Consensual Affine Transformations for Partial Valuation Aggregation

Hermann Schichl,[1] **Meinolf Sellmann**[2]
[1]University of Vienna, Austria, [2]General Electric
hermann.schichl@univie.ac.at, meinolf@gmail.com

Abstract

We consider the task of aggregating scores provided by experts that each have scored only a subset of all objects to be rated. Since experts only see a subset of all objects, they lack global information on the overall quality of all objects, as well as the global range in quality. Inherently, the only reliable information we get from experts is therefore the relative scores over the objects that they have scored each.

We propose several variants of a new aggregation framework that takes this into account by computing consensual affine transformations of each expert's scores to reach a globally balanced view. Numerical comparisons with other aggregation methods, such as rank-based methods, Kemeny-Young scoring, and a maximum likelihood estimator, show that the new method gives significantly better results in practice. Moreover, the computation is practically affordable and scales well even to larger numbers of experts and objects.

Score Aggregation for Partial Valuations

We consider the task of aggregating scores from experts who have only had access to a part of all objects. We focus purely on empirical performance in this paper, for the purpose of providing an effective tool that helps, for example, when aggregating reviewer scores in conference reviewing, the organization of large online learning courses where students grade each others' homework, or when multiple information-retrieval algorithms score webpages with respect to different sources of information (text, images, video, audio, etc). The recurring characteristic is that each expert only considers a part of the universe of objects, for example because the workload for each expert would be too high to score all objects, or because not every expert has the expertise to score every object. This can lead to very specific kinds of biases that need to be actively taken into account in the mechanism that is used to reach a decision on the ranking of objects.

The main contributions of our work are the development of a new, computationally efficient, *practical* approach to reduce partial valuation bias. Moreover, to the best of our knowledge we conduct the first comprehensive numerical comparison of partial value aggregation methods.

Problem Definition

Let us begin our work by introducing some terminology.

Definition 1 *Let $\mathcal{O}$ be a finite set of $N \in \mathbb{N}$ objects, and assume that there exists a function $f : \mathcal{O} \to \mathbb{R}$ assigning a "ground truth" value to each object. Furthermore, let $\mathcal{E}$ be a finite set of experts. For each $o \in \mathcal{O}$, we assume there is a set $\mathcal{E}_o \subseteq \mathcal{E}$ of experts each estimating $f(o)$. For each $o \in \mathcal{O}$ and $e \in \mathcal{E}_o$, denote with $f_o^e \in \mathbb{R}$ the valuation of expert e of object o.*

The objective of the Score Aggregation Problem (SAP) *is to estimate the ground truth valuations for each object as good as possible using only the expert valuations.*

Definition 2 *Let $\mathcal{A}$ be a finite set of alternatives and $\mathcal{E}$ a finite set of voters or experts. Assume that each voter has a strict preference order for [a subset of] the alternatives, i.e., the preference order is transitive and anti-symmetric. Such a ranking can be described by a ranking map $\mathcal{B} \mapsto \{0, \dots, |\mathcal{B}| - 1\}$, where $\mathcal{B} \subseteq \mathcal{A}$. If the maps do not just provide a ranking but a score for each alternative, we speak of a* cardinal voting system.

A collection of alternatives together with the set $\mathcal{V}$ of ranking maps r_e, for $e \in \mathcal{E}$, of expert preferences is called a preference profile $(\mathcal{A}, \mathcal{V})$. *A social choice function (SCF) f, also called a voting scheme, is a map that assigns to each preference profile $(\mathcal{A}, \mathcal{V})$ a ranking map $f(\mathcal{A}, \mathcal{V})$, which represents the aggregate choice of the experts.*

A simple way is to consider the cumulative rank each object has for all rankings where it appears. This ordering is commonly referred to as the *Borda count* (de Borda 1781; Lippman 2013).

Caragiannis et al. (2015) study peer grading in massive online open courses where students grade each other. They prove that the Borda count can result in approximate ground-truth-revealing grading schemes *provided that reviewers are carefully assigned to objects*. They use bundle graphs for this purpose. Complete freedom when assigning experts to objects is required for this method, which somewhat limits it to settings like grading in massive online courses. In conference reviewing, experts typically have pockets of expertise, and we cannot assign a knowledge representation paper to an expert in combinatorial search.

Another rank-based aggregation method is unweighted Kemeny–Young method. It is based on summing, for each

pair of objects, the number of experts who ranked the pair oppositely to the given ordering. The Kemeny–Young rule thus results in a ranking that minimizes the number of pairwise disagreements with the expert votes. (Conitzer, Davenport, and Kalagnanam 2006; Conitzer and Sandholm 2012; Conitzer 2013)

The starting point for our work is the paper by Roos et al. (2011). They introduce a quadratic programming approach that computes a maximum likelihood estimator based on the assumption that each object has one correct score that all experts aim to assess. The assessment is distorted in two ways: First, the experts assess the correct score with some random Gaussian noise of equal variance. Second, because of personal and systemic partial-view bias, the resulting score undergoes an affine transformation.

Consensual Affine Transformations

The inherent problem with the SAP is the limited view that each expert has on the universe of objects. Consequently, even when ignoring personal or methodological biases, experts are simply not in a position to judge how good their set of objects, as a whole, is, nor what the correct range in quality over their objects is, relative to the objects they have not seen. Therefore, it is all right to modify or normalize expert scores before averaging the experts' opinions to arrive at an aggregated vote.

Example: *Assume 5 reviewers each rate 3 out of 5 papers.*

reviewer	paper scores				
1	6	8	10	*	*
2	5	*	6	*	7
3	1	2	*	5	*
4	*	*	2	3	4
5	*	2	*	4	5
average	4	4	6	4	5.3
true	1	3	5	7	9

Table 1: Reviewer scores

The objective scores are given in the last row. Note how every reviewer assesses the relative quality of their sets of papers reasonably well. However, due to a lack of global scale, the mediocre paper 3 rises to the top of the heap, benefiting from the fact that it was reviewed by positive reviewers and reviewers whose other papers were objectively worse, while better papers were reviewed by more negative reviewers and reviewers who got assigned to a set of high-quality papers to begin with. This distorts the global ranking to a point where the second best paper 4 becomes indistinguishable with the worst paper 1. Moreover, looking at reviewer 3, e.g., we find that he or she gives an average score of 2.7 to his/her papers 1, 2, and 4, while the average of the aggregated scores is 4, which indicates significant disagreement between reviewer 3 and the group as a whole as to how good, on average, the papers 1, 2, and 4 actually are.

We develop a general procedure for the modification and aggregation of experts' scores that will prove valuable already in the special case of affine transformations. Consider a finitely parameterized set Φ of monotonically increasing expert modification functions $\varphi^e : \mathbb{R} \to \mathbb{R}$ that we want to use to modify each expert's scores: $\varphi^e(x) := \varphi(x; a_1^e, \ldots, a_m^e)$, the function parameters a_i^e, $i = 1, \ldots, m$, depending on the expert. We set $\varphi^e(f_o^e) =: \varphi_o^e$. Moreover, let Ψ be a family of averaging functions $\psi(x_1, \ldots, x_k) : \mathbb{R}^k \to \mathbb{R}$ that are increasing in every variable and that we intend to use to average the modified experts' valuations to arrive at our final value estimate: $\psi(\varphi_o^{e_1}, \ldots, \varphi_o^{e_k}) =: \psi_o$ for $\mathcal{E}_o = \{e_1, \ldots, e_k\}$.

Then, for every expert e, our task consists in determining the parameters $a_1^e, \ldots, a_m^e$ such that the ordering induced by the ψ_o is close to the ordering induced by the ground-truth values $f(o)$.

We now study the two questions which constitute the core of our contribution: 1. The first question is, which modification functions should we consider? 2. And second, since we do not know the ground truth values, what is a reasonable proxy objective that will indirectly lead to a reasonable ordering of objects?

1. Regarding the type of modification functions, we consider linear monotonic affine transformations as introduced in (Roos, Rothe, and Scheuermann 2011) (in fact, we learned about their work only after making this choice, which at least suggests that linear transformations are natural candidates here). In the following, we will consider $\varphi^e(x) := \varphi(x; a_1^e, a_2^e) = a_1^e x + a_2^e$ for $a_1^e \geq \varepsilon$ with fixed small $0 < \varepsilon \in \mathbb{R}$ to avoid disregarding or reversing an expert's opinion. Note how the two parameters selected for each expert allow the system to scale the range of valuations as well as shift the entire range. This directly addresses the inherent inability of experts to judge the correct range and the average quality of valuations their objects should have within the unknown universe of objects. At the same time, monotonicity and linearity ensure that the modified scores respect the *relative values* experts have assigned to their set of objects.

For the averaging functions ψ, in this paper we choose the arithmetic mean. This gives all reviewers the same influence, which may be important to gain acceptance in conference reviewing. However, note that other choices could easily be made here (and the algorithmic realizations of this framework, which we introduce in the next section, can easily accommodate some of these modifications). For example, one could consider a weighted average to take into account different reviewer confidences. It is also noteworthy that, while not explicitly mentioned in (Roos, Rothe, and Scheuermann 2011), the density function they use is maximized when the true score estimates equal the arithmetic mean of the transformed expert scores. In this regard, by using a simple arithmetic we stay in line with the published literature, even though we arrived there through a different motivation.

2. Regarding the second question, Roos et al. (2011) model reviewer distortions based on a probabilistic model and can thus derive the proxy as a maximum likelihood objective. Aiming to avoid the assumption of a specific proba-

bilistic model, we propose to use transformations that drive a specific notion of consensus among the experts. Namely, with consensus we do not mean to force experts into agreement on individual objects, because we obviously cannot, and do not want to, force experts to agree on each object. However, we can try and find modification functions so that each expert can agree what the *mean score and range of scores* for their set of objects should be, taking into account the views of all experts who have seen these objects.

We want to find modification functions such that, for each expert, the average modified score over all objects they considered is close to the average overall scores for all experts that have scored these same objects. More formally, for each expert e we want

$$\sum_{o \in \mathcal{O}^e} \varphi_o^e \approx \sum_{o \in \mathcal{O}^e} \psi_o, \tag{1}$$

whereby $\mathcal{O}^e := \{o \in \mathcal{O} \mid e \in \mathcal{E}_o\}$.

Analogously, we want the range of modified valuations for each expert to be close to the range of overall modified scores. In more formal terms, for all experts e, we want

$$\max_{o \in \mathcal{O}^e}\{\varphi_o^e\} - \min_{o \in \mathcal{O}^e}\{\varphi_o^e\} \approx \max_{o \in \mathcal{O}^e}\{\psi_o\} - \min_{o \in \mathcal{O}^e}\{\psi_o\}. \tag{2}$$

Example (contd): *Consider Reviewer 3 in our previous example. We may decide to scale and shift the scores provided using the function $\varphi(x) := 2x - 2$ and would arrive at a scoring of 0, 2 and 8 for papers 1, 2 and 4. Note how this is still in accordance with the reviewer's original opinion in terms of the relative score for each paper. The average score for these three papers, according to reviewer 3, is now 3.3. If we also (still sub-optimally) scale other reviewers as shown in Table 2, we get an aggregated score of 0.7, 2.7 and 7.3 for these three papers, with a mean of 3.6. Therefore, the discrepancy between this reviewer's view on the quality of papers 1, 2, and 4 is now much closer to the global view than it was before (recall that it was 2.7 versus 4 before, now it is 3.3 vs 3.6.)*

reviewer	a_1	a_2	new paper scores				
1	1	-5	1	3	5	*	*
2	4	-19	1	*	5	*	9
3	2	-2	0	2	*	8	*
4	2	1	*	*	5	7	9
5	2	-1	*	3	*	7	9
average			0.7	2.7	5	7.3	9
true			1	3	5	7	9

Table 2: Adjusted reviewer scores

Note that, so far, we used the $\approx$ to express our general desire to make certain terms *close* to one another. This vague objective obviously needs further refinement to become a prescriptive method. In the following, we propose four specific realizations of the above framework as optimization models. We stress again that our objective here is to find a *practically efficient method* that can effectively deal with the inherent bias introduced by partial valuations, *without having to make assumptions about the underlying probabilistic model of how experts' scores may be distorted.*

Linear Formulations

Let us begin by considering linear optimization models. The decision variables in the model are obviously the continuous a_k^e for all $e \in \mathcal{E}$ and $k \in \{1, 2\}$. For practical purposes, we will assume that each variable is bounded from above and below. Bounds typically arise from the respective application. E.g., when aggregating reviewers' scores at a conference, we may assume that there is some absolute truth in the valuations given, but we may still consider shifting the scores by up to 10% of the total range and/or scaling the range by up to 10%.

The main issue is now to formulate our objective that averages and ranges should be "close" to one another as loosely expressed in statements (1) and (2). Limiting ourselves to linear models in this subsection, we could simply minimize the sums of absolute deviations. Alternatively, we could consider the maximum norm and minimize the maximal absolute deviation. We propose to do both at the same time to strike a balance between worst and average case:

For all $e \in \mathcal{E}$, we set $M^e := |\mathcal{O}^e|$ and introduce continuous variables d_e with associated constraints $M^e d_e \geq \sum_{o \in \mathcal{O}^e}(\varphi_o^e - \psi_o)$ and $M^e d_e \geq \sum_{o \in \mathcal{O}^e}(\psi_o - \varphi_o^e)$. At optimum, d_e is then the absolute difference between the adjusted average score that this expert assigned to his/her objects, and the adjusted average score that all experts assigned to the same set of objects. We also add a variable Z_1 with constraints $Z_1 \geq d_e$ for all $e \in \mathcal{E}$. At optimum, Z_1 is then $||d_e||_\infty$. Then, we minimize $\sum_e d_e + \frac{|\mathcal{E}|}{2} Z_1 = ||d_e||_1 + \frac{|\mathcal{E}|}{2}||d_e||_\infty$. To make both one-norm and infinity norm operate at roughly the same magnitude, we multiply the maximum norm with the number of terms in the one-norm over 2.

This takes care of statements (1). Note how, so far, we have been able to formulate the problem as a convex linear continuous optimization problem. Unfortunately, when considering statement (2), we lose convexity (and thereby the ability to solve the resulting instances efficiently) because we need to minimize absolute differences of ranges. The first range, that of the modified scores of the respective expert, can be computed efficiently: The monotone linear transformation of expert scores guarantees that the worst and best objects (let us call them $o^e_{\min}$ and $o^e_{\max}$) for each expert remain identical. We can therefore compute the indices of the modified values that define that minimum and maximum in the range for the expert a priori. However, for the range of the aggregate modified scores, the objects that define the minimum and maximum values may very well change. In fact, the whole point of modifying expert scores is to *allow* them to change! Otherwise we might as well use the expert scores at face value and aggregate them to arrive at the joint ordering.

In the following we consider three ways to deal with the problem. One of them leads to an integer problem, two to a (sequence of) continuous problem(s).

Integer Programming Formulation To compute the range of the aggregate modified scores, we introduce continuous variables $x^e_{\min}, x^e_{\max} \in \mathbb{R}$ and binary indicator variables b^e_o for all experts $e \in \mathcal{E}$ and $o \in \mathcal{O}^e$. We constrain these variables by requiring $x^e_{\min} \leq \psi_o$, $x^e_{\min} \geq \psi_o - Mb^e_o$, and $\sum_{o \in \mathcal{O}^e} b^e_o = 1$ for all experts $e \in \mathcal{E}$, $o \in \mathcal{O}^e$ and an appropriately large value for M.[1]

After convergence, $x^e_{\min}$ ($x^e_{\max}$) denotes the minimum (maximum) average (over all experts scoring the same object) adjusted score over all objects scored by expert e. b^e_o is an indicator variable that is 1 if and only if object o has the smallest average (over all experts scoring object o) adjusted score of all objects scored by expert e. For the computation of $x^e_{\max}$, another variable like this is introduced analogously.

We then introduce continuous variables r_e for all experts $e \in \mathcal{E}$ with associated constraints $r_e \geq x^e_{\max} - x^e_{\min} - (\varphi^e_{o^e_{\max}} - \varphi^e_{o^e_{\min}})$ and $r_e \geq \varphi^e_{o^e_{\max}} - \varphi^e_{o^e_{\min}} - (x^e_{\max} - x^e_{\min})$. At optimum and after convergence, r_e is the absolute difference between the range of the adjusted scores that this expert assigned to his/her objects, and the range of the average adjusted scores that all experts assigned to the same set of objects.

As before, we introduce a continuous variable Z_2 for the maximum norm and constrain it with $Z_2 \geq r_e$ for all experts $e \in \mathcal{E}$. The respective part of the objective for the range approximation is then to minimize $\sum_e r_e + \frac{|\mathcal{E}|}{2} Z_2 = ||r_e||_1 + \frac{|\mathcal{E}|}{2} ||r_e||_\infty$.

Linear Continuous Formulation A simple model that leads to a convex linear formulation is to simply ignore statement (2) altogether. Alternatively, we will approximate the range of aggregate modified scores by setting it to $\psi_{o^e_{\max}} - \psi_{o^e_{\min}}$. That is, we use the initially worst and best objects for the expert as approximate representatives of the final objects that mark the aggregate worst and best scores. After we solved this model, the worst and and best scores for an expert may have shifted. We can then *iterate* the process using these new worst and best objects (while carefully adjusting the bounds on the variables a^e_k so that overall we do not alter the scores in excess of the initially given limits) until either an iteration limit is reached or the best and worst objects for all experts no longer change.

Non-linear Continuous Formulation

We used a mixture between one-norms and infinity norms above in order to keep the objective function linear. Another choice is to use a quadratic programming approach and to minimize the Euclidean norms over the average score deviation and the average range deviation for each expert when compared with the entire group of experts. In this last formulation, we therefore minimize $\sum_e d^2_e + \sum_e r^2_e$.

For the r_e, we run into the same problem as before: For a given expert e, the affine transformations of expert scores may lead to a change in $\text{argmax}_{o \in \mathcal{O}^e} \psi_o$ and $\text{argmin}_{o \in \mathcal{O}^e} \psi_o$. Consequently, the objects that define the group worst and group best over the objects that expert e has scored are not stable. We could formulate a non-linear integer program to deal with this. However, key for a successful application of the technology we develop here is computational efficiency.

Consequently, we propose to use the same approach as presented in the second linear continuous approach. Namely, we solve a sequence of quadratic programs. At the beginning we fix, for each expert, the group-worst and group-best objects over the set of objects she scored, and use them to define the size of the group range. At the end of each iteration, we check if any of these have changed for any expert, as a result of the affine transformations applied to the experts' scores. If so, we continue, otherwise we stop. In our experiments we limit the number of iterations to 20. In Table 3 we summarize the four models introduced in this section.

name	type	iterated	objective
NoRange	LP	no	$\sum_e d_e + \frac{\|\mathcal{E}\|}{2} Z_1$.
Aff-LP	LP	yes	$\sum_e d_e + \frac{\|\mathcal{E}\|}{2} Z_1 + \sum_e \tilde{r}_e + \frac{\|\mathcal{E}\|}{2} \tilde{Z}_2$.
Aff-IP	IP	no	$\sum_e d_e + \frac{\|\mathcal{E}\|}{2} Z_1 + \sum_e r_e + \frac{\|\mathcal{E}\|}{2} Z_2$.
Aff-NLP	QP	yes	$\sum_e d^2_e + \sum_e \tilde{r}^2_e$.

Table 3: Four realizations of consensual affine transformations as (iterated) optimization models. $\tilde{Z}$ and $\tilde{r}$ denote the terms whose definition changes in consecutive iterations.

Numerical Results

We have devised four optimization approaches to deal with the biases introduced by partial valuations commonly occurring in conference reviewing. In a set of extensive experiments, we now evaluate and compare the various proposed formulations of the consensual affine transformation approach with each other and the most prominent existing techniques for score aggregation. Ultimately, we want to assess which methods can effectively deal with partial valuation bias.

Metric

There are different ways how we could evaluate the quality of an aggregate estimate. For conference reviewing, the main task is to identify the *top-rated set* of objects. A defect occurs when a paper that should be published is not identified as being part of the top X%, and a paper of lesser quality is put into the program instead. As our error metric, we thus chose to count the number of objects erroneously included in a top set of all objects, when compared to a ground truth top set. We track this "error rate" as the percentage of misplaced objects in the top set. In our experiments we choose 25%, but this choice did not affect the results which were the same for other percentages, both lower and higher.

Benchmarks

One True Value Per Object: The first benchmark follows the randomized model in (Roos, Rothe, and Scheuermann 2011): For a desired number of objects and experts, as well as a target range for the number of scores per object, our

[1] We tried alternative models based on special ordered sets as well, which did not lead to performance improvements.

generator first picks a true score for each object as an integer between 1 and 10 by drawing from a normal distribution with mean 5 and standard deviation 2, then rounding the result to the nearest integer, and clipping the interval at 1 and 10. These values are the ground-truth values for each object.

The generator then assigns random experts to each object, whereby the total number of scores per object is guaranteed to be in the range given, and the total number of scores each expert gives is limited from above by 20% over the average of objects that each expert needs to score if each object had to receive the maximum number of scores.

Finally, the generator picks a scaling factor and shift value for each expert, whereby the maximal distortion is controlled by user-defined ranges for the scaling as well as the shift. The final score each expert gives to her set of objects is then the nearest integer in 1 to 10 of ground truth value of each object times the scaling factor plus the shift value plus noise.

The parameters for this first probabilistic model were set to somewhat resemble the distribution of papers we would expect at a typical conference, with a big body of papers with average quality. However, the particular parameters we chose did not affect our comparative results as additional experiments showed.

Latent Metric: The generator for the second randomized benchmark essentially follows the procedure of the first. However, instead of picking one "true" score for each object, it picks one score for each *object/expert pair* to which the randomized linear distortion and the noise are applied to generate the benchmark instance.

To generate a "true" valuation of each expert for each object reviewed, we choose *three latent values* for each object. The idea is that there are multiple metrics (this could be novelty, impact, and quality of evaluation) that underlie the overall score for each object. For each of the latent metrics, in this benchmark we assume there exists one true score. However, different reviewers value different metrics differently strongly. For each expert, we therefore choose three non-negative weights that add up to 1 to reflect the relative importance of each metric. We arrive at the "true" score for an object/expert pair by computing the expert-dependent convex combination of the true object scores regarding the latent metrics. Then, the affine and noise distortion works exactly as in the generator for the first benchmark.

Instances Derived From Real Conferences: To generate a different, more realistic type of benchmark, we were fortunate to obtain real-world data from two AI conferences (we also tried to obtain data from previous AAAI conferences, unfortunately to no avail). The first had 41 submissions (objects) and 33 reviewers (experts). The second was a bit larger, it received 71 submissions and had 67 reviewers. As ground truth, we experiment with two setups: In the first, we use the unaltered conference data as ground truth, distort it (see below), and then see how well we can re-construct the original data. In the second setup, we compute "true" scores for each paper by consensually scaling and shifting reviewers' scores and using the aggregate scores to define an ordering on all submissions. We then aim to recover the top-rated objects according to this ordering. To distort the ground truth, we shift and scale the experts' scores and add noise to each individual score, followed by rounding the result to the nearest integer in 1 to 10.

The resulting benchmark instances differ in two important aspects from the instances that were generated completely automatically. First, and most importantly, the *bipartite graph structure* of which experts score which objects is given by the two real-world examples and can thus reflect mechanisms like bidding as well as areas of expertise, etc. Secondly, contrary to the one-true-value generated data, the value from which we generate experts' scores by noisy affine distortion is *not identical* for all experts scoring the same object. Instead, we use the "true" scores for each expert/object pair as derived from the original data. Consequently, even if a score aggregation method can find a perfect consensual affine transformation to undo the distortions, on individual objects, we still find experts disagreeing by more than just the noise that was added to generate the benchmark instance. Note that the latent-metric generated benchmark also has this latter property.

Competitors

Apart from the four methods introduced in this paper, we compare with simple averaging (still most commonly used in conference systems), the Borda count (for which Caragiannis et al. (2015) showed that it can lead to provably good performance when the assignments of experts to objects permits), with un-scored and thus un-ranked candidates receiving zero points, as well as the unweighted and weighted Kemeny-Young methods. Our last competitor is the quadratic programming method from Roos et al. (2011). Note that our first automatically generated benchmark set is specifically constructed so that the assumptions for their method are met perfectly.

In summary, we compare the following nine approaches: Ave (order objects by average score received), Borda (rank objects for each reviewer, then sort objects by average rank received), KY (use Kemeny-Young scoring to order objects), wKY (use weighted Kemeny-Young scoring to order objects), MLE (Roos et al.'s maximum likelihood estimator), NoRange (use linear continuous affine transformations without penalizing range difference, then sort objects by mean adjusted score received), Aff-LP (use affine transformations computed by iterated linear programming to adjust experts' scores, then sort objects by mean adjusted score received), Aff-IP (use consensual linear affine transformations to adjust experts' scores, then sort objects by mean adjusted score received), and Aff-NLP (use affine transformations computed by iterated quadratic programming to adjust experts' scores, then sort objects by mean adjusted score received).

All approaches have been implemented in C++ using the Gnu g++ compiler 4.4.5 (Red Hat 4.4.5-6) and were run on Intel Xeon CPU X3430 processors at 2.40GHz. Whenever optimization was needed, we used Ilog Cplex 12.6.

Small Test Sets

We start our comparison on small generated data sets. These instances are simple and small enough to run all nine competitors.

In Table 4 we show the results. In this experiment, we set the distortion limits to $[\frac{5}{6}, \frac{6}{5}]$ for scaling and $[-1.8, 1.8]$ for shifting (which corresponds to 20% of the total score range in each direction). The noise was chosen uniformly at random in $[-0.72, 0.72]$ (or 40% of the maximal shifting distortion).

We observe that Kemeny-Young scoring takes orders of magnitude more computation time than any of the other methods. The reason why it is so much more computationally intensive than, e.g., the integer linear model we use to compute consensual affine transformations, is that, in Kemeny-Young scoring, we need to determine the full ordering of all objects using integer variables in the model. This translates into a number of binary variables that is quadratic in the number of objects. In Aff-IP, on the other hand, we only need binary variables to set the worst and best object for each expert. Since each expert scores at most all objects (and usually much fewer), and there are usually way fewer experts than objects, the corresponding integer programming models are significantly easier to solve in practice.

Method	Rank	Error		CPU Time
	μ	μ [%]	σ [%]	μ [s]
Ave	6.78	9.62	5.91	0.01
Borda	6.55	8.46	5.30	0.01
KY	6.46	7.44	4.60	7.38K
wKY	6.46	7.44	4.60	7.3K
MLE	3.14	0.23	1.16	0.02
NoRange	3.29	0.55	1.76	0.01
Aff-LP	3.66	1.27	3.22	0.06
Aff-IP	3.38	0.70	1.92	47.1
Aff-NLP	3.40	0.70	2.26	0.25

(a) 50 Objects, 30 Experts, One True Value

Method	Rank	Error		CPU Time
	μ	μ [%]	σ [%]	μ [s]
Ave	4.57	13.1	8.84	0.02
Borda	5.16	14.9	8.32	0.02
KY	5.82	17.2	8.24	18.1K
wKY	5.82	17.2	8.24	18.3K
MLE	4.32	12.2	8.07	0.03
NoRange	3.66	9.72	7.22	0.03
Aff-LP	4.09	11.4	8.09	0.03
Aff-IP	4.14	11.6	7.53	92.4
Aff-NLP	4.10	11.35	7.01	0.06

(b) 40 Objects, 20 Experts, Latent Metric

Method	Rank	Error		CPU Time
	μ	μ [%]	σ [%]	μ [s]
Ave	3.85	13.7	7.63	0.01
Borda	6.05	18.2	3.97	0.01
KY	6.55	22.0	8.13	135K
wKY	6.55	22.0	8.13	135K
MLE	4.9	15.9	6.43	0.01
NoRange	3.5	11.4	5.24	0.02
Aff-LP	3.5	11.4	5.13	0.03
Aff-IP	3.85	12.1	5.10	277
Aff-NLP	3.15	10.6	3.69	0.14

(c) 50 Objects, 30 Experts, Latent Metric

Table 4: Mean rank, error mean and std. dev., and runtime as average over 100 instances each for various score aggregation methods run on generated data, where each paper received between 3 and 5 reviews each, and 25% of papers get accepted.

To address the complexity of Kemeny-Young scoring, approximation algorithms (in the theorist sense of the term, meaning these algorithms come with performance guarantees) have been developed for Kemeny-Young scoring (Kenyon-Mathieu and Schudy 2007). The problem with these approaches is that the approximation guarantees regard the objective function value instead of the ordering itself. Consequently, the objects that end up in the top set could differ substantially, even though the overall stress in the system may not be much more than it would be for the optimal Kemeny-Young solution.

More fundamentally, though, the experiments show that the approximation target itself, the Kemeny-Young ordering, does not exhibit great strength when dealing with partial valuation bias, which we must assume exists in our application, given that experts only get to see a small subset of the universe of all objects, even when setting other considerations such as personality differences aside. We observe that Kemeny-Young scoring performs marginally better than averaging and simple rank aggregation in the Borda count on the one-true-value benchmarks, and worse than both on the latent-metric benchmarks. The MLE from Roos et al. (2011) and all affine consensual transformation methods, on the other hand, are significantly more effective at reducing partial valuation bias, whereby all but the integer programming based method work very fast. In fact, we ran a comparison on the one-value benchmark with 2,000 objects and 500 experts (distortion 20% and noise 40%, 4 to 5 scores per object). Aff-NLP had an error rate under 0.6% and ran in less than 8 minutes on average. Since the Aff-IP is slower and does not appear to give any significant benefit over the other consensual-affine transformation methods anyway, we exclude the Kemeny-Young methods and Aff-IP from further experiments.

Scaling Experiments

We now scale both generated benchmarks, varying the number of objects between 20 and 100 in increments of 10, and the number of experts between 10 and the respective number of objects, also in increments of 10. In Figure 1, we visualize the results of over 464,000 experiments (100 runs per data point). Distortion and noise limits, as well as the range of scores per object, are the same as before.

The one-true-value benchmark was constructed specifically to meet the assumptions of the MLE from Roos et al. (2011). Consequently, the maximum likelihood method sets the gold standard for this data set. Averaging and the Borda count both cannot compete and have markedly worse error rates, whereby the Borda count gets better when there are fewer reviewers.[2] This is quite natural as this implies that each expert sees a larger percentage of the entire universe of

[2]Not shown here: Our data shows that the MLE error rate declines with the ratio of number of experts over number of objects.

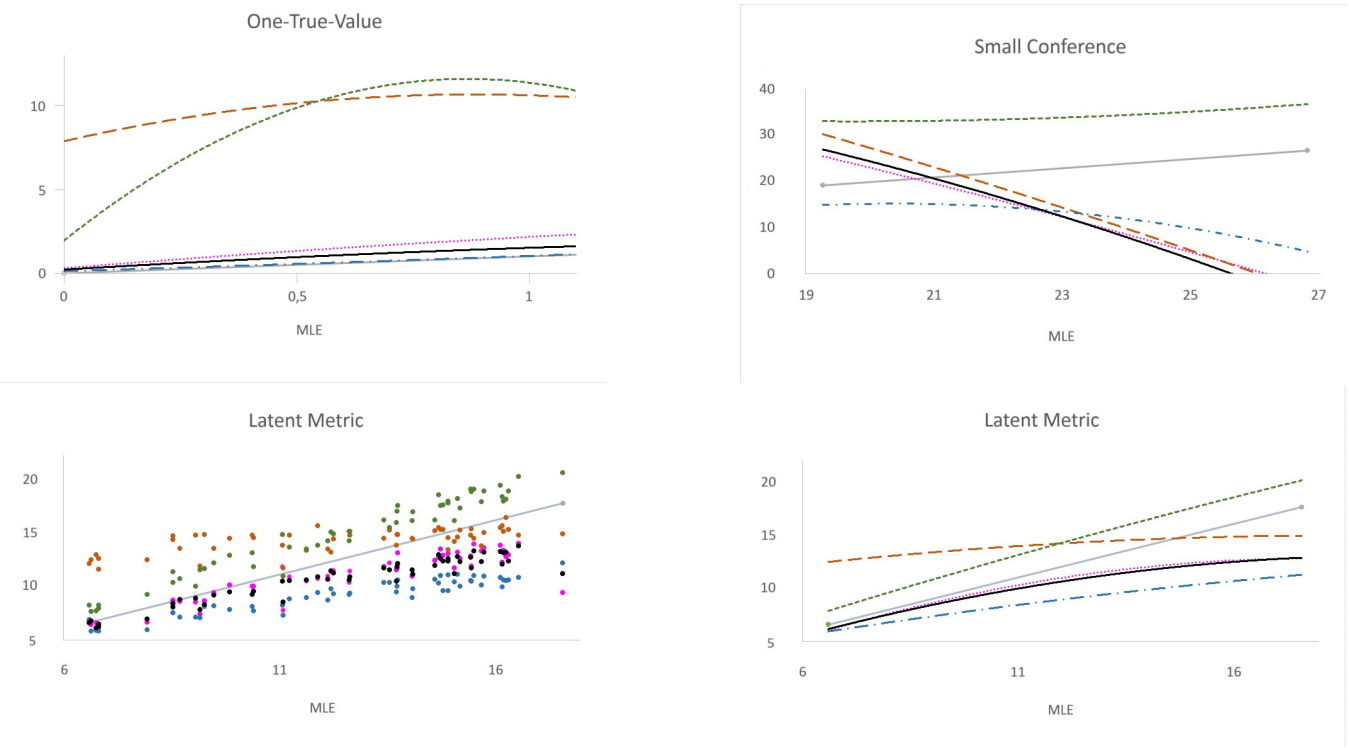

Figure 1: One-true-value (top left), small conference (top right), and latent-metric benchmark experiments (scatter plots on bottom left with 100 instances per data point, second order polynomials on right). We give error rates (y-axis in %) for six score aggregation methods (MLE as gray line, Ave in brown, Borda in green, NoRange in blue, Aff-Cont in purple, and Aff-NLP in black) over the MLE error rate (x-axis in %) when varying the number of objects and experts.

objects, which makes the partial ranking of objects more informative.

We find that the NoRange method, where we ignore differences in expert ranges, is statistically not worse than the MLE in this experiment. Moreover, when we consensually shift and scale by repeatedly solving linear or non-linear programs in the Aff-Cont and Aff-NLP methods, respectively, we observe only slightly worse error rates.

On the latent-metric benchmark, we see how brittle maximum likelihood methods can become when the underlying probabilistic assumptions are not met perfectly. Even a slight deviation from the original model (recall that now each object has three latent values instead of just one, out of which we form ground-truth object/expert scores by convex-combining the latent object scores consistently for each expert) leads to much less convincing performance. Statistically worse performs only the simple Borda count. Not even the simplistic averaging method is overall worse than the MLE. Best performing on this benchmark, however, are the consensual affine transformation methods, whereby the NoRange method is the best, with Aff-NLP trailing marginally in second place.

One reason why the maximum likelihood estimator may not work well in case different experts actually truly disagree is shown in the following example:

		Paper evaluation							
Reviewer	# reviews	1	2	3	4	5	6	7	8
1	4	1	–	–	2	–	2	–	1
2	6	2	–	2	1	2	1	–	2
3	4	–	2	2	–	2	–	2	–
4	4	1	–	–	2	–	2	–	1
5	4	–	2	1	–	1	–	2	–

In this case, an optimal solution to the QP introduced in (Roos, Rothe, and Scheuermann 2011) is to set scaling and shifting values $\hat{p}, \hat{q}$ to 0 for reviewers 1, 2, 4, and 5, while reviewer 3's values are scaled by a factor of 5 and then shifted by -10. The resulting objective costs are obviously 0 and thus optimal, with all objects receiving the same score of 0. That means, the MLE-based ordering in this example is completely arbitrary, due to the marginalization of expert opinions which run contrary to one another. Note that this is not the result of some pathological example but inherent to the MLE optimization itself which will always benefit from washing out actual differences of opinion which, in its motivating probabilistic model, should not exist in the first place. Due to the variable bounds, consensual affine transformations naturally preserve contrary ratings and thus prevent the undesirable outcome of arbitrary object orderings. In the example above, NoRange and Aff-NLP both return the meaningful order $1 = 8 < 4 = 6 < 3 = 5 < 2 = 7$.

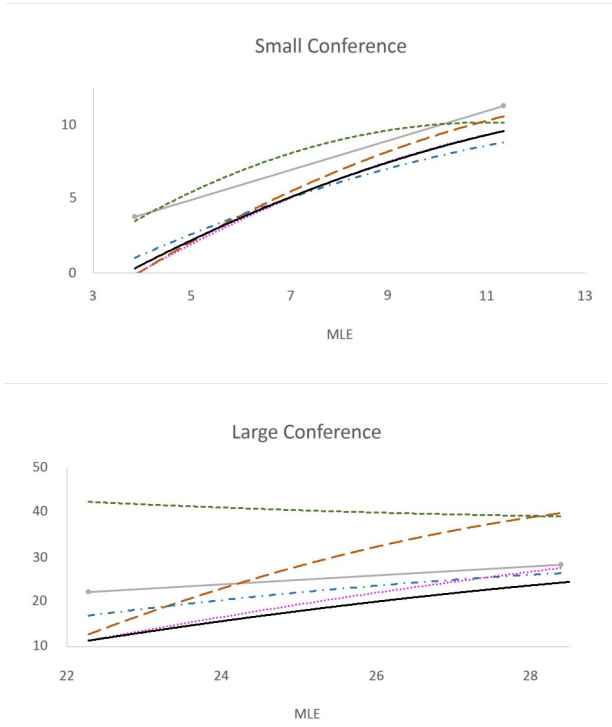

Figure 2: Fitted polynomials for conference benchmark experiments (smaller conference on left). Axes and color coding of aggregation methods as in Figure 1.

Real-world Inspired Benchmarks

While they were able to show the need for methods that are tailor-made to deal with partial valuation bias, the previous experiments lack realism in the sense that the matching of experts to objects is rarely random. We now use real-world data from two AI conferences. As this setting does not allow us to scale the problem size, we instead vary distortion and noise intervals between 5% and 30%, which leads to over 400,000 values per conference that we analyze.

Figure 1 shows the results on the smaller conference when using the first setup where we use the unaltered, original reviewer values as ground truth. While we can see again clearly that neither Borda nor the MLE work competitively, and consensual affine transformations work best, we notice that, strangely, averaging as well as the various consensual methods have an error pattern that is reversed from MLE's. The problem here is that the experimental setup is flawed: Recall that we strive to eliminate systematic bias that arises from partial valuations. In this setup, we start with a ground truth *that still contains this bias*. We then add even more bias, which we consequently strive to remove again. Eventually, we compare with the original, biased data. Clearly, to assess if any noise filter works, we need to conduct the final comparison with a noise-free pattern so that we can properly assess if the noise filter works. Otherwise, we cannot distinguish undesirable effects of our filter from its actual function of removing bias.

In Figure 2, we present the results on the conference data, starting from de-noised ground truth valuations. We observe that the MLE, which performed well by design on the one-value benchmark, but struck out on the latent-value benchmark and the unaltered conference data, can again not compete with consensual affine transformation-based methods.

These experiments confirm that the consensual affine transformation techniques are not only computationally affordable, but also lead to the lowest error rates when selecting top objects for all methods that we compared.

Conclusion

We studied aggregation methods for partial valuation aggregation. In the first extensive comparison of this kind, we showed that standard methods, like Borda count or Kemeny-Young aggregation, perform poorly for this task. Moreover, we found that a maximum likelihood approach designed specifically for the problem of ordering conference papers works very well for the noise model it was designed for, but generalizes poorly. We introduced various optimization approaches to compute consensual affine transformations of the experts' scores. Our tests on automatically generated and real-life data showed that these scale well and yield better results than other methods. Overall best perform the NoRange and Aff-NLP methods, whereby the first often leads to the best results and the latter works overall most robustly, giving the best or second best performance on all benchmarks we tested.

References

Caragiannis, I.; Krimpas, G. A.; and Voudouris, A. A. 2015. Aggregating partial rankings with applications to peer grading in massive online open courses. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2015, Istanbul, Turkey, May 4-8, 2015*, 675–683.

Conitzer, V., and Sandholm, T. 2012. Common voting rules as maximum likelihood estimators. *arXiv preprint arXiv:1207.1368*.

Conitzer, V.; Davenport, A.; and Kalagnanam, J. 2006. Improved bounds for computing kemeny rankings. In *AAAI*, volume 6, 620–626.

Conitzer, V. 2013. The maximum likelihood approach to voting on social networks. In *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*, 1482–1487. IEEE.

de Borda, J. C. 1781. Mémoire sur les élections au scrutin. *Histoire de l'Academie Royale des Sciences*.

Kenyon-Mathieu, C., and Schudy, W. 2007. How to rank with few errors. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, 95–103. ACM.

Lippman, D. 2013. *Voting theory*. Creative Commons BY-SA. chapter 2.

Roos, M.; Rothe, J.; and Scheuermann, B. 2011. How to calibrate the scores of biased reviewers by quadratic programming. In *AAAI*.