Regularized Maximum Likelihood Estimation of the Subjective Quality from Noisy Individual Ratings

Regularized Maximum Likelihood Estimation of the Subjective Quality from Noisy Individual Ratings

Lohic Fotio Tiotsop[1], Antonio Servetti[1], Marcus Barkowsky[2], Enrico Masala[1]
[1]*Control and Computer Engineering Department, Politecnico di Torino, Torino, Italy*
[2]*Deggendorf Institute of Technology, University of Applied Sciences, Deggendorf, Germany*
lohic.fotiotiotsop@polito.it, antonio.servetti@polito.it, marcus.barkowsky@th-deg.de, enrico.masala@polito.it

***Abstract*—Despite several approaches to recover the ground truth subjective quality score from noisy individual ratings in subjective experiments have been explored in the literature, there is still room for improvement, in particular in terms of robustness to noise. This paper proposes a new approach that combines the traditional maximum likelihood estimation framework with a newly proposed regularization term, based on information theory concepts, that is meant to underweight surprising ratings of the quality of a given stimulus, looked at as a noise manifestation, in the final analytical expression of the recovered subjective quality. Computational experiments show the higher robustness to noise of our proposal when compared to three state-of-the-art methods.**

***Index Terms*—subjective experiment, ground truth subjective quality, maximum likelihood, regularization**

I. Introduction and related work

The need for large scale subjectively annotated datasets is of paramount importance to researchers working on designing new, more reliable, quality estimation algorithms. However, it is very difficult to fully eliminate the multiple influence factors that can generate noise on individual ratings in large scale subjective experiments. For instance, the content of a certain stimulus might affect the emotional state of a subject, yielding an incorrect evaluation of its quality. This source of noise is obviously not under the control of the designer of the experiment. Also, researchers frequently resort to crowdsourcing platforms: in this case nothing guarantees that all participants understood and followed the instructions provided in these platforms. This is another example of a source of noise that the designer of the experiment cannot control.

Several approaches to eliminate the noise before using the data have been proposed [1]–[5]. The most trivial one, for instance, consists in simply averaging the individual opinion scores in order to mitigate the effects of individual expectations. This yields the well known Mean Opinion Score (MOS). The MOS is however known to be particularly sensitive to outliers, i.e., peculiar subjects. A number of approaches to identify such outlier subjects and remove their ratings from the dataset before computing the MOS have been proposed, e.g., the ITU-T BT.500 Rec [1]. These approaches are however perceived by several authors [2]–[4] as over-killing, since by removing all the ratings of a given subject, one throws away also those that were correctly expressed.

More recent approaches avoid subject removal and can be classified under two main categories. The first category assumes that a subject only occasionally provides inconsistent votes [4]. Hence the subject behavior can be modeled with a mixture of two discrete probability distributions: the first one models correct evaluations, while the second accounts for occasional inconsistent ratings of the subject. The second category [2], [3] instead assumes that a subject is permanently characterized by an intrinsic bias and a certain level of inconsistency. The model [3] implemented in the Sureal software [6], used in the results section, follows this latter line of thought. In both categories assumptions are made on the probability distribution underlying individual ratings and statistical methods, typically the Maximum Likelihood Estimation (MLE) framework, are used to recover the subjective quality as one of the parameter of the assumed distribution.

This work proposes an approach that is similar to the more recent approaches from the point of view of avoiding subjects removal, but the key innovative point is that no assumptions are made on the discrete probability distribution underlying the subject behavior. Our main idea is to use standard mathematical tools to find a way to underweight those ratings which are "surprising" for a given stimulus, and hence, can be interpreted as a noise manifestation. In particular, in addition to the classical MLE framework, we rely on an information theory concept to measure how surprising is an event, then use it to define an optimization problem yielding our estimation of the subjective quality from noisy individual opinion scores. We call our proposal Regularized Maximum Likelihood Estimation (RMLE) of the subjective quality.

The results show that our approach yields a more stable subjective quality estimation from noisy individual opinion scores when compared to state-of-the-art methods.

II. Our Approach to Subjective Quality Recovery

A. Notation and Motivation

Let us assume a subjective experiment has been carried out by asking a set $\mathcal{J}$ of subjects to evaluate a set $\mathcal{I}$ of stimuli on a discrete quality scale offering a set $\mathcal{K}$ of possible opinion scores. Let also denote by O_i^j the rating of the subject $j \in \mathcal{J}$ on the quality of the stimulus $i \in \mathcal{I}$; and by n_{ik} the number of subjects that voted $k \in \mathcal{K}$ for the stimulus $i \in \mathcal{I}$.

The MOS of the stimulus $i \in \mathcal{I}$ is defined as:

$$MOS_i = \sum_{j \in \mathcal{J}} \frac{1}{|\mathcal{J}|} \cdot O_i^j = \sum_{k \in \mathcal{K}} \frac{n_{ik}}{|\mathcal{J}|} \cdot k \tag{1}$$

The first equality in Eq (1) highlights the fact that when using the MOS, all individual ratings have the same weight (i.e., $\frac{1}{|\mathcal{J}|}$). Thus, each rating has the same importance independently from the fact that it is reliable or not. In other words, by weighting each possible opinion score k on the quality scale with the fraction $\frac{n_{ik}}{|\mathcal{J}|}$, one yields a ground truth subjective quality estimator (the MOS) that attributes same importance to noisy and noiseless opinion scores.

Our idea is to determine a more robust way to weight the different opinion scores offered by the quality scale, i.e., giving less weights to opinion scores that are potentially noisy. We define the recovered quality Q_i of the stimuli $i \in \mathcal{I}$ as:

$$Q_i = \sum_{k \in \mathcal{K}} q_{ik} \cdot k \tag{2}$$

where each weight q_{ik} is different from the fractions $\frac{n_{ik}}{|\mathcal{J}|}$ in Eq (1) and will be determined in the next section.

B. Mathematical Formulation of Our Approach

Let us consider the weight q_{ik} as the unknown probability of choosing the opinion score $k \in \mathcal{K}$ when rating the stimuli $i \in \mathcal{I}$. Given a dataset of ratings, the probability of obtaining exactly the observed data, also known as the likelihood function in statistics, can be expressed as:

$$L(q) = \prod_{i \in \mathcal{I}} \prod_{k \in \mathcal{K}} q_{ik}^{n_{ik}} \tag{3}$$

where q denotes a vector containing all the values $q_{ik}, \quad \forall i \in \mathcal{I}, \quad k \in \mathcal{K}$.

The logarithm of $L(q)$, known as the Log-Likelihood function, is given by:

$$LL(q) = \sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}} n_{ik} \cdot \log(q_{ik}) \tag{4}$$

The MLE framework suggests that the values of q_{ik} for which the function $LL(q)$ is maximized are the desired estimates of the weights $q_{ik}, \quad \forall i \in \mathcal{I}, \quad k \in \mathcal{K}$

However, such a maximization of the function $LL(q)$ would estimate each q_{ik} as the fraction $\frac{n_{ik}}{|\mathcal{J}|}$. But, as stated before, this is not a solution particularly robust to noisy ratings.

Therefore, we introduced a regularization term, to be added to $LL(q)$, before formulating the optimization problem that will yield to the desired weights q_{ik}.

Our idea is to design a regularization term that penalizes "surprising" events, i.e., opinion scores on the quality scale that appear to be chosen with low frequency for a given stimulus. In fact, we do believe that noisy ratings occurs only occasionally, while consistent ratings are concentrated on a set of opinion scores frequently chosen. Formally, we measure how surprising is the choice of the opinion score $k \in \mathcal{K}$ for the stimulus $i \in \mathcal{I}$ through the value S_{ik} defined as:

$$S_{ik} = -log\left(\frac{n_{ik}}{|\mathcal{J}|}\right) \tag{5}$$

It is worth noting that the quantification of how surprising is an event through the logarithm of its probability is a consolidated approach in information theory [7]. The definition in Eq (5) is not therefore a peculiarity of this work.

We then defined the following regularization term

$$R(q) = -\sum_{i \in \mathcal{I}} \sum_{k \in \mathcal{K}} S_{ik} \cdot q_{ik} \tag{6}$$

to be added to the Log-likelihood function $LL(q)$ to formulate the following optimization problem whose solution provides the weights $q_{ik} \quad \forall i \in \mathcal{I} \quad k \in \mathcal{K}$. Thus the formulation becomes:

$$\begin{aligned} &\max_{q} LL(q) + \lambda \cdot R(q) \\ &s.t. \quad \sum_{k \in \mathcal{K}} q_{ik} = 1 \quad \forall i \in \mathcal{I} \end{aligned} \tag{7}$$

where λ is the regularization coefficient.

In the following an interpretation of the optimization problem in Eq (7) is provided. For a given stimulus $i \in \mathcal{I}$, S_{ik} assumes large values for less frequently chosen opinion scores $k \in \mathcal{K}$, i.e. those for which n_{ik} tends to 0. By subtracting the term $R(q)$, each value S_{ik} is looked at by the optimization problem as a virtual cost to be paid on the objective function depending on the value that is attributed to the weight q_{ik} of the opinion score k when recovering the quality of the stimuli i. Therefore, in order to maximize the objective function, for each stimuli i, not frequently chosen opinion scores (those with large value of S_{ik}) and hence potentially noisy ones, receive less weight (lower value of q_{ik}) in the optimal solution.

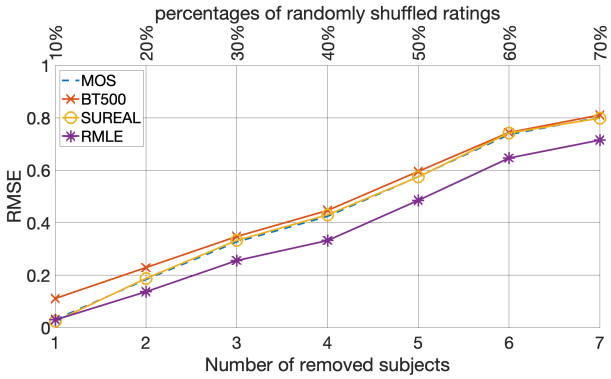
As already mentioned, the $|\mathcal{K}|$ weights $q_{ik}, \quad k \in \mathcal{K}$ associated with a given stimuli $i \in \mathcal{I}$ can be assimilated to the actual distribution of the ratings for that stimuli. Such a distribution could allow to perform statistical tests to verify whether two stimuli have qualities that differ significantly.

To determine a suitable λ value we start from the following considerations. The λ value has to be: i) proportional to the number of stimuli, to account for the noise caused by subjects' fatigue; ii) inversely proportional to the number of subjects, since the larger is the number of subject, the more the dataset is informative and hence the Log-likelihood function $LL(q)$ must have more importance than the regularization term $R(q)$; iii) proportional to the number of possible opinion scores available on the quality scale as one expects subjects to vote more consistently when facing less choices. A typical example is the greater reliability of subjects in pair comparison-based tests.

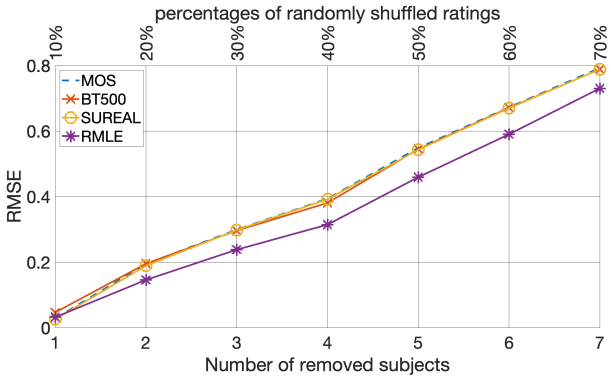
In our experiments, λ was set to

$$\lambda = \frac{1}{2} \cdot \frac{|\mathcal{I}||\mathcal{K}|}{|\mathcal{J}|} \tag{8}$$

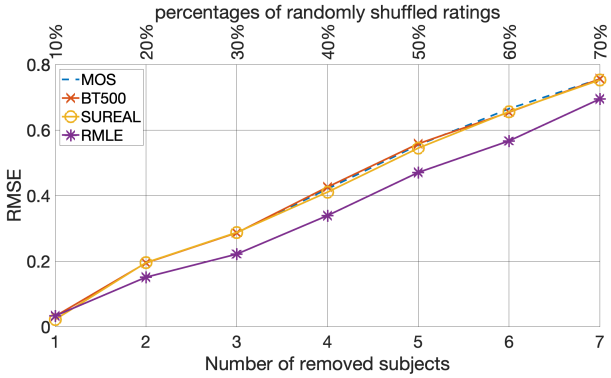
The constant $\frac{1}{2}$ was experimentally found to be a proportionality factor that guarantees the greatest robustness. We are aware that this approach should be considered preliminary and could be further refined. However, preliminary results seems encouraging and we will consider further refinements in a future work.



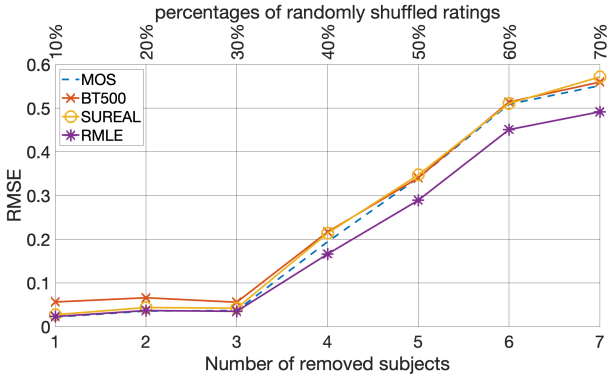
(a) VQEGHD1



(b) VQEGHD3



(c) VQEGHD5



(d) NETFLIX PUBLIC

Fig. 1. Robustness of the different methods to the noise caused by the reduction of the number of subjects and a random shuffle of a certain percentage of opinion scores. The experiments were run with 20 different seeds and the average RMSE is shown for each recovery method.

III. Results

We compared the proposed RMLE approach to three state-of-the-art approaches, i.e., the MOS, the ITU-T BT.500 and the MLE model implemented by the Sureal software [6], in terms of robustness to the noise. The experiments are conducted on four datasets, i.e., the VQEGHD1, VQEGHD3, VQEGHD5 [8] and the Netflix public dataset [3].

The four considered datasets were obtained from subjective experiments conducted in highly controlled environments under conditions specified by the ITU-T Recommendations, thus minimizing the sources of noise. Therefore, following the approach of [3], [4], we considered, as ground truth, the subjective quality recovered by each method on the original datasets, i.e. without adding noise to individual opinion scores.

We then evaluated how much each approach is robust to the noise synthetically added to the dataset. The noise was added to the individual ratings in each of the considered datasets by using approaches similar to those adopted in [3], [4], i.e., reducing the number of subjects and randomly shuffling, i.e. permuting at random, a certain percentage of opinion scores of the remaining subjects. Then, the different quality recovery methods were run on the noisy data. The Root Mean Square Error (RMSE) between the recovered quality from noisy ratings and the ground truth quality, i.e., the one recovered under noiseless conditions, was then computed for each quality recovery method.

The results are shown in Figure 1. The less noisy condition consisted in removing one subject and randomly shuffling 10% of the subjects' opinion scores. Then two subjects were removed and 20% of the ratings was randomly shuffled. We proceeded that way until the most noisy situation in which 7 subjects were removed and 70% of the ratings was shuffled at random. In all noisy conditions the RMLE recovered a quality score with the lowest RMSE with respect to the ground truth quality. In fact, the curve associated with the proposed RMLE approach lies below the ones of all the other quality recovery methods.

IV. Conclusions

In this work, a new approach to recover the ground truth subjective quality from potentially noisy individual opinion scores was proposed. The novelty of the approach relies on designing a specific regularization term to be used in a likelihood maximization framework. The new regularization term is designed to force the attribution of less importance to potentially noisy ratings. Computational experiments showed that our proposal offers greater robustness when dealing with noisy opinion scores compared to other state-of-the-art methods.

REFERENCES

[1] ITU-T Rec. BT.500, "Methodology for the subjective assessment of the quality of television pictures," Jan. 2012.

[2] L. Janowski and M. Pinson, "The accuracy of subjects in a quality experiment: A theoretical subject model," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2210–2224, 2015.

[3] Z. Li and C. G. Bampis, "Recover subjective quality scores from noisy measurements," in *2017 Data Compression Conference (DCC)*, April 2017, pp. 52–61.

[4] J. Li, S. Ling, J. Wang, and P. Le Callet, "A probabilistic graphical model for analyzing the subjective visual quality assessment data from crowdsourcing," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3339–3347.

[5] S. Pezzulli, M. G. Martini, and N. Barman, "Estimation of quality scores from subjective tests: beyond subjects' mos," *IEEE Transactions on Multimedia*, 2020.

[6] Netflix, "The sureal software," https://github.com/Netflix/sureal, May 2017.

[7] F. M. Reza, *An introduction to information theory*. Courier Corporation, 1994.

[8] VQEG, "Report on the validation of video quality models for high definition video content (v. 2.0)," http://bit.ly/2Z7GWDI, Jun. 2010.