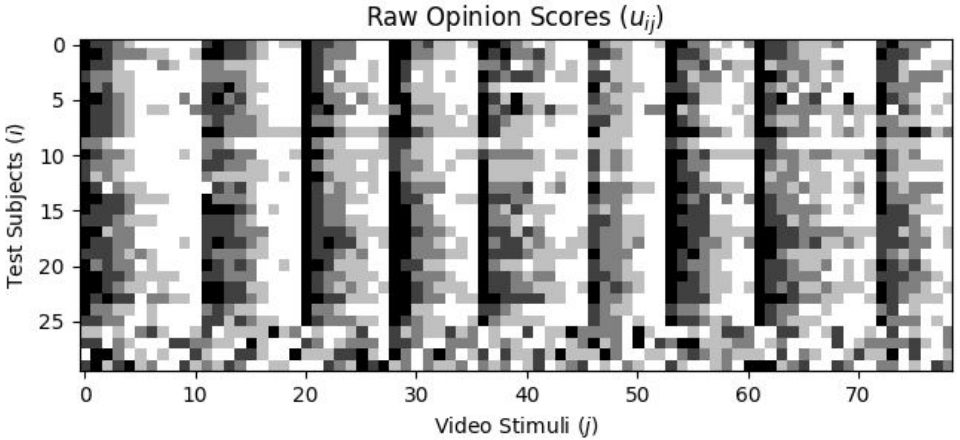
Goal: estimate true quality scores of video stimuli from noisy raw ratings

Black: lowest score
White: highest score



Subject bias

Last 4 Outliers

*NFLX Public Dataset

Proposed Solver

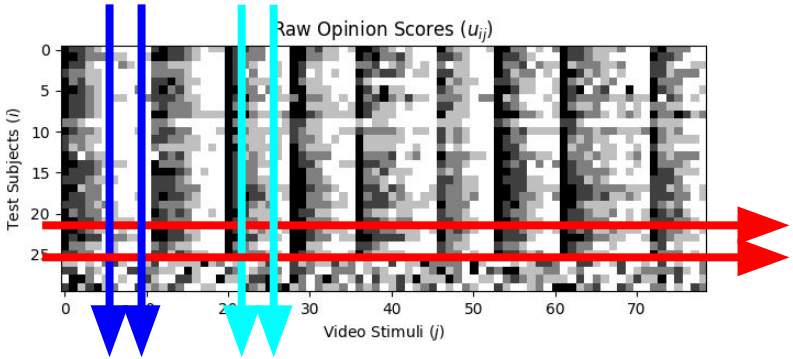
- Input:
  - u_{ijr} for subject $i = 1,\dots,I$, stimulus $j = 1,\dots,J$ and repetition $r = 1,\dots,R$
  - Stop threshold ψ^{thr}.
- Initialize $\{\psi_j\} \leftarrow \{MOS_j\}$, where $MOS_j = (\sum_{ir} 1)^{-1} \sum_{ir} u_{ijr}$.
- Initialize $\{\Delta_i\} \leftarrow \{BIAS_i\}$, where $BIAS_i = (\sum_{jr} 1)^{-1} \sum_{jr} (u_{ijr} - MOS_j)$.
- Loop:
  - $\{\psi_j{}^{prev}\} \leftarrow \{\psi_j\}$.
  - $\epsilon_{ijr} \leftarrow u_{ijr} - \psi_j - \Delta_i$ for $i = 1,\dots,I, j = 1,\dots,J$ and $r = 1,\dots,R$.
  - $v_i \leftarrow \sigma_i\{\epsilon_{ijr}\}$, where $\sigma_i\{\epsilon_{ijr}\} = \sqrt{(\sum_{jr} 1)^{-1} \sum_{jr} (\epsilon_{ijr} - \epsilon_i)^2 - \epsilon_i{}^2}$ and $\epsilon_i = (\sum_{jr} 1)^{-1} \sum_{jr} \epsilon_{ijr}$, for $i = 1,\dots,I$.
  - $\psi_j \leftarrow (\sum_{ir} v_i{}^{-2})^{-1} \sum_{ir} v_i{}^{-2} (u_{ijr} - \Delta_i)$, for $j = 1,\dots,J$.
  - $\Delta_i \leftarrow \left(\sum_{jr} 1\right)^{-1} \sum_{jr} (u_{ijr} - \psi_j)$, for $i = 1,\dots,I$.
  - If $\sqrt{\sum_{j=1}^{J} (\psi_j - \psi_j{}^{prev})^2} < \psi^{thr}$, break.
- Output: $\{\psi_j\}$, $\{\Delta_i\}$, $\{v_i\}$.

Alternating Projection (AP) Solver

1. Video by video, estimate MOS by averaging over subjects
2. Subject by subject, estimate subject bias by comparing against the MOS

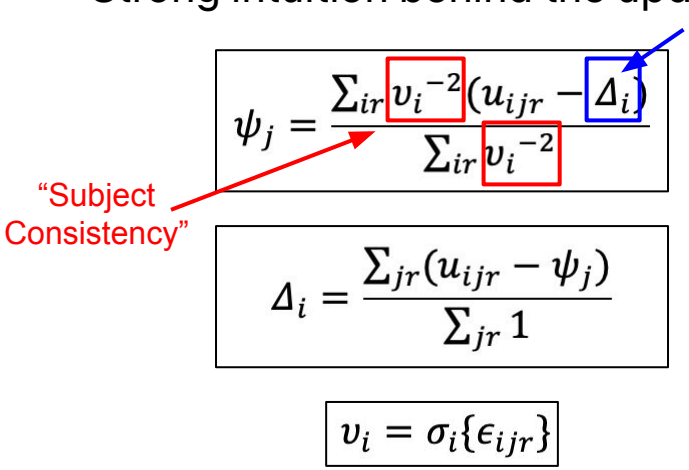
In a loop:

a. Subject by subject, estimate subject inconsistency as the std of the residue of raw scores
b. Repeat step 1 (with weighting).
c. Repeat step 2.
d. If solution stabilizes, break



Proposed Solver - Interpretation

- Strong intuition behind the updating steps



Quality are weighted by "subject consistency" (v_i^{-2}) after the subject bias (Δ_i) is removed. The "subject consistency" is the inverse of the (squared) subject inconsistency (v_i^2)*.

Subject bias (Δ_i) as the mean of the opinion scores after the true quality (ψ_j) removed.

Subject inconsistency as the standard deviation of the estimation residue (ϵ_i).

- The new method can be interpreted as computing the **bias-subtracted consistency-weighted MOS**

*In practical implementation, we add a small ε to make the denominator non-zero.

Proposed Text Change in ITU-T P.913

represent the larger pool of all people. Thus, their scores can be aggregated without applying any scaling or fitting function.

12.6 Improve MOS or DMOS data quality under challenging test conditions

Very often a subjective test needs to be run under challenging conditions. For example, in a crowdsourcing test, the subjects are exposed to an environment that is less controlled than in a laboratory. In a large-scale test conducted by multiple laboratories, inter-lab variability could result in large variance of the ratings collected. Traditional data analysis tools provided by [ITU-T P.910], [b-ITU-T P.911] and [ITU-R BT.500-13] often do not work well under such circumstances. In this clause, an advanced data analysis technique is described, which has shown improvement on the data quality of the MOS or DMOS calculated. See [b-Li, 2017] [b-Li, 2020] for equations, software and evidence for this technique's validity. A reference Python implementation can also be found in Appendix III.

The intuition behind this technique is the following. It is useful to explicitly model each subject's behaviour; in particular, a subject's bias and consistency are two prominent human factors that affect the subject's votes. Through an iterative procedure, this technique tries to jointly estimate the true quality of each PVS and the bias and consistency of each subject. The estimated true quality of each PVS can be interpreted as a "bias-removed consistency-weighted MOS". Compared to the post-screening of subjects described in clause 11.4, which either keep or reject all votes of a subject ("hard rejection"), this technique can be described as "soft rejection". That is, for an outlier subject who votes inconsistently, the subject's votes would carry a small weight, hence contributing little to the overall MOS.

A byproduct of this technique is the estimation of each test subject's bias and consistency. These are valuable information for a subject's suitability for performing subjective tests, hence can be used to screen subjects for future tests. For example, if a subject has shown to vote highly inconsistently, he/she may be excluded from future sessions.

This technique can be considered as generalizing the subject-bias removal described in clause 12.4 (notice the similarity between the two).

First, estimate the MOS for each PVS:

$$\mu_{\psi_j} = \frac{1}{I_j}\sum_{i=1}^{I_j} o_{ij}$$

where:

- o_{ij} is the observed rating for subject i and PVS j;
- I_j is the number of subjects that rated PVS j;
- μ_{ψ_j} estimates the MOS for PVS j, given the source stimuli and subjects in the experiment.

Second, estimate subject bias:

$$\mu_{\Delta_i} = \frac{1}{J_i}\sum_{j=1}^{J_i}\left(o_{ij} - \mu_{\psi_j}\right)$$

where:

- μ_{Δ_i} estimates the overall shift between the ith subject's scores and the true values (i.e., opinion bias)
- J_i is the number of PVSs rated by subject i.

Third, do the following in a loop:

- Record the current estimate of the MOS for each PVS:

$$\mu_{\psi_j}^{c} = \mu_{\psi_j}$$

- Calculate the residue in each observed rating not explained by the MOS and the subject bias:

$$r_{ij} = o_{ij} - \mu_{\psi_j} - \mu_{\Delta_i}$$

- Estimate the subject inconsistency (i.e., the reciprocal of consistency) as the per-subject standard deviation of the residues:

$$\sigma_{r_i} = \sqrt{\frac{1}{J_i}\sum_{j=1}^{J_i}\left(r_{ij} - \mu_{r_j}\right)^2}$$

where:

$$\mu_{r_i} = \frac{1}{J_i}\sum_{j=1}^{J_i} r_{ij}$$

- Estimate the new MOS for each PVS as the bias-removed consistency-weighted mean ratings:

$$\mu_{\psi_j} = \frac{\sum_{i=1}^{I_j}\sigma_{r_i}^{-2}\left(o_{ij} - \mu_{\Delta_i}\right)}{\sum_{i=1}^{I_j}\sigma_{r_i}^{-2}}$$

where:

- $\sigma_{r_i}^{-2}$ is the (squared) consistency of subject i;
- $o_{ij} - \mu_{\Delta_i}$ is the bias-removed rating of subject i on PVS j.

- Estimate the new subject bias the same way as before:

$$\mu_{\Delta_i} = \frac{1}{J_i}\sum_{j=1}^{J_i}\left(o_{ij} - \mu_{\psi_j}\right)$$

- Terminate the loop if:

$$\sum_{j=1}^{J}\left(\mu_{\psi_j} - \mu_{\psi_j}^{c}\right)^2 < 10^{-16}.$$

Once the procedure ends, the final MOS of PVS j is simply μ_{ψ_j}. The standard deviation of score (SOS) for PVS j is computed as:

$$\mathrm{SOS}_j = \frac{\sigma_{r_j}}{\sqrt{I_j}}$$

where

$$\sigma_{r_j} = \sqrt{\frac{1}{I_j}\sum_{i=1}^{I_j}\left(r_{ij} - \mu_{r_j}\right)^2}$$

and:

$$\mu_{r_j} = \frac{1}{I_j}\sum_{i=1}^{I_j} r_{ij}.$$

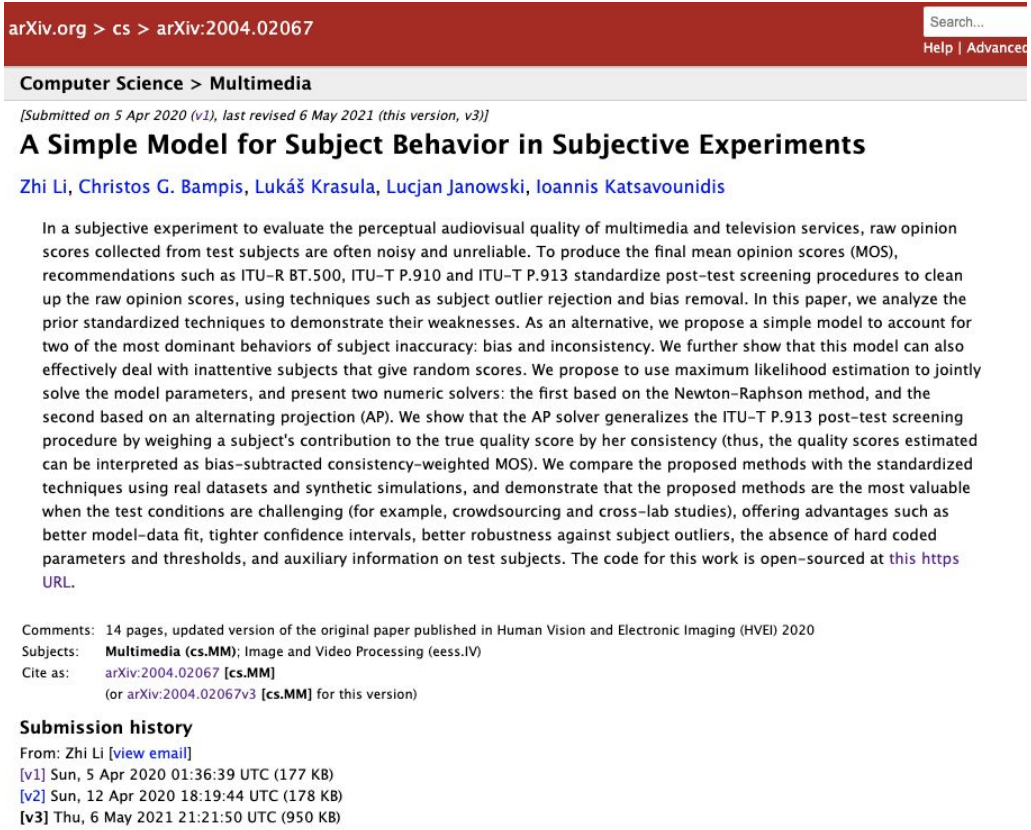
The DMOS and the corresponding SOS can be calculated similarly.

13 Elements of subjective test reporting

Reports on subjective testing are more effective when descriptions of both mandatory and optional elements defining the test are included. A full description of all the elements of the subjective test supports the conclusions from the test.

Adolph, Martin Deleted: 03
Adolph, Martin Deleted: 2016
Adolph, Martin Deleted: 03
Adolph, Martin Deleted: 2016

Publication & Open Source Code



arXiv.org > cs > arXiv:2004.02067

Search... Help | Advanced

Computer Science > Multimedia

[Submitted on 5 Apr 2020 (v1), last revised 6 May 2021 (this version, v3)]

A Simple Model for Subject Behavior in Subjective Experiments

Zhi Li, Christos G. Bampis, Lukáš Krasula, Lucjan Janowski, Ioannis Katsavounidis

In a subjective experiment to evaluate the perceptual audiovisual quality of multimedia and television services, raw opinion scores collected from test subjects are often noisy and unreliable. To produce the final mean opinion scores (MOS), recommendations such as ITU-R BT.500, ITU-T P.910 and ITU-T P.913 standardize post-test screening procedures to clean up the raw opinion scores, using techniques such as subject outlier rejection and bias removal. In this paper, we analyze the prior standardized techniques to demonstrate their weaknesses. As an alternative, we propose a simple model to account for two of the most dominant behaviors of subject inaccuracy: bias and inconsistency. We further show that this model can also effectively deal with inattentive subjects that give random scores. We propose to use maximum likelihood estimation to jointly solve the model parameters, and present two numeric solvers: the first based on the Newton-Raphson method, and the second based on an alternating projection (AP). We show that the AP solver generalizes the ITU-T P.913 post-test screening procedure by weighing a subject's contribution to the true quality score by her consistency (thus, the quality scores estimated can be interpreted as bias-subtracted consistency-weighted MOS). We compare the proposed methods with the standardized techniques using real datasets and synthetic simulations, and demonstrate that the proposed methods are the most valuable when the test conditions are challenging (for example, crowdsourcing and cross-lab studies), offering advantages such as better model-data fit, tighter confidence intervals, better robustness against subject outliers, the absence of hard coded parameters and thresholds, and auxiliary information on test subjects. The code for this work is open-sourced at this https URL.

Comments: 14 pages, updated version of the original paper published in Human Vision and Electronic Imaging (HVEI) 2020
Subjects: **Multimedia (cs.MM)**; Image and Video Processing (eess.IV)
Cite as: arXiv:2004.02067 **[cs.MM]**
(or arXiv:2004.02067v3 **[cs.MM]** for this version)

Submission history

From: Zhi Li [view email]
[v1] Sun, 5 Apr 2020 01:36:39 UTC (177 KB)
[v2] Sun, 12 Apr 2020 18:19:44 UTC (178 KB)
[v3] Thu, 6 May 2021 21:21:50 UTC (950 KB)

Whitepaper available at: https://arxiv.org/abs/2004.02067

Source code - *free and open-sourced*: https://github.com/Netflix/sureal/tree/master/itut_p913_demo