A Statistical Approach to Calibrating the Scores of Biased Reviewers: The Linear vs. the Nonlinear Model[1]

Magnus Roos[2] and **Jörg Rothe**[2] and **Joachim Rudolph**[3] and **Björn Scheuermann**[4] and **Dietrich Stoyan**[5]

Abstract. Two methods are proposed for aggregating the scores of reviewers in a peer-reviewing rating system. Both methods are of a statistical nature. The simpler method, which is based on a classical statistical approach from the field of linear models, uses the analysis of variance and can thus be realized by means of existing statistical software. The more advanced method, which is a slight modification of the method proposed by Roos et al. [13], uses a nonlinear model and numerical optimization based on a least-squares approach. Under reasonable statistical assumptions, both approaches—the linear and the nonlinear one—can be seen as using the maximum likelihood principle. Application of either method implies also an evaluation of the reviewers. An application example with real conference data shows the power of the statistical methods, compared with the common naive approach of simply taking the average scores.

1 Introduction

Evaluation of persons, papers, products, etc. is a fundamental social activity. For example, students are evaluated by teachers, scientific papers by journal/conference reviewers, and sportsmen by referees, e. g., in figure skating and gymnastics. Even if all reviewers in a rating system are subjectively fair, some of them may be biased and produce scores systematically too high or too low. If then not all objects are reviewed by all reviewers, it becomes complicated to aggregate the scores given to the same objects in a fair way.

The present paper focuses on the problem of ranking scientific papers submitted to conferences, where usually the relative number of reviews per paper is small. The common procedure applied by popular conference management systems such as EasyChair[6] and ConfMaster[7] is described as (quoting from the EasyChair website): "When computing the average score, weight reviews by reviewer's confidence." This means that all scores given to a paper are simply averaged, possibly weighted by reviewer-specific weights, the confidence levels of the reviewers, which again are very subjective because every reviewer evaluates only him- or herself. Under these conditions it may happen that by good luck a weak scientific paper goes to some lenient or generous reviewers, whereas a good paper goes to a harsh reviewer and some normal reviewers. Then the weak paper might be accepted, but the good one is rejected.

The present paper aims to improve the common "naive" (as Lauw et al. [9] call it) approach where the overall scores of all objects are obtained by simply averaging all given scores of the object. Of course, paper scores can only provide some guidance on paper acceptance; the final decision is usually made on deeper considerations.

It is assumed here that external information about the reviewers is not used, such as weighting the scores. There is also no separate "training" phase in order to characterize the reviewers' tendencies. Instead, the proposed methods apply cross-classification techniques to determine the characteristics of both the reviewers and the judged objects simultaneously in one step. All reviewers are assumed to be "honest," to exercise their best judgments, without any personal relation to certain objects. Nevertheless, some reviewers may be biased in giving systematically high or low scores. As long as all papers are evaluated by all reviewers, this is not an obstacle to fair score aggregation by averaging. However, if there are only a few reviews per paper, problems are likely to arise. The following toy example taken from [8] shows what can happen.

Example 1 *Consider the data in Table 1. There are five reviewers (r_i) and five papers (p_j). The original scores y_{ij} from [8] are here multiplied by* 10 *and are thus in the range from 0 to 10. Consisting of only* 15 *scores in total, this data set is very small.*

Table 1. Data for a toy example taken from [8].

	p_1	p_2	p_3	p_4	p_5
r_1	6	6	6	–	–
r_2	3	–	–	4	–
r_3	3	–	–	–	4
r_4	–	3	3	4	4
r_5	–	3	3	4	4

The naive approach results in the same average score of 4.0 *for all five papers. This seems to be highly questionable: in their preliminary discussion, Lauw et al. [8] point out that reviewer r_1 is very likely to be lenient, causing too high aggregated scores for papers p_1, p_2, and p_3. In Section 2.2, this example is to be continued to show the results that can be obtained by means of statistical methods.*

Related Work

Preference aggregation is a wide field that has been intensely studied by various scientific communities, ranging from multiagent systems

[1] This work was supported in part by DFG grants RO 1202/12-1 and RO 1202/15-1, the European Science Foundation's EUROCORES program LogICCC, an SFF grant of HHU Düsseldorf, ARC grant DP110101792, and a DAAD-PPP grant in the PROCOPE project.

[2] Institut für Informatik, Heinrich-Heine-Universität Düsseldorf, 40225 Düsseldorf, Germany, email: {roos, rothe}@cs.uni-duesseldorf.de

[3] Institut für Sozialwissenschaften, Humboldt-Universität zu Berlin, 10099 Berlin, Germany

[4] Institut für Informatik 4, Universität Bonn, 53113 Bonn, Germany, email: scheuermann@cs.uni-bonn.de

[5] Institut für Stochastik, TU Bergakademie Freiberg, 09596 Freiberg, Germany, email: stoyan@math.tu-freiberg.de

[6] `http://www.easycair.org`

[7] `http://www.confmaster.net`

to computational social choice. The topic of this paper—aggregating the scores in reviewing scientific papers—has also been investigated, although from other angles and using different methods. For example, Douceur [4] encoded the 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 [6] presents an algebraic framework to study the problem of aggregating individual scores.

The present paper uses methods of analysis of variance from the field of statistics, see [7]. The setting is called *two-way classification* there, where one "way" relates to reviewers and the other to papers. This classical statistical approach from the field of linear models is adapted here. This leads to fairer overall scores for the papers, where "fairer" in a technical sense refers to the fact that the proposed method leads to unbiased estimators for certain model parameters (see Section 2.2 for details). At the same time in parallel, the method also allows for an evaluation of the reviewers.

The papers by Lauw et al. [8, 9] tackle the same problem as the present paper, yet with quite a different approach. They apply a so-called "differential model," which is an ad-hoc nonlinear model. Their model includes an unknown model parameter α, which appears not to be statistically estimable. No random errors occur in this model, although in real review processes such effects are well conceivable to play a role.

We will first present the simple linear approach in Section 2.2. It can be realized by existing statistical software. This approach is then refined in Section 2.3 by a nonlinear method, which applies techniques from quadratic programming. Under some statistical assumptions, both approaches—the linear and the nonlinear one—can be seen as using the maximum likelihood principle.

The nonlinear model is inspired by a solution to the offline synchronization problem in broadcast networks, as discussed by Scheuermann et al. [14]. 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. "Reviewers" in the present paper take the role of "network nodes" there, the role of "papers" here corresponds to "network packet transmissions" there, and "review scores" here are in line with "reception timestamps" there. However, the setting and assumptions in Scheuermann et al. differ in some central aspects. In particular, random packet reception time delays, which correspond to random components in review scores, follow exponential distribtions and are not Gaussian. More technically, the resulting optimization problem is linear in [14], while it is (semi-definite) quadratic here.

There is a significant body of existing work in the area of preference aggregation, i. e., on the question how to aggregate individual preferences into a common, global ranking. Some of these works use related estimators in different settings. For example, Conitzer and Sandholm [3], Conitzer, Rognlie, and Xia [1], and Xia et al. [18, 17] apply maximum likelihood estimation to model the "noise" in voting. Relatedly, Pini et al. [12] study the issue of aggregating partially ordered preferences with respect to Arrovian impossibility theorems. However, their framework differs from the model used here: 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 a notion of distance.

2 Models

2.1 Basic Assumptions

In the reviewing process considered, reviewers not only comment on the weaknesses and strengths of the papers, but give a score to each paper reviewed. The following analysis focuses on only the scores. These scores are assumed to be integers, to which situation most evaluation processes can be transformed, even if decimal numbers with one or two decimals are given. High scores mean good quality.

There are I reviewers r_i and J papers p_j. For each pair (i, j), there exists a binary number e_{ij}, where $e_{ij} = 1$ means that reviewer r_i reviews paper p_j, and $e_{ij} = 0$ otherwise. The matrix $(e_{ij})_{1\le i\le I, 1\le j\le J}$ is called *incidence matrix*. Let $E = \{(i, j) \mid e_{ij} = 1\}$. The scores corresponding to pairs $(i, j) \in E$ are denoted by y_{ij}.

2.2 The Linear Model

Adapting the classical statistical linear modeling approach, the following model is used:

$$y_{ij} = \mathscr{D}\left(\mu + \alpha_i + \beta_j + \varepsilon_{ij}\right) \quad \text{for } (i, j) \in E. \tag{1}$$

Here, $\mathscr{D}$ is a discretization operator that transforms any real number x into the integer score $\mathscr{D}(x)$. The other symbols have the following meanings:

- μ is the overall mean of all scores given,
- α_i is the mean difference between the scores of reviewer r_i and μ,
- β_j is the mean difference between the scores of paper p_j and μ,
- ε_{ij} is a random error for $(i, j) \in E$.

The α_i are closely related to the "leniencies" of reviewers discussed by Lauw et al. [8, 9], and the β_j to their paper "qualities." The idea is that reviewer r_i does not assign a score to paper p_j based on its true quality β_j (which r_i does not know), but based on r_i's own noisy view of p_j's quality, which is $\beta_j + \varepsilon_{ij}$. This judgment is then linearly shifted according to the reviewer's "leniency". Simplifying more general models, it is assumed that there is no interaction between reviewers and papers (which, if desired, could be expressed by parameters $(\alpha\beta)_{ij}$).

The strategy in the following is to ignore the discretization in the statistics and to assume that the discretized data belong to the truly linear model

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad \text{for } (i, j) \in E. \tag{2}$$

with

$$\mathbf{E}\,\varepsilon_{ij} \equiv 0 \quad \text{and} \quad \mathbf{var}\,\varepsilon_{ij} \equiv \sigma^2 \quad \text{for } (i, j) \in E, \tag{3}$$

where the ε_{ij} are independent and **E** and **var** denote the expectation and variance, respectively. The error of this simplified approach will be discussed in the full version of this paper.

Model (2) is called *two-way classification* in the analysis of variance, see, e. g., the book by Draper and Smith [5].

As mentioned above, the naive estimators of the sums $\mu + \beta_j$, here denoted by $\widehat{\mu + \beta_j}$, are the averages of all review scores assigned to the respective paper:

$$\widehat{\mu + \beta_j} = \overline{y}_{*j} = \frac{1}{n_{*j}} \sum_{i:(i,j)\in E} y_{ij}, \tag{4}$$

where n_{*j} is the number of reviews for paper p_j. No serious statistician will use them, since these estimators are not unbiased and better estimators are possible.

Theory says that only the differences of the effects α_i and β_j can be estimated without bias. Fortunately, for the problem of ranking papers it completely suffices to have estimates of the differences $\beta_j - \beta_1$. And for evaluating the reviewers, estimates of the differences $\alpha_i - \alpha_1$ are fully sufficient. Thus, one may assume that

$$\sum_{i=1}^{I} \alpha_i = 0 \qquad \text{and} \qquad \sum_{j=1}^{J} \beta_j = 0. \tag{5}$$

In many statistical textbooks such as [5] and [15], it is assumed that for each pair (i, j) a fixed, strictly positive number n of observations is given (where, in typical settings, $n \gg 1$). If so, least-squares estimates of μ, α_i, and β_j are easy to determine. They directly follow from the means

$$\overline{y}_{**} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}, \qquad \overline{y}_{i*} = \frac{1}{J}\sum_{j=1}^{J} y_{ij}, \quad \text{and} \quad \overline{y}_{*j} = \frac{1}{I}\sum_{i=1}^{I} y_{ij}$$

as $\mu = \overline{y}_{**}$, $\alpha_i = \overline{y}_{i*} - \overline{y}_{**}$, and $\beta_j = \overline{y}_{*j} - \overline{y}_{**}$. These estimators are unbiased. In this case, the naive approach is the best. However, in the situation typical for peer reviewing, the "observation" counts n_{ij} are 0 (reviewer i does not review paper j) or 1 (reviewer i reviews paper j). (Note that $n_{ij} = 2$ would mean that reviewer i reviews paper j twice, independently.) We are confronted with a so-called "incomplete" (and "unbalanced") experimental design. The corresponding theory is described by Koch [7, Sections 3.4.2 and 3.4.3]. The case of interest here is there referred to as *two-way cross-classification*.

The parameters are estimated by the least-squares approach, i. e., the sum over all

$$(y_{ij} - \mu - \alpha_i - \beta_j)^2$$

is minimized. To this end, Koch [7] describes numerical approaches based on normal equations. Standard statistical software offers various ways to obtain estimators of the α_i, the β_j, and of μ, which differ in the so-called reparametrization conditions.

The model variance σ^2 is estimated by the mean squared error, which is the sum of quadratic deviations $(y_{ij} - \hat{y}_{ij})^2$ with $\hat{y}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j$ divided by their number minus one. The estimators obtained are unbiased and in some sense "best." In the case of normally distributed ε_{ij}, the least-squares estimators are also maximum likelihood estimators.

For the practical statistical analysis, the statistical software package IBM-SPSS Statistics 20 (which we abbreviate by SPSS), procedure UNIANOVA, was used. The procedure UNIANOVA does not use the conditions (5), but it is preset such that α_I and β_J are set to zero in the model discussed here.

Alternatively, also the program that will be mentioned in the next section (see Algorithm 1) can be used by setting $\gamma_i = 1$ in (6) below, which leads to (1). Both programs yield identical results.

The parameters determined by SPSS can easily be transformed into the parameters μ, α_i, and β_j. Simulations and direct calculation of model parameters are easily possible based on the matrix module of SPSS.

Example 2 (continuing Example 1) *Table 2 shows the values for the parameters in the linear model; the parameter μ is estimated as* 4.0*. The model parameters indicate that reviewer r_1 indeed has to be considered as lenient, while the other reviewers are estimated to have the same degree of rigor. The papers are now divided into two classes: p_1, p_2, and p_3 seem to be weaker papers with lower scores, while the other two papers appear to be of the same higher quality. It cannot surprise that Lauw et al. [8] arrive at the same conclusions for this extremely simple example.*

Table 2. Parameters for the toy example from [8].

$i,\ j$	α_i	β_j
1	2.4	−0.4
2	−0.6	−0.4
3	−0.6	−0.4
4	−0.6	0.6
5	−0.6	0.6

Note that in the example above, the estimated parameter values exactly reproduce the scores from Table 1 when used in (2) with all $\varepsilon_{ij} = 0$. Essentially, this means that no random deviations at all are necessary to "explain" the reviewers' scores. Therefore, this example has to be considered extremely simple.

2.3 The Nonlinear Model

The linear model from the previous section is now refined to a nonlinear model, which modifies the method proposed by Roos et al. [13] so as to generalize (1) to

$$y_{ij} = \mathscr{D}\left(\mu + \gamma_i(\alpha_i + \beta_j + \varepsilon_{ij})\right) \quad \text{for } (i,j) \in E \tag{6}$$

with positive parameters γ_i. For the special case of $\gamma_i \equiv 1$, (6) coincides with (1). The term $\gamma_i(\alpha_i + \beta_j + \varepsilon_{ij})$ models the interaction between reviewer r_i and paper p_j; γ_i is a proportionality factor; and μ, α_i, β_j, and ε_{ij} have the same meaning as in the linear case.

Reviewer r_i's perceived, noisy quality level $\beta_j + \varepsilon_{ij}$ is, just like in the linear model, added to this reviewer's systematic bias α_i. In addition, though, the result is transformed by multiplication with the reviewer-specific scaling factor γ_i. This factor models r_i's individual rigor: in essence, γ_i describes by how much reviewer i's review score changes, given a fixed change in (perceived) paper quality.

Even though this nonlinear model is relatively simple, it allows to capture a wide range of reviewer characteristics.

An assumption similar to $\sum_{i=1}^{I} \alpha_i = 0$ in the linear case (see Section 2.2) is now done by

$$\alpha_I \quad = \quad 0. \tag{7}$$

This leads to a problem slightly smaller than that with $\sum_{i=1}^{I} \alpha_i = 0$. Both restrictions are possible and plausible, and the results can simply be transformed to each other by choosing a suitable parameter μ. The aim is to estimate the parameters α_i, β_j, γ_i, and μ. Again the least-squares approach is used, which minimizes the sum of squared errors ε_{ij},

$$\sum_{(i,j)\in E} \left(\frac{y_{ij}}{\gamma_i} - \frac{\mu}{\gamma_i} - \alpha_i - \beta_j \right)^2. \tag{8}$$

Since this does not affect the optimization itself, in this setting μ can be set to zero. After getting the result, one may shift the values so that a condition like $\sum_{j=1}^{J} \beta_j = 0$ as in (5) is fulfilled. It is easy to see that the resulting parameter estimators are maximum likelihood estimators if the errors ε_{ij} are i. i. d. Gaussian as in (3).

Numerically, the minimization procedure is carried out by means of a direct optimization program such as a so-called quadratic program, see, e. g., the book by Nocedal and Wright [11]. In general, a *quadratic program* (QP) is an optimization problem of the form:

$$\text{minimize} \qquad \frac{1}{2} x^T Q x + c^T x \tag{9}$$

$$\text{subject to} \qquad Ax \geq b, \tag{10}$$

where (letting $\mathbb{Q}$ denote the set of rational numbers) $x \in \mathbb{Q}^n$, $Q \in \mathbb{Q}^{n\times n}$, $c \in \mathbb{Q}^n$, $A \in \mathbb{Q}^{m\times n}$, and $b \in \mathbb{Q}^m$. The solution of a QP is a vector x that minimizes the expression in (9), simultaneously fulfilling all constraints in (10).

With the simple substitution $\tilde{\gamma}_i = 1/\gamma_i$ in (8) one obtains

$$\sum_{(i,j)\in E} \left(y_{ij}\tilde{\gamma}_i - \mu\tilde{\gamma}_i - \alpha_i - \beta_j\right)^2, \tag{11}$$

which can be transformed into the form of a QP as required by (9) and (10). In the following, the estimators of α_i, β_j, and $\tilde{\gamma}_i$ are denoted by $\hat{\alpha}_i$, $\hat{\beta}_j$, and $\hat{\gamma}_i$. With respect to the QP discussed so far, note that a trivial solution can be achieved by setting $\hat{\gamma}_i$, $\hat{\alpha}_i$, and $\hat{\beta}_j$ each to zero, which clearly is not reasonable. Assuming typical reviewers to be "rational", one may require the normalization constraint:

$$\frac{1}{I}\sum_{i=1}^{I}\hat{\gamma}_i \quad = \quad 1. \tag{12}$$

Defining a vector $x = \left(\hat{\beta}_1,\ldots,\hat{\beta}_J,\hat{\gamma}_1,\ldots,\hat{\gamma}_I,\hat{\alpha}_1,\ldots,\hat{\alpha}_I\right)^T$, containing the variables to estimate, one obtains the QP:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}x^T Q x \\ \text{subject to} & Ax \geq b \end{array} \tag{13}$$

with a square matrix Q (see lines 2–13 of Algorithm 1 below), and a matrix A representing the normalization constraint (12).

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., [16]. In this specific QP, the matrix Q is at least positive semi-definite, i. e., all eigenvalues of A are nonnegative, because it can be written as $H \cdot H^T$ (see Algorithm 1 below for the definition of matrix H). Analogously to the linear model in Section 2.2, one does not have any global, absolute "reference" to which the overall scores could be adjusted. This leads to an additional degree of freedom in the optimization, which precludes obtaining a unique maximum. In fact, a similar issue also occurred in the work of Scheuermann et al. [14], and along similar lines as there it is easy to overcome: one may set $\alpha_I = 0$, thus using one reviewer as a "fixed" reference point. In this paper, the last reviewer is picked for this constraint, see Equation (7). 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 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 the resulting QP, one can use existing solvers such as MINQ [10], a MATLAB script for bound constrained indefinite quadratic programming. Algorithm 1 illustrates this approach. The scores y_{ij} for $(i,j) \in E$ are assumed to be nonnegative for line 5 to work. Any negative number (e. g., -1) at position (i,j) in the input matrix M indicates that reviewer r_i did not review submission p_j (i. e., $(i,j) \notin E$). M thus encodes both E and the review scores y_{ij}. Note that the resulting estimated scores in $\hat{\beta}$ may exceed the interval of the input scores. This can, however, be overcome by subsequently scaling to results as desired, as discussed above; this yields the scaled score estimates, in the following denoted by β_j^*, for all submissions.

Algorithm 1 Computing the estimated scores

1: **Input:** $M \in \mathbb{Q}^{m\times n}$ *// M contains the given scores.*
2: $H = \begin{bmatrix}0\end{bmatrix} \in \mathbb{Q}^{(2m+n)\times(m\cdot n)}$
3: **for** $j \in \{1,2,\ldots,m\}$ **do**
4: **for** $k \in \{1,2,\ldots,n\}$ **do**
5: **if** $M_{(j,k)} \geq 0$ **then**
6: $H_{(k,(k-1)\cdot m+j)} = 1$
7: $H_{(n+j,(k-1)\cdot m+j)} = -M_{(j,k)}$
8: $H_{(n+m+j,(k-1)\cdot m+j)} = 1$
9: **end if**
10: **end for**
11: **end for**
12: remove the last row from H *// normalization*
13: $Q = 2 \cdot H \cdot H^T$
14: $h_1 = \begin{pmatrix}0 & \cdots & 0\end{pmatrix} \in \mathbb{Q}^n$
15: $h_2 = \begin{pmatrix}1 & \cdots & 1\end{pmatrix} \in \mathbb{Q}^m$
16: $h_3 = \begin{pmatrix}0 & \cdots & 0\end{pmatrix} \in \mathbb{Q}^{m-1}$
17: $A = \begin{bmatrix} h_1 & \frac{1}{m}\cdot h_2 & h_3 \\ h_1 & -\frac{1}{m}\cdot h_2 & h_3 \end{bmatrix}$
18: $b = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$
19: solve: $\min \frac{1}{2}x^T Q x$ subject to $Ax \geq b$
20: $\hat{\beta} = \begin{pmatrix}x_1 & \cdots & x_n\end{pmatrix}^T$
21: **Output:** $\hat{\beta} \in \mathbb{Q}^n$

3 A Case Study

The following discusses data from the *Third International Workshop on Computational Social Choice* (COMSOC-2010) that took place in September 2010 in Düsseldorf, Germany [2]. There were 57 submissions (where submissions that had to be rejected on formal grounds are disregarded) and 20 reviewers. Every submission was reviewed by at least two reviewers; a third reviewer was assigned to some submissions later on, and one paper was even reviewed by four reviewers. (The fact that these extra reviews were somehow related to the evaluation of the papers in the first two reports is ignored in the following.) Table 3 shows the data, the results of the reviewing process. It contains the scores given by the reviewers to the papers, where "–" means "no review." As is common in EasyChair, the scores were integers between -3 and 3, which are here shifted to the integers between 1 and 7, where 7 is the best possible score.

Table 4 shows the main results of applying the methods proposed in this paper to real conference data: the estimated COMSOC-2010 paper scores obtained by the two approaches presented here, which are closely related to the β_j. The acceptance threshold of the conference was around 4.5, based on the naive approach. This led to acceptance of a total of 40 submissions, while 17 were rejected.

Table 5 shows the parameters α_i and γ_i of the reviewers, which allow to evaluate them as well. This is simpler in the linear than in the nonlinear approach. According to the linear approach, reviewer 7 with $\alpha_7 = 2.3662$ is the most lenient reviewer. In the nonlinear approach, the relatively large value of $\gamma_7 = 6.1283$ also leads to high review scores even if the paper quality is only moderate. By contrast, reviewer r_{19} with $\alpha_{19} = -0.8523$ (in the linear model) has some tendency of being harsh. The parameters in the nonlinear approach, $\alpha_{19} = -0.6411$ and $\gamma_{19} = 1.8889$, allow for a more differentiated representation of this reviewer's mapping of paper quality to review score.

The differences in modeling and reducing reviewer bias between the approaches results in different paper rankings. Consider, for ex-

Table 3. Input data from the review process for COMSOC-2010. The scores of 20 reviewers for 57 papers are shown. (Note that the data matrix given here is transposed compared with Table 1.) The papers are ordered with respect to their rank obtained by the naive approach.

	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9	r_{10}	r_{11}	r_{12}	r_{13}	r_{14}	r_{15}	r_{16}	r_{17}	r_{18}	r_{19}	r_{20}
p_1	–	–	–	–	–	–	–	–	–	–	7	–	–	–	7	–	–	–	–	7
p_2	7	–	–	–	–	–	–	–	–	7	–	–	–	–	–	–	–	–	–	–
p_3	–	–	–	–	–	–	–	–	–	–	7	–	–	–	7	–	–	–	–	7
p_4	–	–	–	–	–	–	–	–	–	–	–	7	–	–	7	–	–	–	–	–
p_5	–	7	–	–	–	–	–	–	–	–	–	–	–	6	–	–	–	–	–	–
p_6	–	–	–	–	–	–	–	–	–	–	–	7	–	6	–	–	–	–	–	–
p_7	–	–	–	–	–	–	–	–	–	7	–	–	–	–	6	–	–	–	–	–
p_8	–	–	–	–	–	–	–	–	–	–	–	–	7	–	–	–	–	–	6	–
p_9	–	–	–	–	–	–	–	–	–	–	–	–	–	–	7	–	–	–	–	6
p_{10}	–	–	–	–	6	–	–	–	–	–	–	–	–	–	–	7	–	–	–	–
p_{11}	6	–	–	–	–	–	–	–	–	–	–	7	–	–	–	–	–	–	–	–
p_{12}	7	–	–	–	–	–	–	–	–	6	–	–	–	6	–	–	–	–	–	–
p_{13}	–	–	–	–	–	–	–	–	–	–	–	–	–	–	7	–	–	–	5	–
p_{14}	6	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	6	–	–
p_{15}	–	6	–	–	–	6	–	–	–	–	–	–	–	–	–	–	–	–	–	–
p_{16}	–	6	–	–	–	–	–	–	–	–	6	–	–	–	–	–	6	–	–	–
p_{17}	–	–	–	–	–	–	6	–	–	6	–	–	–	–	–	–	–	–	–	–
p_{18}	–	–	–	–	–	–	–	–	6	–	–	–	–	6	–	–	–	–	–	–
p_{19}	6	–	–	–	–	–	–	–	–	–	–	–	–	–	6	–	–	–	–	–
p_{20}	–	–	–	6	–	–	–	–	–	–	–	–	–	–	–	6	–	–	–	–
p_{21}	–	–	–	–	6	–	–	–	–	–	–	–	6	–	–	–	–	–	–	–
p_{22}	–	–	6	–	–	–	–	–	–	–	–	–	–	–	–	6	–	–	–	–
p_{23}	–	–	–	–	5	–	–	–	–	–	–	–	–	–	–	–	–	–	7	–
p_{24}	7	–	–	–	5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
p_{25}	–	–	–	–	5	–	–	–	6	–	–	–	–	–	–	–	–	–	–	–
p_{26}	–	–	–	6	–	–	5	–	–	–	–	–	–	–	–	–	–	–	–	–
p_{27}	–	6	5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
p_{28}	–	–	–	–	–	–	–	5	–	6	–	–	–	–	–	–	–	–	–	–
p_{29}	–	–	–	–	–	–	–	–	–	–	–	5	–	–	–	6	–	–	–	–
p_{30}	–	–	6	–	–	–	–	–	5	–	–	–	–	–	–	–	–	–	–	–
p_{31}	–	–	–	–	–	–	–	5	–	–	–	–	–	–	–	6	–	–	–	–
p_{32}	–	–	–	–	–	5	6	–	–	–	–	–	–	–	–	–	–	–	–	–
p_{33}	–	5	–	–	–	–	–	6	–	–	–	–	–	–	–	–	–	–	–	–
p_{34}	–	–	–	5	–	–	6	–	–	–	–	–	–	–	–	–	–	–	–	–
p_{35}	–	–	–	–	–	–	–	5	6	–	–	–	–	–	–	–	–	–	–	–
p_{36}	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	5	–	6
p_{37}	–	–	–	–	–	–	–	–	5	–	–	–	5	–	–	–	–	–	–	–
p_{38}	–	–	–	–	–	7	–	–	–	–	5	–	–	3	–	–	–	–	–	–
p_{39}	–	–	–	–	–	–	–	–	–	–	7	–	–	–	–	–	–	–	3	4
p_{40}	–	–	–	–	–	–	–	–	–	–	–	–	5	–	–	–	–	–	–	4
p_{41}	–	4	–	5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
p_{42}	–	–	–	–	–	4	–	–	–	–	3	–	–	–	–	6	–	–	–	–
p_{43}	–	–	–	–	5	–	–	–	–	–	3	5	–	–	–	–	–	–	–	–
p_{44}	–	–	4	–	–	6	–	–	–	–	3	–	–	–	–	–	–	–	–	–
p_{45}	–	–	–	–	–	–	–	–	–	–	2	5	–	–	–	–	5	–	–	–
p_{46}	–	–	3	–	–	6	–	–	–	–	–	–	–	3	–	–	–	–	–	–
p_{47}	–	–	–	–	–	–	3	–	–	–	–	–	–	–	–	–	–	4	–	–
p_{48}	–	–	–	5	–	–	–	2	–	–	–	–	–	–	–	–	–	–	–	–
p_{49}	–	–	–	–	–	–	–	–	5	–	–	–	–	–	–	–	–	–	2	–
p_{50}	–	–	–	–	–	–	–	–	–	–	3	–	4	–	–	–	3	–	–	–
p_{51}	–	1	–	–	–	–	7	–	–	–	–	–	–	–	–	–	1	4	–	–
p_{52}	–	–	–	–	–	–	–	4	–	2	–	–	–	–	–	–	–	–	–	–
p_{53}	–	–	3	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3	–	–
p_{54}	–	–	–	–	–	–	–	3	–	–	–	–	–	–	–	–	–	3	–	–
p_{55}	–	–	–	2	–	–	–	–	–	–	–	–	–	–	–	–	–	3	–	–
p_{56}	–	–	–	–	–	–	–	–	–	–	–	1	–	–	–	2	–	–	–	–
p_{57}	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1	–	1	–

Table 4. The scores in all three approaches. The β_j in the linear approach are shifted by $\mu_{lin} = 0.6698$ and the nonlinear β_j by $\mu_{nonlin} = 2.8864$ in order to achieve the same average scores as in the naive approach. Note that this means a slightly modified righthand side in (5).

Number of paper	Naive approach score	Naive approach rank	Linear model score	Linear model rank	Nonlinear model score	Nonlinear model rank
1	7.000	1	7.557	1	6.549	7
2	7.000	2	6.831	8	6.132	15
3	7.000	3	7.557	2	6.549	8
4	7.000	4	6.315	15	7.538	2
5	6.500	5	7.305	3	6.230	13
6	6.500	6	6.815	9	6.957	5
7	6.500	7	6.602	10	5.150	29
8	6.500	8	7.195	4	7.229	3
9	6.500	9	6.965	6	6.477	10
10	6.500	10	6.249	17	7.651	1
11	6.500	11	6.123	19	6.352	11
12	6.333	12	6.588	12	6.179	14
13	6.000	13	6.891	7	6.482	9
14	6.000	14	5.552	28	5.913	19
15	6.000	15	5.697	25	5.194	28
16	6.000	16	6.598	11	5.462	22
17	6.000	17	5.124	33	5.078	31
18	6.000	18	6.528	13	6.550	6
19	6.000	19	5.989	20	4.922	34
20	6.000	20	5.783	24	5.039	32
21	6.000	21	6.303	16	7.205	4
22	6.000	22	6.483	14	5.227	25
23	6.000	23	7.130	5	6.323	12
24	6.000	24	6.228	18	5.931	18
25	5.500	25	5.846	22	5.971	16
26	5.500	26	4.162	43	4.719	36
27	5.500	27	5.964	21	5.218	26
28	5.500	28	5.509	31	5.456	23
29	5.500	29	4.644	38	4.405	47
30	5.500	30	5.687	26	4.806	35
31	5.500	31	4.917	34	5.210	27
32	5.500	32	4.095	46	4.550	39
33	5.500	33	5.791	23	5.660	21
34	5.500	34	4.162	44	4.513	43
35	5.500	35	5.514	30	5.962	17
36	5.500	36	5.527	29	5.784	20
37	5.000	37	4.911	35	4.691	37
38	5.000	38	5.243	32	4.999	33
39	4.667	39	5.644	27	5.444	24
40	4.500	40	4.769	36	5.089	30
41	4.500	41	4.264	41	4.647	38
42	4.333	42	3.796	47	4.507	44
43	4.333	43	4.668	37	4.532	41
44	4.333	44	4.271	40	4.204	50
45	4.000	45	4.349	39	4.544	40
46	4.000	46	4.136	45	4.434	45
47	3.500	47	3.718	48	4.183	51
48	3.500	48	2.344	54	4.247	48
49	3.500	49	3.047	49	3.855	53
50	3.333	50	2.936	51	4.235	49
51	3.250	51	3.009	50	4.515	42
52	3.000	52	4.238	42	4.430	46
53	3.000	53	2.903	52	3.614	55
54	3.000	54	2.729	53	3.649	54
55	2.500	55	1.702	56	2.973	56
56	1.500	56	0.644	57	−3.745	57
57	1.000	57	2.034	55	3.962	52

Table 5. The reviewers' parameters. Note that the zeros in the α columns of the last row result from the normalization according to (7).

Number i of reviewer	Linear model α_i	Nonlinear model α_i	γ_i
1	0.9511	4.0540	0.9190
2	0.1620	−0.7132	3.0569
3	0.2494	0.3388	2.1501
4	1.6499	1.3767	1.7379
5	−0.0676	10.3078	0.3896
6	1.7839	0.2520	2.7372
7	2.3662	−0.7730	6.1283
8	0.5962	0.9447	1.4482
9	0.7156	9.8857	0.4435
10	0.7260	−0.8439	3.3569
11	−0.0703	-0.2022	2.2902
12	1.1419	7.9980	0.5621
13	0.8011	4.3951	0.7558
14	−0.4330	0.3932	1.4713
15	0.4097	12.2399	0.4336
16	1.9088	11.6348	0.4356
17	0.1235	−0.7309	4.0056
18	1.2852	2.7802	0.9436
19	−0.8523	−0.6411	1.8889
20	0	0	1.8305

ample, papers p_{17} and p_{23}: p_{17} was (by good luck for the authors) reviewed by reviewers r_7 and r_{10}. As noted above, reviewer r_7 tends to be lenient; the same appears to apply (though to a lesser extent) to reviewer r_{10}. Thus, in the naive approach, paper p_{17} is likely to have been ranked higher than merited. Paper p_{23} was reviewed by r_5 and r_{19}. Reviewer r_5 seems to be neutral with at most a slight tendency of being harsh, reviewer r_{19} exhibits a more distinct tendency towards harshness. Thus, in the two approaches presented here, paper p_{23} is assigned better scores and jumps from rank 23 in the naive approach to rank 5 in the linear and to rank 12 in the nonlinear model. The corresponding mean squared errors (where $n = 116$ is the total number of reviews) are 0.4533 for the linear model and 0.1739 for the nonlinear model. It is not surprising that the additional parameters γ_i reduce the error.

4 Conclusions

In this paper, we introduced two statistical methods for fairer rating (and thus, ranking) of scientific papers based on scores of potentially biased, partially blindfolded reviewers. These methods work well also in cases where each paper is reviewed only by a small number of reviewers; in particular, there is no need for every reviewer to assess each paper. This approach clearly improves on the classical, naive, yet currently common method of averaging the individual reviewers' scores. The linear approach can be carried out by means of existing statistical standard software. The nonlinear approach, however, allows for a more detailed modeling of the behavior of reviewers. On the other hand, it requires more sophisticated software tools to be carried out. The authors assume that Section 3 provides sufficient information for its use, and they offer their help in analyzing data based on a data table like Table 3. We applied both methods to real data from a scientific conference, and pointed out some effects and implications that are visible in the results. This displays their potential to improve decision-making in peer-reviewed scientific publication venues.

Acknowledgments: We thank the M-PREF-2012 reviewers as well as the AAAI-2011 reviewers for their helpful comments.

References

[1] V. Conitzer, M. Rognlie, and L. Xia, 'Preference functions that score rankings and maximum likelihood estimation', in *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pp. 109–115. IJCAI, (July 2009).

[2] *Proceedings of the 3rd International Workshop on Computational Social Choice*, eds., V. Conitzer and J. Rothe, Universität Düsseldorf, 2010.

[3] V. Conitzer and T. Sandholm, 'Common voting rules as maximum likelihood estimators', in *Proceedings of the 21st Annual Conference on Uncertainty in Artificial Intelligence*, pp. 145–152. AUAI Press, (2005).

[4] J. Douceur, 'Paper rating vs. paper ranking', *ACM SIGOPS Operating Systems Review*, **43**, 117–121, (2009).

[5] N. Draper and H. Smith, *Applied Regression Analysis*, Wiley Series in Probability and Statistics, John Wiley & Sons, 3rd edn., 1998.

[6] R. Haenni, 'Aggregating referee scores: An algebraic approach', in *Proceedings of the 2nd International Workshop on Computational Social Choice*, eds., U. Endriss and P. Goldberg, pp. 277–288. University of Liverpool, (2008).

[7] K. Koch, *Parameter Estimation and Hypothesis Testing in Linear Models*, Springer, 2nd edn., 1999.

[8] H. Lauw, E. Lim, and K. Wang, 'Summarizing review scores of "unequal" reviewers', in *Proceedings of the 7th SIAM International Conference on Data Mining*. SIAM, (April 2007).

[9] H. Lauw, E. Lim, and K. Wang, 'Bias and controversy in evaluation systems', *IEEE Transactions on Knowledge and Data Engineering*, **20**, 1490–1504, (2008).

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

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

[12] M. Pini, F. Rossi, K. Venable, and T. Walsh, 'Aggregating partially ordered preferences', *Journal of Logic and Computation*, **19**(3), 475–502, (2009).

[13] M. Roos, J. Rothe, and B. Scheuermann, 'How to calibrate the scores of biased reviewers by quadratic programming', in *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, pp. 255–260. AAAI Press, (August 2011).

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

[15] R. Sokal and F. Rohlf, *Biometry: The Principles and Practice of Statistics in Biological Research*, W. H. Freeman, 4th edn., 2012.

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

[17] L. Xia and V. Conitzer, 'A maximum likelihood approach towards aggregating partial orders', in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pp. 446–451. IJCAI, (July 2011).

[18] L. Xia, V. Conitzer, and J. Lang, '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*, pp. 399–408. IFAAMAS, (May 2010).