_6 \succeq S_7 \succeq S_4 \succeq S_5 \succeq S_2 \succeq S_1 \succeq S_3.$$

This arrangement contradicts the order of absolute paper quality, even though this order is strictly preserved in each

	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9
R_1	.40	.44		.67		.81	.89	.95	.99
R_2	.15	.19	.32	.45	.49	.62			
R_3		.21		.39		.61		.79	
R_4	.20		.20		.60		.70		.80
R_5			.10		.20		.30	.40	.50
average	.25	.28	.21	.50	.43	.68	.63	.71	.76

Table 1: Review scores for Example 1.

individual reviewer's scoring. The contradiction is a result of the reviewers' biases in assigning review scores, and of the inappropriateness of using a score average for global comparison in such a setting.

Tasks like this—aggregating individual preferences in a partial-information model—may occur in other contexts as well. In a more general setting, we are given a set of agents who each will give a score to some (but in general not all) of the given alternatives. Our goal then is to achieve "globally comparable" scores for all alternatives, based on the "local" (i.e., partial) scores of the agents that may be biased one way or the other. That is, assuming we have m agents and n alternatives, we are looking for a function g mapping the set of all $m \times n$ matrices with entries (i, j) (representing either agent i's score for alternative j—a rational number—or indicating that agent i does not evaluate alternative j) to $\mathbb{Q}^n$, where $\mathbb{Q}^n$ denotes the set of n-tuples of rational numbers. Given such a matrix M, $g(M) = \vec{z}$ is the global score vector we wish to compute. This mapping g should be as "fair" as possible from a global perspective, which means that better alternatives should receive higher scores. Of course, it is difficult to say what "better" actually means here or how one could formally define it. Intuitively stated, our goal is, for each alternative to, on the one hand, minimize the deviation of the desired global score from a presumed (objectively) "ideal" score for this alternative and, on the other hand, let the global score reflect the preferences of the individual agents who have evaluated this alternative.

For the sake of concreteness, however, and since this is the main application example motivating our study, we will henceforth focus on the particular task of aggregating the scores that the reviewers in a reviewing process assign to the submissions (i.e., we will not speak of agents and alternatives henceforth).

Our approach pursued here is to formulate a maximum likelihood estimator, which estimates both the "ideal" score of each submission and the bias of each individual reviewer. This estimator leads to a quadratic program, the solution of which essentially transforms the individual reviewers' scores into global scores. By fixing an acceptance threshold the program committee can then partition the submissions into those to be accepted and those to be rejected. Alternatively, by arranging the submissions according to their global scores, one could also simply rank them (in case one is even interested in a ranking).

Related Work

Preference aggregation is a wide field that has been intensely studied by various scientific communities, ranging from multiagent systems to computational social choice. The topic of this paper—aggregating the scores in peer reviewing—has also been investigated, although from different angles and using different methods. For example, Douceur (2009) encoded this aggregation problem into a corresponding problem on directed multigraphs and focuses on rankings (i.e., ordinal preferences) rather than ratings (i.e., cardinal preferences obtained by assigning review scores). By contrast, Haenni (2008) presents an algebraic framework to study the problem of aggregating individual scores. Our approach of using maximum likelihood estimators to formulate a quadratic program for solving this problem efficiently is, to the best of our knowledge, novel.

Our model is inspired by the offline time synchronization problem in broadcast networks, as discussed by Scheuermann et al. (2009). In that work, the problem of synchronizing timestamps in a set of event log files is addressed, where each log file has been generated with a different, potentially deviating local clock. However, our setting and our assumptions here differ in some central aspects. For example, while the time delays when an event is recorded in a log file in (Scheuermann et al. 2009) are (certainly reasonably) assumed to be always positive, the review score assigned by a reviewer may deviate in both directions. More specifically, the time delay when an event is recorded is assumed to be exponentially distributed, whereas we assume a Gaussian distribution. The resulting model is thus quite different: While Scheuermann et al. (2009) had to solve a specific linear program, we obtain a (semi-definite) quadratic program here.

Maximum likelihood estimators have been used in other contexts of preference aggregation as well. For example, Conitzer and Sandholm (2005), Conitzer, Rognlie, and Xia (2009), Xia, Conitzer, and Lang (2010), and Xia and Conitzer (2011) applied maximum likelihood estimation to model the "noise" in voting. Relatedly, Pini et al. (2009) study the issue of aggregating partially ordered preferences with respect to Arrovian impossibility theorems. Their framework differs from ours, however, as they consider ordinal preferences, whereas peer-reviewing is commonly based on scores, i.e., on cardinal preferences. Note that cardinal preferences are more expressive than ordinal preferences, as they also provide some notion of distance.

Model and Basic Assumptions

As mentioned above, we focus on a reviewing process as a special kind of preference aggregation with partial information. Our approach may also apply to other preference aggregation scenarios with the same or a similar structure.

In a common reviewing process, the reviewers not only comment on the weaknesses and strengths of the submission under review but also give an overall score. Although usually more information is requested from the reviewers (such as additional scores for criteria like "originality," "significance," "technical correctness," etc., plus a level of their own confidence in their expertise regarding this submission), we want to keep our model simple and thus focus on only the overall score a reviewer assigns to a submission. Furthermore, although scores are usually integer-valued (rarely, half points may be allowed), we do allow rational numbers as scores, thus obtaining a finer grained evaluation.

Let R be a set of m reviewers and S be a set of n submissions. Typically the submissions are distributed among the reviewers and each reviewer has to give a score to each submission assigned to him or her (usually just a small subset of S). Let $E \subseteq R \times S$ represent the set of review assignments, i.e., $(r,s) \in E$ if and only if reviewer r has evaluated submission s. Let this score be denoted by $e_{r,s}$. In order to normalize, we assume the scores to be rational numbers in the range $[0,1]$, where a higher score indicates a better quality.

Since different reviewers may assign different scores to the same submission, we need to find a way for how to make the decision as to whether this submission is accepted or rejected, based on the scores $e_{r,s}$ with $(r,s) \in E$.

We now proceed with introducing our model. Suppose a reviewer assigns a review score to a submission s, given that this submission has "absolute" or "ideal" quality z_s. Of course, z_s will be unknown in practice—our aim and approach later on will be to estimate z_s based on the assigned review scores.

Our stochastic model consists of two central components: a "random" deviation from the ideal, absolute score—essentially a form of "noise" disturbing the reviewer's "measurement" of the paper quality—and a systematic bias. We assume that the "noise" component is independent for individual paper assessments; it includes, e.g., misperceptions in either direction, but potentially also effects like strategic considerations of the reviewer with respect to the submission being evaluated. The systematic bias, by contrast, models the general rigor of the reviewer across all his or her reviews. It characterizes, for instance, whether the reviewer is lenient and assigns good scores even to very poor submissions, or whether this reviewer hardly ever gives a grade on the upper end of the scale.

Following these general lines, there is a random component for each review assignment, which we model by pairwise independent[1] Gaussian random variables $\Delta_{r,s}$ for reviewer r's assessment of submission s, with common variance $\sigma^2 > 0$ and mean $\mu = 0$.[2] Basically, the reviewer will then not assign a score based on the submission's objective quality z_s, but one based on his or her own noisy view of the quality, which is $z_s + \Delta_{r,s}$. This perceived quality will then be mapped to a review score, according to the respective reviewer's individual rigor and systematic bias. We model this by a linear function f_r, which means that the review score $e_{r,s}$ that reviewer r assigns to submission s is given by

$$e_{r,s} = f_r(z_s + \Delta_{r,s}) = p_r \cdot (z_s + \Delta_{r,s}) + q_r. \tag{2}$$

f_r is characterized by two reviewer-specific, unknown parameters p_r and q_r. Since it is reasonable to expect that each reviewer will generally tend to assign better grades to better submissions, we may assume that $p_r > 0$.

Even though this linear model is relatively simple, it allows to capture a wide range of reviewer characteristics—from very lenient up to very rigorous reviewers. The simplicity of our model is a feature we have chosen to put up with deliberately. One might of course add more model parameters to make the model more expressive; however, the number of parameters should be kept low to facilitate computational feasibility.

The Quadratic Programming Method

Our approach to aggregating preferences (that is, review scores) with partial knowledge is based on estimating the parameters z_s, p_r, and q_r for all submissions $s \in S$ and all reviewers $r \in R$ so as to maximize the likelihood of the assigned review scores. The first step to this end is to solve (2) for $\Delta_{r,s}$. Along with the substitutions $\overline{p}_r = 1/p_r$ and $\overline{q}_r = q_r/p_r$ this leads to

$$\Delta_{r,s} \quad = \quad e_{r,s} \cdot \overline{p}_r - \overline{q}_r - z_s. \tag{3}$$

In the following, we will consider $\overline{p}_r$ and $\overline{q}_r$ instead of p_r and q_r. These two representations are obviously equivalent and easily interchangeable, but the substituted variants are mathematically much more easily tractable. The Gaussian distribution with mean μ and variance σ^2 has probability density

$$D_{\mu,\sigma^2}(\Delta_{r,s}) \quad = \quad \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{(\Delta_{r,s} - \mu)^2}{2\sigma^2}\right).$$

Since we assume the $\Delta_{r,s}$ to be independent and $\mu = 0$, the overall probability density for all $\Delta_{r,s}$ for all review assignments in E is given by

$$\begin{aligned} & \prod_{(r,s)\in E} \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\frac{(\Delta_{r,s})^2}{2\sigma^2}\right) \\ = & \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{|E|} \cdot \exp\left(-\frac{\sum_{(r,s)\in E} (\Delta_{r,s})^2}{2\sigma^2}\right). \end{aligned}$$

We may now substitute according to (3) and obtain

$$\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{|E|} \cdot \exp\left(-\frac{\sum_{(r,s)\in E} (e_{r,s} \cdot \overline{p}_r - \overline{q}_r - z_s)^2}{2\sigma^2}\right).$$

The maximum likelihood estimate for the parameters z_s, $\overline{p}_r$, and $\overline{q}_r$ is the assignment of values $\widehat{z}_s$, $\widehat{p}_r$, and $\widehat{q}_r$ that maximizes this expression.

As usual, in order to find this maximum, we take the logarithms (because this will not affect the maximum). We thereby arrive at the problem of maximizing the expression

$$\begin{aligned} & |E| \ln \frac{1}{\sigma\sqrt{2\pi}} + \ln\exp\left(-\frac{\sum_{(r,s)\in E} (e_{r,s} \cdot \widehat{p}_r - \widehat{q}_r - \widehat{z}_s)^2}{2\sigma^2}\right) \\ = \quad & |E| \ln \frac{1}{\sigma\sqrt{2\pi}} - \frac{\sum_{(r,s)\in E} (e_{r,s} \cdot \widehat{p}_r - \widehat{q}_r - \widehat{z}_s)^2}{2\sigma^2}. \end{aligned}$$

[1] This assumption in particular requires our method to be only applied prior to reviewer discussions.

[2] Our approach would still be computationally tractable if we allowed for arbitrary, unknown means μ_r for each reviewer. However, we may assume, without loss of generality, that $\mu_r = 0$ for each $r \in R$ because any situation where $\mu_r \neq 0$ is equivalent to a case where $\mu_r = 0$ and an accordingly adjusted systematic bias component of the respective reviewer.

with respect to $\widehat{z}_s$, $\widehat{p}_r$, and $\widehat{q}_r$. Observe that the assignment for these variables maximizing the above expression does not depend on the value of σ. Maximizing the above expression is equivalent to minimizing

$$\sum_{(r,s)\in E} (e_{r,s} \cdot \widehat{p}_r - \widehat{q}_r - \widehat{z}_s)^2 . \quad (4)$$

Note that this is a quadratic optimization problem, which can be formulated as a so-called quadratic program, see, e.g., (Nocedal and Wright 2006). In general, a *quadratic program* (QP) is an optimization problem of the form:

$$\text{minimize} \quad \frac{1}{2}x^T Qx + c^T x \quad (5)$$

$$\text{subject to} \quad Ax \geq b, \quad (6)$$

where $x \in \mathbb{Q}^n$, $Q \in \mathbb{Q}^{n\times n}$, $c \in \mathbb{Q}^n$, $A \in \mathbb{Q}^{m\times n}$, and $b \in \mathbb{Q}^m$. The solution of a QP is a vector x that minimizes the expression in (5), simultaneously fulfilling all constraints in (6).

With respect to our specific QP as discussed so far, note that the solution would be trivial by setting all $\widehat{p}_r$, $\widehat{q}_r$, and $\widehat{z}_s$ to zero—which is obviously not a sensible result. If we assume the reviewers to be sort of rational on average, though, then we may require

$$\frac{1}{m}\sum_{r\in R} \widehat{p}_r \quad = \quad 1 \quad (7)$$

as a normalization constraint.

If we write the variables to estimate in a vector

$$x = (\widehat{z}_1 \quad \dots \quad \widehat{z}_n \quad \widehat{p}_1 \quad \dots \quad \widehat{p}_m \quad \widehat{q}_1 \quad \dots \quad \widehat{q}_m)^T ,$$

we obtain the following QP:

$$\text{minimize} \quad \frac{1}{2}x^T Qx$$

$$\text{subject to} \quad Ax \geq b$$

with a square matrix Q (see lines 2–13 of Algorithm 1), and a matrix A representing the normalization constraint (7). To have contraints of the form $Ax \geq b$ (instead of $Ax = b$) we simply "double" our normalization constraint (7) to

$$\frac{1}{m}\sum_{r\in R} \widehat{p}_r \geq 1 \qquad \text{and} \qquad -\frac{1}{m}\sum_{r\in R} \widehat{p}_r \geq -1.$$

A QP with a positive definite matrix Q has a unique solution and can be solved in polynomial time using interior-point methods, see, e.g., (Wright 1997). In our specific QP, the matrix Q is at least positive semi-definite,[3] because it can be written as $H \cdot H^T$ (see Algorithm 1 for the definition of matrix H). Unfortunately, though, it turns out that Q is *not* positive definite. So, let us have a closer look at why a unique solution actually cannot exist in general for our problem.

Observe that if all the absolute qualities z_s, $s \in S$, of the submissions were decreased by the same amount ξ, and the individual biases of each reviewer (parameters q_r, $r \in R$) were shifted accordingly, then we would end up with exactly the same review scores $e_{r,s}$ and thus an identical optimization problem. So, clearly, the fact that we do not have any global, absolute "reference" to which the overall scores could be adjusted results in an additional degree of freedom in the optimization, which prevents us from obtaining a unique maximum. In fact, a similar issue occurred in (Scheuermann et al. 2009), and along similar lines as there it is easy to overcome: For one arbitrarily picked reviewer r^* we may set $q_{r^*} = 0$, thus using this reviewer as a "fixed" reference point. The choice of r^* is not critical, as the optimization result will only be shifted and scaled accordingly. We suggest to eliminate this effect by normalizing the resulting scores to the interval $[0,1]$ in an additional step; then, all choices of r^* yield identical results, and the estimated scores $\widehat{z}_s$ are unique.

Yet, also with this modification it is *still* possible to come up with pathological instances where the solution is not unique. This lies in the nature of the problem: For instance, it is clearly impossible to estimate the absolute quality of a submission which did not receive at least one review. Similarly, it is impossible to compare the relative "rigor" of two groups of reviewers, if there is no paper that has been reviewed by at least one reviewer out of each of the two groups. In general, such ambiguities are easily identified and can always be resolved by introducing additional constraints as needed (or, alternatively, by assigning additional reviews). This then yields a positive definite matrix Q and consequently a unique solution of the QP.

To solve our QP, we can use existing solvers like, for example, MINQ (Neumaier 1998), a MATLAB script for "bound constrained indefinite quadratic programming." For given scores $e_{r,s}$ corresponding to the review assignments $(r,s) \in E$, Algorithm 1 illustrates our approach. We assume the scores to be nonnegative for line 5 to work (e.g., in $[0,1]$). Any negative number (e.g., -1) at position (r,s) in the input matrix M indicates that reviewer r did not review submission s. M thus encodes both E and the review scores $e_{r,s}$. Note that the resulting estimated scores in $\widehat{z}$ can exceed the interval of the input scores. This will, however, be overcome by subsequently scaling to results in $[0,1]$, as discussed above; this yields the scaled score estimates, in the following denoted by z^*_s, for all submissions $s \in S$.

	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9
average	.25	.28	.21	.50	.43	.68	.63	.71	.76
z^*	0	.05	.18	.40	.53	.68	.75	.92	1

Table 2: Average and normalized scores for Example 2.

Example 2 *Let us now continue Example 1 from the introduction. We use the review scores from Table 1 and build a QP as explained above. Solving this problem with MINQ, we are able to compute a solution containing the estimates $\widehat{z}_s$ in addition to the parameters $\widehat{p}_r$ and $\widehat{q}_r$, for each $s \in \{S_1,\dots,S_9\}$ and $r \in \{R_1,\dots,R_5\}$. The resulting scaled estimates z^*_s are given in Table 2 (along with the average review scores from Table 1 above, for convenience). Indeed we notice that according to the normalized review scores obtained by using our method, the expected ranking of the*

[3] A matrix $A \in \mathbb{Q}^{n\times n}$ is said to be *positive definite* if all eigenvalues of A are positive. A is said to be *positive semi-definite* if its eigenvalues are nonnegative.

Algorithm 1 Computing the estimated scores

Input: $M \in \mathbb{Q}^{m \times n}$ *// M contains the given scores.*
$H = [0] \in \mathbb{Q}^{(2m+n)\times(m \cdot n)}$
for $j \in \{1, 2, \ldots, m\}$ **do**
  for $k \in \{1, 2, \ldots, n\}$ **do**
    if $M_{(j,k)} \geq 0$ **then**
      $H_{(k,(k-1)\cdot m+j)} = 1$
      $H_{(n+j,(k-1)\cdot m+j)} = -M_{(j,k)}$
      $H_{(n+m+j,(k-1)\cdot m+j)} = 1$
    end if
  end for
end for
remove the last row from H *// normalization*
$Q = 2 \cdot H \cdot H^T$
$h_1 = (0 \quad \cdots \quad 0) \in \mathbb{Q}^n$
$h_2 = (1 \quad \cdots \quad 1) \in \mathbb{Q}^m$
$h_3 = (0 \quad \cdots \quad 0) \in \mathbb{Q}^{m-1}$
$A = \begin{bmatrix} h_1 & \frac{1}{m}\cdot h_2 & h_3 \\ h_1 & -\frac{1}{m}\cdot h_2 & h_3 \end{bmatrix}$
$b = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$
solve: $\min \frac{1}{2}x^T Q x$ subject to $Ax \geq b$
$\widehat{z} = (x_1 \quad \cdots \quad x_n)^T$
Output: $\widehat{z} \in \mathbb{Q}^n$

submissions is preserved. It is now trivial to identify and accept the three top submissions to be presented at the 3rd International Three Papers Get Accepted Conference.

Discussion

In the example above we saw that our proposed algorithm was able to reconstruct the global order according to the "absolute" paper qualities. This raises the question of whether this property holds in general. Apparently, if we permit an arbitrary level of "noise" in the reviews—in the form of very large $\Delta_{r,s}$ values—, then this "noise" will at some point dominate the "signal," and submissions may switch order with respect to their absolute qualities. So, we should slightly reformulate this question: If the noise is sufficiently small, will the order of the global paper qualities z_s be preserved in the estimated global scores $\widehat{z_s}$?

To approach this question, let us first consider the case where there is no noise at all, i.e., where $\Delta_{r,s} = 0$ for all $(r,s) \in E$. For simplicity, let us also assume that there is a unique solution to the optimization problem (as argued above, this can always be achieved). Since all $\Delta_{r,s}$ are zero, it holds that

$$\forall (r,s) \in E : e_{r,s} = p_r z_s + q_r. \tag{8}$$

Thus, the objective function (4) of the QP is equal to zero if the estimated values $\widehat{p}_r$, $\widehat{q}_r$, and $\widehat{z_s}$ are each equal to the corresponding correct values $\overline{p}_r$, $\overline{q}_r$, and z_s. Since the objective function is always nonnegative, this value is optimal. Since we assume a unique optimum, our estimator will therefore correctly determine both all absolute paper qualities and the parameters of all reviewers' bias functions.[4]

Now consider any two submissions, s_1 and s_2, which both have been evaluated by the same reviewer r. Since we assumed that all $\Delta_{r,s}$ values are zero, this reviewer's scores are given by (8). Since $p_r > 0$, these scores represent the same order as the absolute qualities z_{s_1} and z_{s_2}, i.e., $e_{r,s_1} < e_{r,s_2}$ if and only if $z_{s_1} < z_{s_2}$. As this will hold for *any* reviewer in the setting without "noise," and since the estimated quality levels $\widehat{z_s}$ for all submissions are then, as argued above, equal to the correct values z_s, it holds that the local order of all reviewer scores is preserved in the resulting global ranking of the submissions.

Let us now consider the case where the $\Delta_{r,s}$ are not all equal to zero. Observe that each $e_{r,s}$ is a continuous function of $\Delta_{r,s}$ (refer to (2)), and that the objective function of the QP is a continuous function of all $e_{r,s}$, and thus also of all $\Delta_{r,s}$. Intuitively speaking, a small change of the $\Delta_{r,s}$ values will therefore also cause only a small change in the solution of the QP. That is, the point where the objective function is minimal will "move" continuously with a changed input. As a result, the estimates $\widehat{z_s}$ are also a continuous function of the $\Delta_{r,s}$.

More formally, let $\varepsilon = \min\{|z_{s_1} - z_{s_2}| : s_1, s_2 \in S, s_1 \neq s_2\}$. Then, as long as no estimated review score deviates by more than $\varepsilon/2$ from the corresponding correct value, there cannot be a change in order. Since the estimates are, as stated before, a continuous function of the $\Delta_{r,s}$, it follows from the definition of continuity that there exists some $\delta > 0$ (depending on ε) such that if $|\Delta_{r,s}| < \delta$ for all $(r,s) \in E$ then $|z_s - \widehat{z_s}| < \varepsilon/2$ for all $s \in S$. Consequently, if the $\Delta_{r,s}$ do not become too large, the estimated scores $\widehat{z_s}$ will not deviate far enough from the correct values z_s to cause a change in the resulting ranking of the submissions.

A Case Study

We evaluated data from the "*Third International Workshop on Computational Social Choice*" (COMSOC-2010) that took place in September 2010 in Düsseldorf, Germany (Conitzer and Rothe 2010). There were 57 submissions (where submissions that had to be rejected on formal grounds are disregarded) and 21 reviewers. Every submission was reviewed by at least two reviewers; a third reviewer was assigned to some submissions later on. The results are shown in Table 3 for accepted submissions and in Table 4 for rejected submissions. In each row of both tables, (a) the first column gives the number of submissions that received the same final score, resulting as the average of the single reviewers' overall scores for this submission (weighted by the reviewers' confidence level); (b) the second and third columns give this final average score and the corresponding rank of the paper; (c) the fourth and fifth columns give the score(s) and rank(s) our algorithm produces based on the same overall scores of the reviewers.

Scores are here assumed to be integers in the range $[-3, 3]$, so when applying our method we also re-normalized

[4]For simplicity, we neglect the effects of the normalization here, which, as argued above, may result in a scaling and shifting which applies equally to all reviewers and scores.

#	PC decision score	PC decision rank	our approach score(s)	our approach rank(s)
4	3.0	1	2.79, 2.79, 2.48, 2.41	3, 4, 6, 10
3	2.7	5	2.50, 2.44, 1.97	5, 8, 16
4	2.5	8	3.00, 2.45, 2.41, 1.90	1,7, 9,17
1	2.4	12	2.21	14
1	2.3	13	2.37	12
9	2.0	14	2.38, 2.23, 1.85, 1.65, 1.56, 1.55, 1.03, 0.99, 0.70	11, 13, 18, 20, 21, 23, 30, 31, 35
2	1.8	23	2.81, 2.06	2, 15
4	1.6	25	1.76, 1.35, 1.31, -0.03	19, 26, 28, 39
2	1.5	29	1.48, 0.70	25, 34
5	1.4	31	1.56, 1.50, 1.04, -0.25, -0.29	22, 24, 29, **44**, **46**
1	1.3	36	1.33	27
2	1.0	37	0.75, 0.55	33, 36
1	0.9	39	0.88	32
1	0.7	40	0.40	38

Table 3: Conference data: accepted papers.

the results to that range. Using an acceptance threshold of 0.6, a total of 40 submissions were accepted and 17 were rejected by the program committee (PC). As one can see, our method provides a different ranking of the papers. In particular, if the PC again accepted 40 submissions, according to the ranking resulting from our method two originally accepted submissions would now be rejected (see the boldfaced entries in Table 3), whereas two originally rejected submissions would now be accepted (see the boldfaced entries in Table 4).

#	PC decision score	PC decision rank	our approach score(s)	our approach rank(s)
1	0.5	41	-0.12	42
1	0.3	42	-0.27	45
1	0.0	43	0.46	**37**
1	-0.1	44	-0.09	**40**
1	-0.2	45	-0.16	43
3	-0.6	46	-0.47, -0.59, -1.05	47, 48, 50
5	-1.0	49	-0.11, -0.77, -1.35, -1.42, -1.49	41, 49, 52, 53, 54
1	-1.5	54	-2.37	56
1	-2.4	55	-3.00	57
1	-2.7	56	-1.17	51
1	-3.0	57	-2.12	55

Table 4: Conference data: rejected papers.

Conclusions

We have presented a novel method—using maximum likelihood estimation and quadratic programming—to calibrate the scores of potentially biased, partially blindfolded reviewers in peer reviewing that is arguably superior to the currently common method of simply taking the average of the individual reviewers' scores. We have discussed some critical points in applying our method and proposed ways of how to handle them, and we have applied it empirically using real-world data. An interesting task for future research is to compare our method, analytically and empirically, to other mechanisms of preference aggregation in a partial-information model. Also, it would be nice to specify theoretical properties a ranking method like ours should satisfy and then compare such methods based on these properties.

Acknowledgments

We thank the reviewers for their helpful comments and Dietrich Stoyan for interesting discussions. This work was supported in part by DFG grant RO 1202/12-1 in the European Science Foundation's EUROCORES program LogICCC.

References

Conitzer, V., and Rothe, J., eds. 2010. *Proceedings of the Third International Workshop on Computational Social Choice*. Universität Düsseldorf.

Conitzer, V., and Sandholm, T. 2005. Common voting rules as maximum likelihood estimators. In *Proceedings of the 21st Annual Conference on Uncertainty in Artificial Intelligence*, 145–152. AUAI Press.

Conitzer, V.; Rognlie, M.; and Xia, L. 2009. Preference functions that score rankings and maximum likelihood estimation. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, 109–115. IJCAI.

Douceur, J. 2009. Paper rating vs. paper ranking. *ACM SIGOPS Operating Systems Review* 43:117–121.

Haenni, R. 2008. Aggregating referee scores: An algebraic approach. In Endriss, U., and Goldberg, P., eds., *Proceedings of the 2nd International Workshop on Computational Social Choice*, 277–288. University of Liverpool.

Neumaier, A. 1998. MINQ – general definite and bound constrained indefinite quadratic programming. WWW document. http://www.mat.univie.ac.at/˜neum/software/minq.

Nocedal, J., and Wright, S. 2006. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, 2nd edition.

Pini, M.; Rossi, F.; Venable, K.; and Walsh, T. 2009. Aggregating partially ordered preferences. *Journal of Logic and Computation* 19(3):475–502.

Scheuermann, B.; Kiess, W.; Roos, M.; Jarre, F.; and Mauve, M. 2009. On the time synchronization of distributed log files in networks with local broadcast media. *IEEE/ACM Transactions on Networking* 17(2):431–444.

Wright, S. 1997. *Primal-Dual Interior-Point Methods*. SIAM.

Xia, L., and Conitzer, V. 2011. A maximum likelihood approach towards aggregating partial orders. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*. IJCAI. To appear.

Xia, L.; Conitzer, V.; and Lang, J. 2010. Aggregating preferences in multi-issue domains by using maximum likelihood estimators. In *Proceedings of the 9th International Joint Conference on Autonomous Agents and Multiagent Systems*, 399–408. IFAAMAS.