A Simple Model for Subject Behavior in Subjective Experiments

Zhi Li (Netflix), Christos G. Bampis (Netflix), Lucjan Janowski (AGH, Poland) and Ioannis Katsavounidis (Facebook)

Abstract

In a subjective experiment to evaluate the perceptual audiovisual quality of multimedia and television services, raw opinion scores offered by subjects are often noisy and unreliable. Recommendations such as ITU-R BT.500, ITU-T P.910 and ITU-T P.913 standardize post-processing 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 (aka systematic error) and inconsistency (aka random error). We further show that this model can also effectively deal with inattentive subjects that give random scores. We propose to use maximum likelihood estimation (MLE) to jointly estimate the model parameters, and present two numeric solvers: the first based on the Newton-Raphson method, and the second based on alternating projection. We show that the second solver can be considered as a generalization of the subject bias removal procedure in ITU-T P.913. We compare the proposed methods with the standardized techniques using real datasets and synthetic simulations, and demonstrate that the proposed methods have advantages in better model-data fit, tighter confidence intervals, better robustness against subject outliers, shorter runtime, the absence of hard coded parameters and thresholds, and auxiliary information on test subjects. The source code for this work is open-sourced at https://github.com/Netflix/sureal.

Introduction

Subjective experiment methodologies to evaluate the perceptual audiovisual quality of multimedia and television services have been well studied. Recommendations such as ITU-R BT.500 [1], ITU-T P.910 [2] and ITU-T P.913 [3] standardize the procedures to conduct subjective experiments and post-process raw opinion scores to yield the final mean opinion scores (MOS). To account for the inherently noisy and often unreliable nature of test subjects, the recommendations have included corrective mechanisms such as subject rejection (BT.500), subject bias removal (P.913), and criteria for establishing the confidence intervals of the MOS (BT.500, P.910 and P.913). The standardized procedures are not without their own limitations. For example, in BT.500, if a subject is deemed an outlier, all opinion scores for that subject are discarded, which could be excessive. The BT.500 procedure also incorporates a number of hard coded parameters and thresholds, which may not be suitable for all conditions and subjective tests.

As an alternative, we propose a simple model to account for two of the most dominant behaviors of test subject inaccuracy: *bias* and *inconsistency*. In addition, this model can effectively deal with inattentive subject outliers that give random scores. Compared to BT.500-style subject rejection, the proposed model can be thought as performing "soft" subject rejection as it explicitly models subject outliers as having large inconsistencies, and

thus is able to diminish the effect from the estimated true quality scores. To solve for the model parameters, we propose to jointly optimize its likelihood function, also known as maximum likelihood estimation (MLE) [4]. We present two numeric solvers: 1) a Newton-Raphson (NR) solver [5], and 2) an Alternating Projection (AP) solver. We further show that the AP solver can be considered as a weighted and iterative generalization of the subject bias removal procedure in P.913. The AP solver also has the advantage of *having no hard coded parameters and thresholds*.

One of the challenges is to fairly compare the proposed methods to its alternatives. We evaluate the proposed simple model and its numerical solvers separately. To evaluate the model's fit to real datasets, we use Bayesian Information Criterion (BIC) [6], where the winner can be characterized as *having a good fit to data while maintaining a small number of parameters*. We also compare the confidence intervals of the estimated quality scores, where a tighter confidence interval implies a higher confidence in the estimation. To evaluate the model's robustness against subject outliers, we perform a simulation study on how the true quality score's root mean squared error (RMSE) changes compared to the clean case as the number of outliers increases. Lastly, to validate that the numerical solvers are indeed accurate, we use synthetic data to compare the recovered parameters against the ground truth.

The rest of the paper is organized as follows. Section Prior Art and Standards discusses prior art and standards. The proposed model is presented in Section Proposed Model, followed by the two numerical solvers in Section Proposed Solvers, and the calculation of confidence intervals in Section Confidence Interval. Section Experimental Results presents the experimental results.

The source code of this work is open-sourced on Github [7].

Prior Art and Standards

Raw opinion scores collected from subjective experiments are known to be influenced by the inherently noisy and unreliable nature of human test subjects [8]. To compensate for the influence of individuals, a common practice is to average the raw opinion scores from multiple subjects, yielding a MOS per stimulus. Standardized recommendations incorporate more advanced corrective mechanisms to further compensate for test subjects' influence, and criteria for establishing the confidence intervals of MOS.

- ITU-R BT.500 Recommendation [1] defines methodologies including double-stimulus impairment scale (DSIS) and double-stimulus continuous quality scale (DSCQS), and a corresponding procedure for subject rejection (ITU-R BT.500-14 Section A1-2.3.1) prior to the calculation of MOS. Video by video, the procedure counts the number of instances when a subject's opinion score deviates by a few sigmas (i.e. standard deviation), and rejects the subject if

the occurrences are more than a fraction. All scores corresponding to the rejected subjects are discarded, which could be considered excessive. On the other hand, our experiment shows that, in the presence of many outlier subjects, the procedure is only able to identify a portion of them. Another drawback of this approach is that it incorporates a number of hard coded parameters and thresholds to determine the outliers, which may not be suitable for all conditions. It also establishes the corresponding way of calculating the confidence interval (ITU-R BT.500-14 Section A1-2.2.1).

- ITU-T P.910 Recommendation [2] defines methodologies including absolute category rating (ACR), degradation category rating DCR (equivalent to DSIS), absolute category rating with hidden reference (ACR-HR) and the corresponding differential MOS (DMOS) calculation, and recommends using the BT.500 subject rejection and confidence interval calculation procedure in conjunction.

- ITU-T P.913 Recommendation [3] defines a procedure to remove subject bias (ITU-T P.913 Section 12.4) before carrying out other steps. It first finds the mean score per stimulus, and subtract it from the raw opinion scores to get the residual scores. Then it averages the residue scores on a per-subject basis to yield an estimate of each subject's bias. The subject bias is then removed from the raw opinion scores. For P.913 to possess resistance to subject outliers, it needs to be combined with BT.500-style subject rejection. Yet, by doing so, it poses similar weaknesses as BT.500.

For completeness, below we give mathematical descriptions of the subject rejection method standardized in ITU-R BT.500-14 and the subject bias removal method in ITU-T P.913. Let u_{ijr} be the opinion score voted by subject i on stimulus j in repetition r. Note that, in BT.500-14, the notation j is used to indicate test condition and k is used to indicate sequence/image; in this paper, the test condition and sequence/image are combined and collectively represented by the stimulus notation j. Let μ_{jr} denote the mean value over scores for stimulus j and for repetition r, i.e. $\mu_{jr} = (\sum_i 1)^{-1} \sum_i u_{ijr}$. Similarly, $m_{n,jr}$ denotes the n-th order central moment over scores for stimulus j and repetition r, i.e. $m_{n,jr} = (\sum_i 1)^{-1} \sum_i (u_{ijr} - \mu_{jr})^n$. Lastly, σ_{jr} denotes the sample standard deviation for stimulus j and repetition r, i.e. $\sigma_{jr} = \sqrt{((\sum_i 1) - 1)^{-1} \sum_i (u_{ijr} - \mu_{jr})^2}$. In the previous, the term $\sum_i 1$ indicates the number of observers that have offered an opinion score for a given stimulus/repetition, jr. This number of observers could be the same, I, or different per stimulus, if a subjective experiment has been designed in such a way. The subject rejection procedure in ITU-R BT.500-14 Section A1-2.3 can be summarized in Algorithm 1.

ITU-T P.913 does not consider repetitions, so the notation u_{ij} denotes the opinion score voted by subject i on stimulus j. The subject bias removal procedure in ITU-T P.913 Section 12.4 can be summarized in Algorithm 2.

Proposed Model

We propose a simple yet effective model to account for two of the most dominant effects of test subject inaccuracy: subject bias and subject inconsistency. We further show that this model can effectively deal with inattentive subjects that give random scores, without invoking explicit subject rejection. The proposed

---

Algorithm 1 ITU-R BT.500 Subject Rejection [1]

- Input: u_{ijr} for $i = 1, ..., I$, $j = 1, ..., J$ and $r = 1, ..., R$.
- Initialize $p(i) \leftarrow 0$ and $q(i) \leftarrow 0$ for $i = 1, ..., I$.
- For $j = 1, ..., J$, $r = 1, ..., R$:
    - Let $Kurtosis_{jr} = \frac{m_{4,jr}}{(m_{2,jr})^2}$.
    - If $2 \leq Kurtosis_{jr} \leq 4$, then $\varepsilon_{jr} = 2$; otherwise $\varepsilon_{jr} = \sqrt{20}$.
    - For $i = 1, ..., I$:
        * If $u_{ijr} \geq \mu_{jr} + \varepsilon_{jr}\sigma_{jr}$, then $p(i) \leftarrow p(i) + 1$.
        * If $u_{ijr} \leq \mu_{jr} - \varepsilon_{jr}\sigma_{jr}$, then $q(i) \leftarrow q(i) + 1$.
- Initialize $Set_{rej} = \emptyset$.
- For $i = 1, ..., I$:
    - If $\frac{p(i)+q(i)}{\sum_{jr} 1} \geq 0.05$ and $\left| \frac{p(i)-q(i)}{p(i)+q(i)} \right| < 0.3$, then $Set_{rej} \leftarrow Set_{rej} \cup \{i\}$.
- Output: Set_{rej}.

Algorithm 2 ITU-T P.913 Subject Bias Removal [3]

- Input:
    - u_{ij} for subject $i = 1, ..., I$, stimulus $j = 1, ..., J$.
- For $j = 1, ..., J$:
    - Estimate MOS of stimulus as $MOS_j = (\sum_i 1)^{-1} \sum_i u_{ij}$.
- For $i = 1, ..., I$:
    - Estimate subject bias as $BIAS_i = (\sum_j 1)^{-1} \sum_j (u_{ij} - MOS_j)$.
- Calculate the subject bias-removed opinion scores $r_{ij} = u_{ij} - BIAS_i$, $i = 1, ..., I$, $j = 1, ..., J$.
- Use r_{ij} instead of u_{ij} as the opinion scores to carry out the remaining steps.

---

model is a simplified version of [9] without considering the ambiguity of video content. Compared to the previously proposed model, the solutions to the simplified model are more efficient and stable.

We assume that each opinion score u_{ijr} can be represented by a random variable as follows:

$$U_{ijr} = \psi_j + \Delta_i + \upsilon_i X, \tag{1}$$

where ψ_j is the true quality of stimulus j, Δ_i represents the bias of subject i, the non-negative term υ_i represents the inconsistency of subject i, and $X \sim N(0,1)$ are i.i.d. Gaussian random variables. The index r represents repetitions.

It is important to point out that a subject with erroneous behaviors can be modeled by a large inconsistency value υ_i. The erroneous behaviors that can be modeled include but are not limited to: subject giving random scores, subject being absent-minded for a portion of a session, or software issue that randomly shuffles a subject's scores among multiple stimuli. By successfully estimating υ_i and accounting its effect to calculating the true quality score, we can compensate for subject outliers without invoking BT.500-style subject rejection.

Given a collection of opinion scores $\{u_{ijr}\}$ from a subjective experiment, the task is to solve for the free parameters $\theta =$

$(\{\psi_j\}, \{\Delta_i\}, \{v_i\})$, such that the model fits the observed scores the best. This can be formulated as a maximum likelihood estimation (MLE) problem. Let the log-likelihood function be

$$L(\theta) = \log P(\{u_{ijr}\} | \{\psi_j\}, \{\Delta_i\}, \{v_i\}),$$

i.e. a monotonic measure of the probability of observing the given raw scores, for a set of these parameters. We can solve the model by finding $\hat{\theta}$ that maximizes $L(\theta)$, or $\hat{\theta} = \arg\max L(\theta)$. This problem can be numerically solved by the proposed Newton-Raphson method or the Alternating Projection method, to be discussed in Section Proposed Solvers.

It is important to notice that the recoverability of $\{\psi_j\}$ and $\{\Delta_i\}$ in (1) is up to a constant shift. Formally, assume $\hat{\theta} = (\{\hat{\psi}_j\}, \{\hat{\Delta}_i\}, \{\hat{v}_i\})$ is a solution that maximizes $L(\theta)$, one can easily show that $(\{\hat{\psi}_j + C\}, \{\hat{\Delta}_i - C\}, \{\hat{v}_i\})$ where $C \in \mathbb{R}$, is another solution that achieves the same maximum likelihood value $L(\hat{\theta})$. This implies that the optimal solution is not unique. In practice, we can enforce a unique solution, by adding a constraint that forces the mean subject bias to be zero, or

$$\sum_i \Delta_i = 0.$$

This intuitively makes sense, since bias is relative - saying everyone is positively biased is equivalent to saying that no one is positively biased. It is also equivalent to assuming that the sample of observers that offer opinion scores in a subjective experiment are truly random and do not consist of only "expert" viewers or "lazy" viewers that tend to offer lower or higher opinion scores, as a whole. There is always the possibility, once a subjective test establishes that the population from where subjects were recruited have such a collective bias, to change the condition and thus properly estimate what the true "typical" observer, drawn from a more representative pool that would vote.

Lastly, one should keep in mind that it is always possible to use more complicated models than (1) to capture other effects in a subjective experiment. For example, [9] considers content ambiguity, and [10, 11] considers per-stimulus ambiguity. There are also environment-related factors that could induce biases. Additionally, the votes are influenced by the voting scales chosen, for example, continuous vs. discrete [12]. Our hope is that the proposed model strikes a good balance between the model complexity and explanatory power. In Section Model-Data Fit, we show that the proposed model yields better model-data fit than BT.500 and P.913 used today.

Proposed Solvers

Let us start by simplifying the form of the log-likelihood function $L(\theta)$. We can write:

$$
\begin{aligned}
L(\theta) &= \log P(\{u_{ijr}\} | \{\psi_j\}, \{\Delta_i\}, \{v_i\}) \\
&= \log \prod_{ijr} P(u_{ijr} | \psi_j, \Delta_i, v_i) \quad (2) \\
&= \sum_{ijr} \log P(u_{ijr} | \psi_j, \Delta_i, v_i) \\
&= \sum_{ijr} \log f(u_{ijr} | \psi_j + \Delta_i, v_i) \\
&\cong \sum_{ijr} -\log v_i - \frac{(u_{ijr} - \psi_j - \Delta_i)^2}{2v_i^2} \quad (3)
\end{aligned}
$$

Algorithm 3 Proposed Newton-Raphson (NR) solver

- Input:
  - u_{ijr} for subject $i = 1, ..., I$, stimulus $j = 1, ..., J$ and repetition $r = 1, ..., R$.
  - Refresh rate α.
  - Stop threshold ψ^{thr}.
- Initialize $\{\Delta_i\} \leftarrow \{0\}$, $\{\psi_j\} \leftarrow \{MOS_j\}$, $\{v_i\} \leftarrow \{RSD_i\}$.
- Loop:
  - $\{\psi_j^{prev}\} \leftarrow \{\psi_j\}$.
  - $\Delta_i \leftarrow (1 - \alpha) \cdot \Delta_i + \alpha \cdot \Delta_i^{new}$ where $\Delta_i^{new} = \Delta_i - \frac{\partial L(\theta)/\partial \Delta_i}{\partial^2 L(\theta)/\partial \Delta_i^2}$ for $i = 1, ..., I$.
  - $v_i \leftarrow (1 - \alpha) \cdot v_i + \alpha \cdot v_i^{new}$ where $v_i^{new} = v_i - \frac{\partial L(\theta)/\partial v_i}{\partial^2 L(\theta)/\partial v_i^2}$ for $i = 1, ..., I$.
  - $\psi_j \leftarrow (1 - \alpha) \cdot \psi_j + \alpha \cdot \psi_j^{new}$ where $\psi_j^{new} = \psi_j - \frac{\partial L(\theta)/\partial \psi_j}{\partial^2 L(\theta)/\partial \psi_j^2}$ for $j = 1, ..., J$.
  - If $\left(\sum_{j=1}^{J} (\psi_j - \psi_j^{prev})^2 \right)^{\frac{1}{2}} < \psi^{thr}$, break.
- Output: $\{\psi_j\}$, $\{\Delta_i\}$, $\{v_i\}$.

---

where (2) uses the independence assumption on opinion scores, $f(x | \mu, v)$ is the Gaussian density function with mean μ and standard deviation v, and $\cong$ denotes equal with omission of constant terms.

Note that not every subject needs to vote on each stimulus in every repetition. Our proposed solvers can effectively deal with subjective tests with incomplete data where some observations u_{ijr} are missing. Denote by $\star$ the missing observations in an experiment. All summations in this paper are ignoring the missing observations $\star$, that is, $\sum_{ijr}$ is equivalent to $\sum_{ijr:u_{ijr} \neq \star}$, and so on.

Newton-Raphson (NR) Solver

With (3), the first- and second-order partial derivatives of $L(\theta)$ can be derived (see the first Appendix). We can apply the Newton-Raphson rule [5] $a^{new} \leftarrow a - \frac{\partial L/\partial a}{\partial^2 L/\partial a^2}$ to update each parameter a in iterations. We further use a refresh rate parameter α to control the speed of innovation and to avoid overshooting. Note that other update rules can be applied, but using the Newton-Raphson rule yields nice interpretability.

Also note that the NR solver finds a local optimal solution when the problem is non-convex. It is important to initialize the parameters properly. We choose zeros as the initial values for $\{\Delta_i\}$, the mean score $MOS_j = (\sum_{ir} 1)^{-1} \sum_{ir} u_{ijr}$ for $\{\psi_j\}$, and the residue standard deviation $RSD_i = \sigma_i(\{\varepsilon_{ijr}\})$ for $\{v_i\}$, where $\varepsilon_{ijr} = u_{ijr} - MOS_j$ is the "residue", $\sigma_i(\{\varepsilon_{ijr}\}) = \sqrt{(\sum_{jr} 1)^{-1} \sum_{jr} (\varepsilon_{ijr} - \varepsilon_i)^2 - \varepsilon_i^2}$, and $\varepsilon_i = (\sum_{jr} 1)^{-1} \sum_{jr} \varepsilon_{ijr}$. The NR solver is summarized in Algorithm 3. A good choice of refresh rate and stop threshold are $\alpha = 0.1$ and $\psi^{thr} = 1e^{-9}$, respectively, but varying these parameters would not significantly change the result.

The "new" parameters can be simplified to the following

form:

$$\psi_j^{new} = \frac{\sum_{ir} \upsilon_i^{-2}(u_{ijr} - \Delta_i)}{\sum_{ir} \upsilon_i^{-2}}, \tag{4}$$

$$\Delta_i^{new} = \frac{\sum_{jr}(u_{ijr} - \psi_j)}{\sum_{jr} 1}, \tag{5}$$

$$\upsilon_i^{new} = \upsilon_i \frac{\sum_{jr} 2\upsilon_i^2 - 4(u_{ijr} - \psi_j - \Delta_i)^2}{\sum_{jr} \upsilon_i^2 - 3(u_{ijr} - \psi_j - \Delta_i)^2}.$$

Note that there are strong intuitions behind the expressions for the newly estimated true quality ψ_j^{new} and subject bias Δ_i^{new}. In each iteration, ψ_j^{new} is re-estimated, as the weighted mean of the opinion scores u_{ijr} with the currently estimated subject bias Δ_i removed. Each opinion score is weighted by the "subject consistency" υ_i^{-2}, i.e., the higher the inconsistency for subject i, the less reliable the opinion score, hence less the weight. For the subject bias Δ_i^{new}, it is simply the average shift between subject i's opinion scores and the true values.

Alternating Projection (AP) Solver

This solver is called "alternating projection" because in a loop, we alternate between projecting (or averaging) the opinion scores along the subject dimension and the stimulus dimension. To start, we initialize $\{\psi_j\}$ to $\{MOS_j\}$, where $MOS_j = (\sum_{ir} 1)^{-1} \sum_{ir} u_{ijr}$, same as the NR solver. The subject bias $\{\Delta_i\}$ is initialized differently to $\{BIAS_i\}$, where $BIAS_i = (\sum_{jr} 1)^{-1} \sum_{jr}(u_{ijr} - MOS_j)$ is the average shift between subject i's opinion scores and the true values. Note that the calculation of $\{MOS_j\}$ and $\{BIAS_i\}$ matches precisely to the ones in Algorithm 2 (ITU-T P.913). Within the loop, first, the "residue" ε_{ijr} is updated, followed by the calculation of the subject inconsistency υ_i as the residue's standard deviation per subject, or $\sigma_i(\{\varepsilon_{ijr}\}) = \sqrt{(\sum_{jr} 1)^{-1} \sum_{jr}(\varepsilon_{ijr} - \varepsilon_i)^2 - \varepsilon_i^2}$, where $\varepsilon_i = (\sum_{jr} 1)^{-1} \sum_{jr} \varepsilon_{ijr}$. Then, the true quality $\{\psi_j\}$ and the subject bias $\{\Delta_i\}$ are re-estimated, by averaging the opinion scores along either the subject dimension i or the stimulus dimension j. The projection formula precisely matches equations (4) and (5) of the Newton-Raphson method. The AP solver is summarized in Algorithm 4. A good choice of the stop threshold is $\psi^{thr} = 1e^{-8}$.

In sum, the AP solver can be considered as a generalization of P.913 Section 12.4 in the following sense: first, the AP solver is iterative until convergence whereas P.913 only goes through the initialization steps; second, in the AP solver the re-estimation of quality score ψ_j is weighted by the subject consistency υ_i^{-2} whereas in P.913, the re-estimation is unweighted. Please note that weighting multiple random variables by the inverse of their variance is the minimum error parameter estimation, as can be trivially proven through Lagrange multipliers.

Confidence Interval

The estimate of each model parameter $\{\psi_j\}$, $\{\Delta_i\}$, $\{\upsilon_i\}$ is associated with a confidence interval. Using the Cramer-Rao bound [13], the asymptotic 95% confidence intervals for the mean term ψ_j and Δ_i have the form $CI(a) = a \pm 1.96 \left(-\frac{\partial^2 L(a)}{\partial a^2}\right)^{-\frac{1}{2}}$, where their second-order derivatives $\frac{\partial^2 L(a)}{\partial a^2}$ can be found in the first Appendix. The 95% confidence interval for the standard deviation term υ_i has the form $\left(\sqrt{\frac{k}{\chi_k^2(0.975)}}\upsilon, \sqrt{\frac{k}{\chi_k^2(0.025)}}\upsilon\right)$, where

---

Algorithm 4 Proposed Alternating Projection (AP) solver
- Input:
  - u_{ijr} for subject $i = 1, ..., I$, stimulus $j = 1, ..., J$ and repetition $r = 1, ..., R$.
  - Stop threshold ψ^{thr}.
- Initialize $\{\psi_j\} \leftarrow \{MOS_j\}$, $\{\Delta_i\} \leftarrow \{BIAS_i\}$.
- Loop:
  - $\{\psi_j^{prev}\} \leftarrow \{\psi_j\}$.
  - $\varepsilon_{ijr} = u_{ijr} - \psi_j - \Delta_i$ for $i = 1, ..., I$, $j = 1, ..., J$ and $r = 1, ..., R$.
  - $\upsilon_i \leftarrow \sigma_i(\{\varepsilon_{ijr}\})$ for $i = 1, ..., I$.
  - $\psi_j \leftarrow \frac{\sum_{ir} \upsilon_i^{-2}(u_{ijr} - \Delta_i)}{\sum_{ir} \upsilon_i^{-2}}$ for $j = 1, ..., J$.
  - $\Delta_i \leftarrow \frac{\sum_{jr}(u_{ijr} - \psi_j)}{\sum_{jr} 1}$, for $i = 1, ..., I$.
  - If $\left(\sum_{j=1}^{J}(\psi_j - \psi_j^{prev})^2\right)^{\frac{1}{2}} < \psi^{thr}$, break.
- Output: $\{\psi_j\}$, $\{\Delta_i\}$, $\{\upsilon_i\}$.

$\chi_k^2(a)$ is the percent point function (the inverse of CDF) of a chi-square distribution with k degrees of freedom. After simplification, the confidence intervals for ψ_j, Δ_i and υ_i are:

$$CI(\psi_j) = \psi_j \pm 1.96 \frac{1}{\sqrt{\sum_{ir} \upsilon_i^{-2}}}, \tag{6}$$

$$CI(\Delta_i) = \Delta_i \pm 1.96 \frac{\upsilon_i}{\sqrt{\sum_{jr} 1}},$$

$$CI(\upsilon_i) = \left(\sqrt{\frac{k_i}{\chi_{k_i}^2(0.975)}}\upsilon_i, \sqrt{\frac{k_i}{\chi_{k_i}^2(0.025)}}\upsilon_i\right),$$

where $k_i = \sum_{jr} 1$ is the number of samples that subject i has viewed.

There is one fact worth mentioning. Recall that $\sum_{ir}$ is equivalent to $\sum_{ir:u_{ijr} \neq \star}$, where $\star$ represents missing observation. If there is no missing observation, that is, the subjective test has complete data, then the lengths of the confidence intervals for ψ_j, $j = 1, ..., J$ are the same, equal to $\frac{3.92}{\sqrt{\sum_{ir} \upsilon_i^{-2}}}$ (since it is independent of the subscript j). This is very different from the confidence intervals estimated from a conventional approach (for example, plain MOS, or BT.500), where each stimulus has a different confidence interval length (see the second Appendix for a MLE interpretation of the plain MOS). This phenomenon can be explained by the fact that all the true quality parameters ψ_j, $j = 1, ..., J$ are estimated jointly, yielding identical certainty for all the estimated parameters.

Experimental Results

We compare the proposed method (the proposed model and its two numerical solvers) with the prior art BT.500 and P.913 recommendations. For P.913, after subject bias removal, we assume that a BT.500-style subject rejection is carried out, before calculating the MOS and the corresponding confidence intervals. We first illustrate the proposed model by giving visual examples on two datasets: VQEG HD3 dataset [14] (which is the compression-only subset of the larger HDTV Ph1 Exp3 dataset) and the NFLX
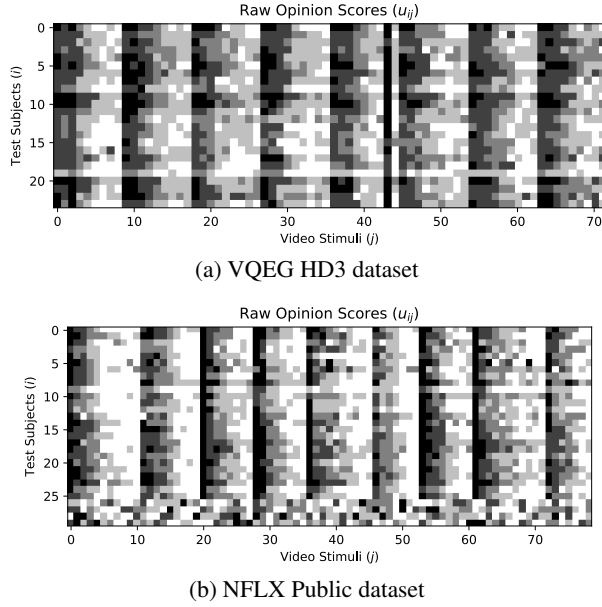
(a) VQEG HD3 dataset

(b) NFLX Public dataset

Figure 1. *Raw opinion scores from (a) the VQEG HD3 dataset and (b) the NFLX Public dataset. Each pixel represents a raw opinion score. The darker the color, the lower the score. The impaired videos are arranged by contents, and within each content, from low quality to high quality (with the reference video always appears last). For the NFLX Public dataset, the last four rows correspond to corrupted subjective data.*

Public dataset [15]. We then validate the model-data fit using the Bayesian Information Criterion (BIC) on 22 datasets, including 20 datasets as part of a different larger experiments: VQEG HDTV Phase I [14]; ITS4S [16]; AGH/NTIA [10, 17]; MM2 [18]; ITU-T Supp23 Exp1 [19]; and ITS4S2 [20]. We also evaluate the confidence intervals on the estimated quality scores on these 22 datasets. Next, we demonstrate that the proposed model is much more effective in dealing with outlier subjects. We then use synthetic data to validate the accuracy of the numerical solvers and the confidence interval calculation. Lastly, we compare the runtime of the various schemes.

Visual Examples

First, we demonstrate the proposed method on the VQEG HD3 and the NFLX Public datasets. Refer to Figure 1 for a visualization of the raw opinion scores. The 44th video of the VQEG HD3 dataset has a quality issue that all its scores are low. The NFLX Public dataset includes four subjects whose raw scores were shuffled due to a software issue during data collection.

Figure 2 shows the recovered quality scores of the four methods compared. The quality scores recovered by the two proposed methods are numerically different from the ones from BT.500 and P.913, suggesting that the recovery is non-trivial. The average confidence intervals by the proposed methods are generally tighter, compared to the ones from BT.500 and P.913, suggesting that the estimation has higher confidence. The NBIC scores, to be discussed in detail in Section Model-Data Fit, represent how well the model fits the data. It can be observed that the proposed model fits the data better than BT.500 and P.913.

Figure 3 shows the recovered subject bias and subject incon-

sistency by the methods compared. On the VQEG HD3 dataset, it can be seen that the 20th subject has the most positive bias, which is evidenced by the whitish horizontal strip visible in Figure 3 (a). On the NFLX Public dataset, the last four subjects, whose raw scores are scrambled, have very high subject inconsistency values. Correspondingly, their estimated biases have very loose confidence intervals. This illustrates that the proposed model is effective in modeling outlier subjects. On the contrary, among the four outlier subjects, both BT.500 and P.910 fail to reject the 28th subject.

The subject bias and inconsistency revealed through the recovery process could be valuable information for subject screening. Unlike BT.500, which makes a binary decision on if a subject is accepted/rejected, the proposed approach characterizes a subject's inaccuracy in two dimensions, along with their confidence intervals, allowing further interpretation and study. How to use the bias and inconsistency information to better screen subjects remains our future work.

Model-Data Fit

Bayesian Information Criterion [6] is a criterion for model-data fit. When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. BIC attempts to balance between the degree of freedom (characterized by the number of free parameters) and the goodness of fit (characterized by the log-likelihood function). Formally, BIC is defined as $BIC = \log(n)|\theta| - 2L(\theta)$, where $n = |\{u_{ijr}\}|$ is the total number of observations (i.e. the number of opinion scores), $|\theta|$ is the number of model parameters, and $L(\theta)$ is the log-likelihood function. One can interpret that the lower the free parameter numbers $|\theta|$, and the higher the log-likelihood $L(\theta)$, the lower the BIC, and hence the better fit. In this work, we adopt the notion of a "normalized BIC", defined as the BIC divided by the number of observations, or:

$$NBIC = \frac{\log(n)|\theta| - 2L(\theta)}{n},$$

as the model fit criterion, for easier comparison across datasets.

Table NBIC shows the NBIC reported on the compared methods on the 22 public datasets. The MOS method is the plain MOS without subject rejection or subject bias removal. $|\theta|$ for MOS and BT.500 is $2J$, where J is the number of stimuli (refer to the second Appendix for a MLE interpretation of the plain MOS). For P.913, $|\theta|$ is equal to $2J + I$, where I is the number of subjects (due to the subject bias term). For the calculation of the log-likelihood function, notice that if subject rejection is applied, only the opinion scores after rejection are taken into account. The result in Table NBIC shows that the proposed two solvers yield better model-data fit than the plain MOS, BT.500 and P.913 approaches.

Confidence Interval of Quality Scores

Table CI shows the average length of the confidence intervals on the compared methods on the 22 public datasets. The smaller the number, the tighter the confidence interval, thus more confident the estimation is. For MOS, BT.500 and P.913, the confidence intervals are calculated based on (8). For BT.500 and P.913, only the opinions scores after rejection are taken into account. For the proposed methods, the confidence intervals are calculated
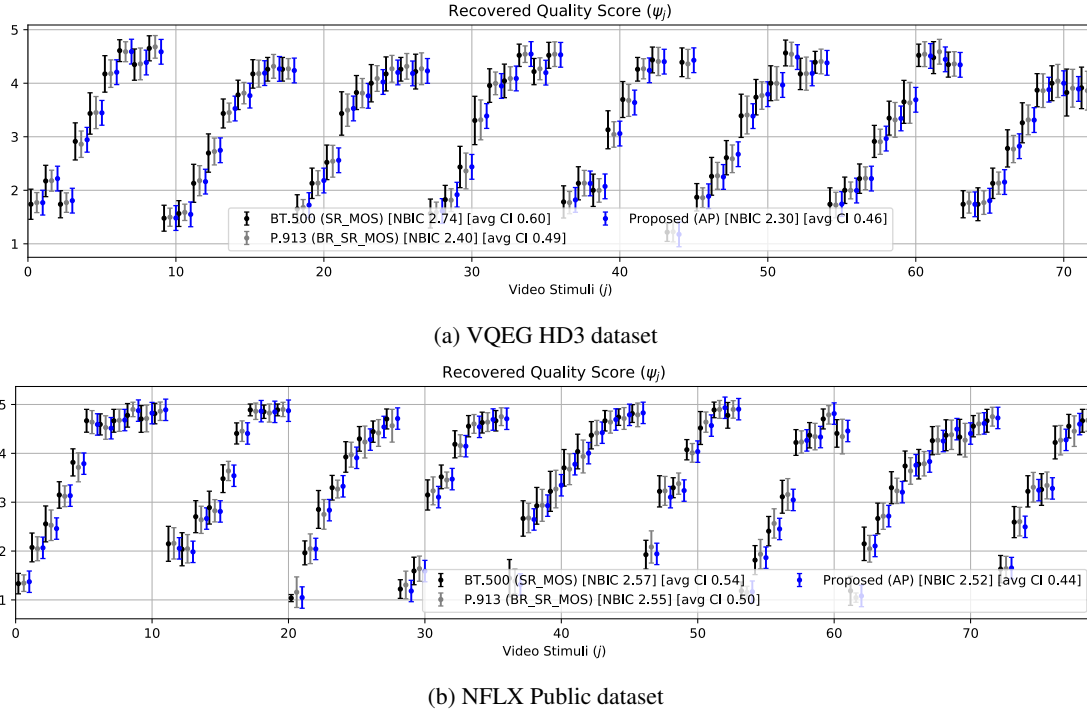
(a) VQEG HD3 dataset

(b) NFLX Public dataset

Figure 2. *Recovered quality score ψ_j and its confidence interval for the four methods compared, on (a) the VQEG HD3 dataset and (b) the NFLX Public dataset. The proposed NR method is not shown in the plots since it virtually produces identical results as the proposed AP method. For each method compared, the NBIC score (see Section Model-Data Fit) and the average length of the confidence interval are reported. (SR: subject rejection; BR: bias removal; avg CI: average confidence interval; NBIC: Bayesian Information Criterion; NR: Newton-Raphson; AP: Alternating Projection.)*

based on (6). It can be observed that the proposed two methods yield tighter confidence intervals compared to the other methods. For some databases BT.500 generates wider confidence interval than the plain MOS. This phenomena can be explained by the fact that subject rejection decreases the number of samples, even though the variance may also be decreased. Overall, the obtained confidence interval can be either narrower or wider.

Robustness against Outlier Subjects

We demonstrate that the proposed method is much more effective in dealing with (corrupted) outlier subjects compared to other methods. We use the following methodology in our reporting of results. For each method compared, we have a benchmark result, which is the recovered quality scores obtained using *that* method - for fairness - on an *unaltered full dataset* (note that for the NFLX Public dataset, unlike the one used in Figure 1, 2 and 3, we start with the dataset where the corruption on the last four subjects has been corrected). We then consider that a number of the subjects are "corrupted" and simulate it by randomly shuffling each corrupted subject's votes among the video stimuli. We then run each method compared on the partially corrupted datasets. The quality scores recovered are normalized by subtracting the mean and dividing by the standard deviation of the scores of the unaltered dataset. The normalized scores are compared against the benchmark, and a root-mean-squared-error (RMSE) value is reported.

Figure 4 reports the results on the two datasets, comparing the proposed method with plain MOS, BT.500, P.913 and the pro-

posed AP solver, as the number of corrupted subjects increases. It can be observed that in the presence of subject corruption, The proposed method achieves a substantial gain over the other methods. The reason is that the proposed model was able to capture the variance of subjects explicitly and is able to compensate for it. On the other hand, the other methods are only able to identify part of the corrupted subjects. Meanwhile, traditional subject rejection employs a set of hard coded parameters to determine outliers, which may not be suited for all conditions. By contrast, the proposed model naturally integrates the various subjective effects together and is solved efficiently by the MLE formulation.

Figure 5 reports the results as we increase the probability of corruption from 0 to 1 while fixing the number of corrupted subjects to 10. It can be seen that as the corruption probability increases, the RMSE increases linearly/near-linearly for other methods, while the RMSE increases much slower for the proposed method, and it saturates at a constant value without increasing further. A simplified explanation is that, since only a subset of a subject's scores is unreliable, discarding all of the subject's scores is a waste of valuable subjective data, while the proposed method can effectively avoid that.

Validation of Solvers and Confidence Interval Calculation

Next, we demonstrate that the NR and AP solvers can accurately recover the parameters of the proposed model. This is shown using synthetic data, where the ground truth of the model parameters are known. In this section, we considered only the
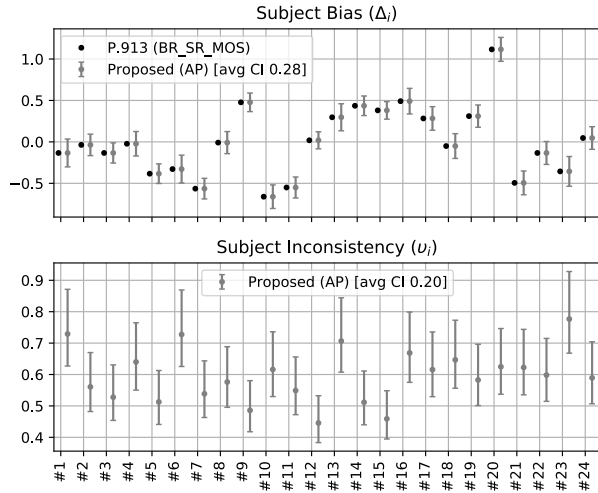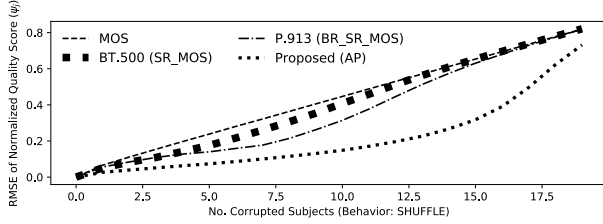
(a) VQEG HD3 dataset
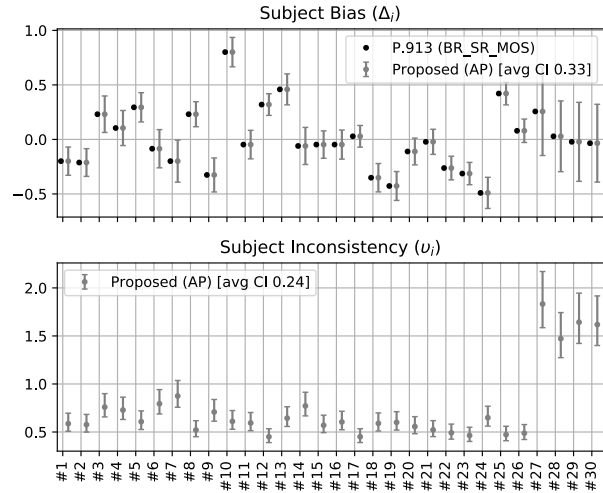
(b) NFLX Public dataset

Figure 3. *Recovered subject bias Δ_i and subject inconsistency υ_i for each subject i, for the methods compared, on (a) the VQEG HD3 dataset and (b) the NFLX Public dataset. The proposed NR method is not shown in the plots since it virtually produces identical results as the proposed AP method. For each method compared, the average length of the confidence interval is reported. (SR: subject rejection; BR: bias removal; avg CI: average confidence interval; NR: Newton-Raphson; AP: Alternating Projection.)*

Table NBIC: Normalized Bayesian Information Criterion (NBIC) reported on the compared methods on public datasets. The NR and AP methods produce identical results. (MOS: plain mean opinion score; NR: Newton-Raphson; AP: Alternating Projection.)

Dataset	MOS	BT.500	P.913	NR/AP
VQEG HD3	2.75	2.74	2.39	**2.30**
NFLX Public	2.97	2.57	2.55	**2.52**
HDTV Ph1 Exp1	2.45	2.46	2.38	**2.20**
HDTV Ph1 Exp2	2.72	2.72	2.52	**2.32**
HDTV Ph1 Exp3	2.72	2.71	2.37	**2.29**
HDTV Ph1 Exp4	2.96	2.96	2.51	**2.27**
HDTV Ph1 Exp5	2.77	2.77	2.47	**2.33**
HDTV Ph1 Exp6	2.51	2.49	2.32	**2.16**
ITU-T Supp23 Exp1	2.91	2.91	2.35	**2.31**
MM2 1	2.80	2.78	2.83	**2.74**
MM2 2	3.89	3.89	3.52	**3.13**
MM2 3	2.48	2.47	2.45	**2.41**
MM2 4	2.74	2.73	2.62	**2.47**
MM2 5	2.90	2.82	2.67	**2.64**
MM2 6	2.81	2.74	2.74	**2.72**
MM2 7	2.73	2.72	2.76	**2.67**
MM2 8	3.00	2.92	2.88	**2.70**
MM2 9	3.27	3.21	2.95	**2.79**
MM2 10	3.04	3.05	2.98	**2.82**
its4s2	3.63	3.63	2.96	**2.59**
its4s AGH	3.15	3.05	2.77	**2.64**
its4s NTIA	2.94	2.91	2.53	**2.38**

Table CI: Average length of confidence intervals of the estimated quality scores reported on the compared methods on public datasets. The NR and AP methods produce identical results. (MOS: plain mean opinion score; NR: Newton-Raphson; AP: Alternating Projection.)

Dataset	MOS	BT.500	P.913	NR/AP
VQEG HD3	0.59	0.60	0.49	**0.46**
NFLX Public	0.62	0.54	0.5	**0.44**
HDTV Ph1 Exp1	0.50	0.61	0.48	**0.46**
HDTV Ph1 Exp2	0.57	0.57	0.53	**0.48**
HDTV Ph1 Exp3	0.56	0.59	0.52	**0.48**
HDTV Ph1 Exp4	0.63	0.63	0.52	**0.47**
HDTV Ph1 Exp5	0.57	0.57	0.53	**0.49**
HDTV Ph1 Exp6	0.50	0.51	0.48	**0.45**
ITU-T Supp23 Exp1	0.61	0.61	0.56	**0.47**
MM2 1	0.59	0.60	0.57	**0.53**
MM2 2	1.21	1.21	1.12	**0.88**
MM2 3	0.47	0.48	0.45	**0.42**
MM2 4	0.58	0.59	0.54	**0.48**
MM2 5	0.63	0.65	0.58	**0.52**
MM2 6	0.62	0.70	0.59	**0.56**
MM2 7	0.60	0.61	0.57	**0.55**
MM2 8	0.76	0.76	0.71	**0.66**
MM2 9	0.84	0.85	0.74	**0.68**
MM2 10	0.77	0.83	0.73	**0.70**
its4s2	0.82	0.82	0.66	**0.60**
its4s AGH	0.68	0.68	0.61	**0.56**
its4s NTIA	0.57	0.58	0.54	**0.48**

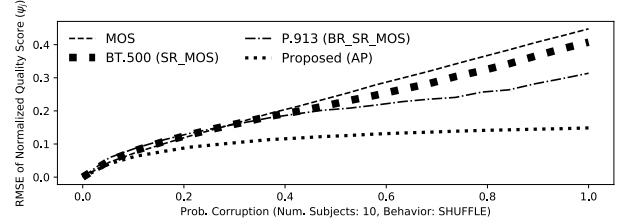(a) VQEG HD3 dataset
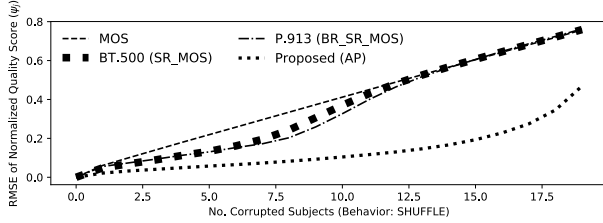
(b) NFLX Public dataset

Figure 4. *RMSE of the (normalized) recovered quality score ψ_j as a function of the number of corrupted subjects, of the proposed method (AP) versus the other methods, of (a) the VQEG HD3 dataset and (b) the NFLX Public dataset. The subject corruption is simulated, in the way that the scores corresponding to a subject are scrambled. The recovered quality score is normalized by subtracting the mean and dividing by the standard deviation of the scores of the unaltered dataset. (MOS: plain mean opinion score; SR: Subject Rejection; BR: Bias Removal; AP: Alternating Projection.)*

(a) VQEG HD3 dataset

(b) NFLX Public dataset

Figure 5. *RMSE of the (normalized) recovered quality score ψ_j as a function of the probability of corruption (fixing the number of corrupted subjects to 10), of the proposed method (AP) versus the other methods, of (a) the VQEG HD3 dataset and (b) the NFLX Public dataset. The subject corruption is simulated, in the way that the scores corresponding to a subject are scrambled. The recovered quality score is normalized by subtracting the mean and dividing by the standard deviation of the scores of the unaltered dataset. (MOS: plain mean opinion score; SR: Subject Rejection; BR: Bias Removal; AP: Alternating Projection.)*

NFLX Public dataset for simulations. The random samples are generated using the following methodology. For each proposed solver, we take the NFLX Public dataset and run the solver to estimate the parameters. The parameters estimated from a real dataset allow us to run simulations with practical settings. We then treat the estimated parameters as the "synthetic" parameters, run simulations to generate synthetic samples according to the model (1). Subsequently, we run the solver again on the synthetic data to yield the "recovered" parameters.
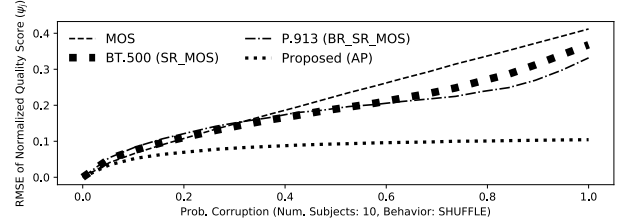
Figure 6 shows the scatter plots of the synthetic vs. recovered parameters, for the true quality ψ_j, subjective bias Δ_i and subject inconsistency υ_i terms. It can be observed that the solvers recover the parameters reasonably well. We have to keep in mind that the synthetic data, differently from usual subjective scores of category rating, are continuous. For discrete data, some specific problems would influence the obtained results as described in [12]. Since those problems are not the main topic of this paper we do not go into more details and leave it as a future topic of research.

Figure 6(a) also shows the recovery result of the BT.500 and P.913. It is noticeable that the recovered subject biases by the AP method and the P.913 subject bias removal are very similar. This should not be surprising, considering that the AP method can be treated as a weighted and iterative generalization of the P.913 method.

Also plotted in Figure 6(b) are the confidence intervals of the recovered parameters. The reported "CI%" is the percentage of occurrences where the synthetic ground truth falls within the confidence interval. By definition, we expect the CI% to be 95% on average. To verify this, we run the same simulation on the 22 public datasets. For each dataset, the simulation is run 100

times with different seeds. The result is shown in Table CI%. We compare the proposed NR and AP methods with the plain MOS. It can be seen that all methods yield CI% to be very close to 95%, but slightly below. The explanation is that both have assumed that the underlying distribution is Gaussian, but with both the mean and standard deviation unknown, one should use a Student's t-distribution instead. If the t-distribution is used, the coefficient can no longer be a fixed value 1.96 but is a function of the number of subjects and repetitions.

For the NR and AP methods, there are occasional cases where the CI% is significantly lower, for example, ψ_j and υ_i for MM2 2 dataset. This is the case where the stimuli and/or subject dimensions are small, yielding non-Gaussian behavior (recall that the confidence interval calculated is asymptotic). In practice, we can introduce a correction term to compensate for the non-Gaussianity.

Runtime and Iterations

Lastly, we evaluate the runtime of the proposed NR and AP methods compared to the others. The experiment was performed on a MacBook Pro (15-inch, 2018) with 2.9 GHz Intel Core i9 with 32 GB 2400 MHz DDR4 memory, macOS version 10.14.6. The schemes compared are implemented in Python, and are open-source on Github [7]. The results of 100 simulations runs (based on the similar methodology as in the previous sections) of each methods are reported in Table Runtime. The results reveal the order of magnitude of the algorithms compared. The plain MOS is typically the fastest, while the BT.500 and P.913 are two magnitude slower. The NR and AP algorithms are three and one magni-
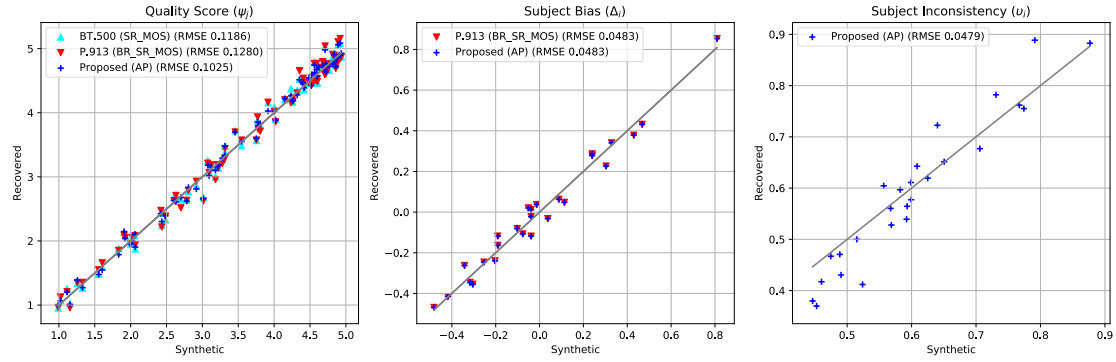
(a) Comapring BT.500, P.913 and AP
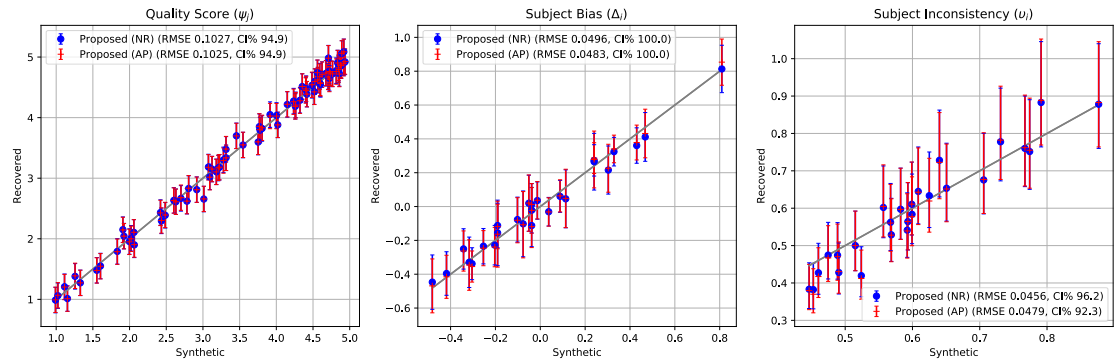
(b) Confidence Intervals of NR and AP

Figure 6. *Validation of the proposed NR and AP solvers using synthetic data. The random samples are generated using the following methodology. For each proposed solver, take the NFLX Public dataset, run the solver to estimate the parameters. Treat the estimated parameters and the "synthetic" parameters, run simulations to generate synthetic samples according to the model (1). Run the solver again on the synthetic data to yield the "recovered" parameters. The x-axis shows the synthetic parameters and the y-axis shows the recovered parameters. (a) Comparing the proposed AP with BT.500 and P.913, (b) Proposed NR and AP with confidence intervals. (NR: Newton-Raphson; AP: Alternating Projection.)*

Table CI%: Average confidence interval coverage (CI%) reported on public datasets. For each proposed solver and each dataset, run the solver to estimate the parameters. Treat the estimated parameters and the "synthetic" parameters, run simulations to generate synthetic samples according to the model (1) (except for MOS, whose samples are generated according to (7)). Run the solver again on the synthetic data to yield the "recovered" parameters and their confidence intervals. The reported "CI%" is the percentage of occurrences when the synthetic ground truth falls within the confidence interval. For each dataset, the simulation is run 100 times with different seeds. Note that for both MOS and the proposed NR and AP methods, the CI% is slightly below 95%, due to the underlying Gaussian assumption used instead of the legitimate Student's t-distrubtion. (MOS: plain mean opinion score; NR: Newton-Raphson; AP: Alternating Projection.)

Dataset	MOS	NR			AP		
	ψ_j	ψ_j	Δ_i	υ_i	ψ_j	Δ_i	υ_i
VQEG HD3	93.3	93.6	93.9	93.0	93.2	94.4	91.9
NFLX Public	94.2	93.7	94.5	93.1	93.5	94.1	92.3
HDTV Ph1 Exp1	93.9	94.1	93.9	93.1	93.8	94.2	91.3
HDTV Ph1 Exp2	93.8	94.0	94.5	92.5	93.8	94.0	91.2
HDTV Ph1 Exp3	93.9	93.9	94.4	92.5	93.7	94.1	90.6
HDTV Ph1 Exp4	93.8	94.0	94.3	91.9	93.8	94.1	90.9
HDTV Ph1 Exp5	93.8	94.1	94.2	92.2	93.9	94.2	90.9
HDTV Ph1 Exp6	93.8	94.0	94.4	92.6	93.9	94.0	91.0
ITU-T Supp23 Exp1	93.8	94.0	94.4	91.2	93.8	94.9	90.0
MM2 1	93.5	92.8	95.4	92.6	92.5	94.0	91.6
MM2 2	92.1	81.5	92.9	80.0	68.1	92.1	75.4
MM2 3	94.4	93.6	95.1	93.4	93.4	94.2	92.0
MM2 4	93.2	93.6	95.6	93.0	93.2	95.1	92.0
MM2 5	93.2	93.2	95.7	92.7	91.8	95.3	91.4
MM2 6	93.6	93.3	95.2	92.8	93.0	94.1	91.4
MM2 7	93.6	93.3	95.2	92.8	92.9	94.2	91.9
MM2 8	93.0	92.4	95.4	88.8	92.2	94.5	87.0
MM2 9	93.2	93.3	94.8	89.1	92.8	94.2	88.1
MM2 10	93.2	93.1	95.7	89.7	92.8	94.5	87.9
its4s2	93.1	94.1	94.6	60.6	94.1	94.2	59.2
its4s AGH	93.6	94.0	94.4	90.4	94.0	94.4	89.7
its4s NTIA	93.9	94.4	94.7	86.1	94.3	95.1	85.6

tude slower, respectively. Notably, the AP runs faster than BT.500 and P.913, and is about 50x faster than the NR. The AP also requires about half the number of iterations to reach convergence than the NR.

Conclusions

In the paper, we proposed a simple model to account for two of the most dominant effects of test subject inaccuracy: subject bias and subject inconsistency. We further proposed to solve the model parameters through maximum likelihood estimation and presented two numerical solvers. We compared the proposed methodology with the standardized recommendations including ITU-R BT.500 and ITU-T P.913, and showed that the proposed methods have advantages in: 1) better model-data fit, 2) tighter confidence intervals, 3) better robustness against subject outliers, 4) shorter runtime, 5) absence of hard coded parameters and thresholds, and 6) auxiliary information on test subjects. We believe the proposed methodology is generally suitable for subjective evaluation of perceptual audiovisual quality in multimedia and television services, and we propose to update the corresponding recommendations with the methods presented.

References

[1] ITU-R BT.500-14 (10/2019): Methodologies for the Subjective Assessment of the Quality of Television Images. `https://www.itu.int/rec/R-REC-BT.500`.

[2] ITU-T P.910 (04/08): Subjective Video Quality Assessment Methods for Multimedia Applications. `https://www.itu.int/rec/T-REC-P.910`.

[3] ITU-T P.913 (03/2016): Methods for the Subjective Assessment of Video Quality, Audio Quality and Audiovisual Quality of Internet Video and Distribution Quality Television in Any Environment. `https://www.itu.int/rec/T-REC-P.913`.

[4] Maximum Likelihood Estimation (Wikipedia). `https://en.wikipedia.org/wiki/Maximum_likelihood_estimation`. [Online; accessed 30-March-2020].

[5] Newton's Method in Optimization (Wikipedia). `https://en.wikipedia.org/wiki/Newton%27s_method_in_optimization`. [Online; accessed 30-March-2020].

[6] Bayesian Information Criterion (Wikipedia). `https://en.wikipedia.org/wiki/Bayesian_information_criterion`. [Online; accessed 30-March-2020].

[7] SUREAL - Subjective Recovery Analysis. `https://github.com/Netflix/sureal`. [Online; accessed 30-March-2020].

[8] Tobias Hoßfeld, Raimund Schatz, and Sebastian Egger. SOS: The MOS is not enough! *2011 3rd International Workshop on Quality of Multimedia Experience, QoMEX 2011*, pages 131–136, 2011.

[9] Zhi Li and Christos G. Bampis. Recover Subjective Quality Scores from Noisy Measurements. `https://arxiv.org/abs/1611.01715`. in Proceedings of the Data Compression Conference, 2017.

[10] L. Janowski and M. Pinson. Subject bias: Introducing a theoretical user model. In *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 251–256, Sep. 2014.

[11] L. Janowski and M. Pinson. The accuracy of subjects in a quality experiment: A theoretical subject model. *IEEE Transactions on Multimedia*, 17(12):2210–2224, Dec 2015.

[12] Lucjan Janowski, Bogdan Ćmiel, Krzysztof Rusek, Jakub Nawała, and Zhi Li. Generalized score distribution. `https://arxiv.org/abs/1909.04369`. in arXiv:1909.04369 [stat.ME], 2019.

[13] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. A Wiley-Interscience publication. Wiley, 2006.

[14] Margaret Pinson, Filippo Speranza, M Barkowski, V Baroncini, R Bitto, S Borer, Y Dhondt, R Green, L Janowski, T Kawano, et al. Report on the validation of video quality models for high definition video content. *Video Quality Experts Group*, 2010.

[15] Netflix Public Dataset. `https://github.com/Netflix/vmaf/blob/master/resource/doc/datasets.md`. [Online; accessed 30-March-2020].

[16] Margaret H. Pinson. Its4s: A video quality dataset with four-second unrepeated scenes. Technical Report NTIA Technical Memo TM-18-532, NTIA/ITS, Feb. 2018.

[17] Margaret H. Pinson and L. Janowski. Agh/ntia: A video quality subjective test with repeated sequences. Technical Report NTIA Technical Memo TM-14-505, NTIA/ITS, June 2014.

[18] M. H. Pinson, L. Janowski, R. Pepion, Q. Huynh-Thu, C. Schmidmer, P. Corriveau, A. Younkin, P. Le Callet, M. Barkowsky, and W. Ingram. The influence of subjects and environment on audiovisual subjective tests: An international study. *IEEE Journal of Selected Topics in Signal Processing*, 6(6):640–651, Oct 2012.

[19] ITU-T P-Series. ITU-T coded-speech database. Technical Report Series P: Telephone Transmission Quality, Telephone Installations, Local Line Networks, ITU-T, Feb. 1998.

[20] Margaret H. Pinson. ITS4S2: An image quality dataset with unrepeated images from consumer cameras,. Technical Report NTIA Technical Memo TM-19-537, NTIA/ITS, Apr. 2019.

Appendix: First- and Second-Order Partial Derivatives of $L(\theta)$

We can derive the first-order and second-order partial derivatives of $L(\theta)$ with respect to ψ_j, Δ_i and υ_i as:

$$\frac{\partial L(\theta)}{\partial \psi_j} = \sum_{ir} \frac{u_{ijr} - \psi_j - \Delta_i}{\upsilon_i^2}$$

$$\frac{\partial L(\theta)}{\partial \Delta_i} = \sum_{jr} \frac{u_{ijr} - \psi_j - \Delta_i}{\upsilon_i^2}$$

$$\frac{\partial L(\theta)}{\partial \upsilon_i} = \sum_{jr} -\frac{1}{\upsilon_i} + \frac{(u_{ijr} - \psi_j - \Delta_i)^2}{\upsilon_i^3}$$

$$\frac{\partial^2 L(\theta)}{\partial \psi_j^2} = -\sum_{ir} \frac{1}{\upsilon_i^2}$$

$$\frac{\partial^2 L(\theta)}{\partial \Delta_i^2} = -\frac{1}{\upsilon_i^2} \sum_{jr} 1$$

$$\frac{\partial^2 L(\theta)}{\partial \upsilon_i^2} = \sum_{jr} \frac{1}{\upsilon_i^2} - \frac{3(u_{ijr} - \psi_j - \Delta_i)^2}{\upsilon_i^4}$$

Appendix: An MLE Interpretation of the Plain MOS

The plain MOS and its confidence interval can be interpreted using the notion of maximum likelihood estimation. Consider the model:

$$U_{ijr} = \psi_j + \upsilon_j X, \tag{7}$$

where U_{ijr} is the opinion score, ψ_j is the true quality of stimulus j and υ_j is the "ambiguity" of j. $X \sim N(0,1)$ is i.i.d. Gaussian. Note that this is different from the proposed model (1) where υ_i is associated with the subjects, not the stimuli. We can define the log-likelihood function for this model as $L(\theta) = \log P(\{u_{ijr}\} | (\{\psi_j\}, \{\upsilon_j\}))$, and solve for $\{\psi_j\}$ and

Table Runtime: Average runtime in seconds and number of iterations (for NR and AP) reported on public datasets. For each proposed solver and each dataset, run the solver to estimate the parameters. Treat the estimated parameters and the "synthetic" parameters, run simulations to generate synthetic samples according to the model (1) (except for MOS, whose samples are generated according to (7)). Run the solver again on the synthetic data. For each dataset, the simulation is run 100 times with different seeds, and the mean is reported. For NR and AP, also reported are the number of iterations. (MOS: plain mean opinion score; NR: Newton-Raphson; AP: Alternating Projection.)

Dataset	Mean Runtime (seconds)					No. Iterations	
	MOS	BT.500	P.913	NR	AP	NR	AP
VQEG HD3	5.2e-4	1.5e-2	1.5e-2	2.1e-1	4.3e-3	26.2	12.1
NFLX Public	5.7e-4	1.8e-2	1.9e-2	2.8e-1	4.5e-3	34.5	11.8
HDTV Ph1 Exp1	7.7e-4	3.3e-2	3.4e-2	2.0e-1	4.6e-3	23.4	10.3
HDTV Ph1 Exp2	7.8e-4	3.3e-2	3.4e-2	2.8e-1	4.9e-3	33.2	11.3
HDTV Ph1 Exp3	7.8e-4	3.3e-2	3.4e-2	2.5e-1	4.7e-3	29.4	10.7
HDTV Ph1 Exp4	7.6e-4	3.3e-2	3.4e-2	3.3e-1	5.0e-3	38.3	11.5
HDTV Ph1 Exp5	7.8e-4	3.3e-2	3.4e-2	2.7e-1	4.7e-3	31.3	10.8
HDTV Ph1 Exp6	7.6e-4	3.3e-2	3.4e-2	2.2e-1	4.6e-3	25.8	10.7
ITU-T Supp23 Exp1	8.1e-4	3.5e-2	3.5e-2	3.4e-1	5.0e-3	36.0	11.6
MM2 1	4.9e-4	1.3e-2	1.3e-2	2.1e-1	4.3e-3	27.4	12.4
MM2 2	4.0e-4	1.0e-2	1.1e-2	5.8e-1	1.4e-2	78.0	54.9
MM2 3	5.3e-4	1.3e-2	1.4e-2	1.8e-1	4.2e-3	23.3	11.6
MM2 4	5.0e-4	1.3e-2	1.4e-2	2.6e-1	4.6e-3	33.4	13.8
MM2 5	5.0e-4	1.3e-2	1.4e-2	2.9e-1	6.0e-3	37.3	19.3
MM2 6	4.8e-4	1.2e-2	1.3e-2	2.2e-1	4.3e-3	28.8	13.1
MM2 7	4.8e-4	1.2e-2	1.3e-2	2.0e-1	4.2e-3	25.6	12.3
MM2 8	4.3e-4	1.1e-2	1.1e-2	2.7e-1	5.5e-3	35.3	18.7
MM2 9	4.3e-4	1.1e-2	1.2e-2	2.8e-1	5.1e-3	36.5	16.8
MM2 10	4.3e-4	1.1e-2	1.2e-2	2.3e-1	4.8e-3	29.8	15.4
its4s2	3.3e-3	2.5e-1	2.5e-1	1.1e+0	1.3e-2	49.8	13.3
its4s AGH	8.7e-4	4.1e-2	4.2e-2	3.5e-1	5.3e-3	39.4	11.6
its4s NTIA	2.6e-3	1.6e-1	1.6e-1	6.4e-1	1.1e-2	46.2	11.3

$\{v_j\}$ that maximize the log-likelihood function, as follows:

$$\psi_j \quad = \quad \frac{\sum_{ir} u_{ijr}}{\sum_{ir} 1},$$

$$v_j \quad = \quad \sqrt{\frac{\sum_{ir} (u_{ijr} - \psi_j)^2}{\sum_{ir} 1}}.$$

The second-order partial derivative w.r.t. to ψ_j is $\frac{\partial^2 L(\theta)}{\partial \psi_j^2} = -\frac{1}{v_j^2} \sum_{ir} 1$. The 95% confidence interval of ψ_j is then:

$$CI(\psi_j) = \psi_j \pm 1.96 \frac{v_j}{\sqrt{\sum_{ir} 1}}. \tag{8}$$

One minor difference between (8) and the 95% confidence interval formula in BT.500-14 Section A1-2.2.1 is that the former uses differential degrees of freedom 0 and the latter uses 1 for the sample standard deviation calculation. In fact, neither one is fully precise. In the most precise way to calculate the confidence interval, one should use a Student's t-distrubiton with a differential degrees of freedom 1 (see Section Validation of Solvers and Confidence Interval 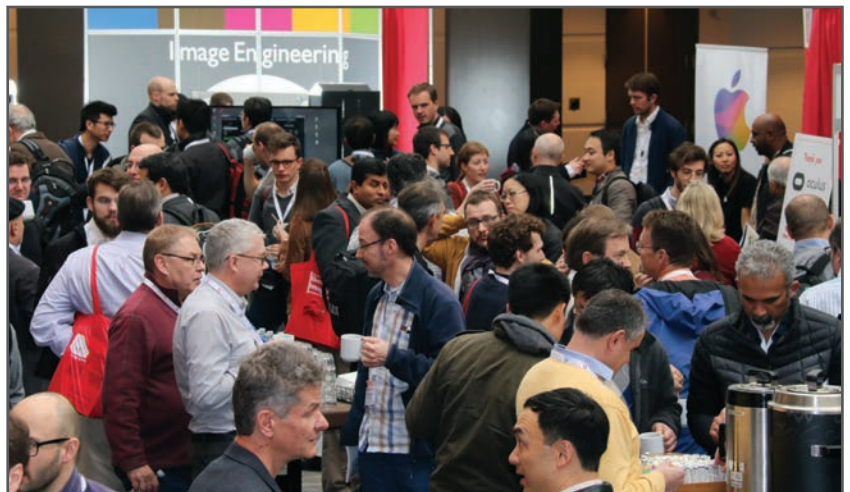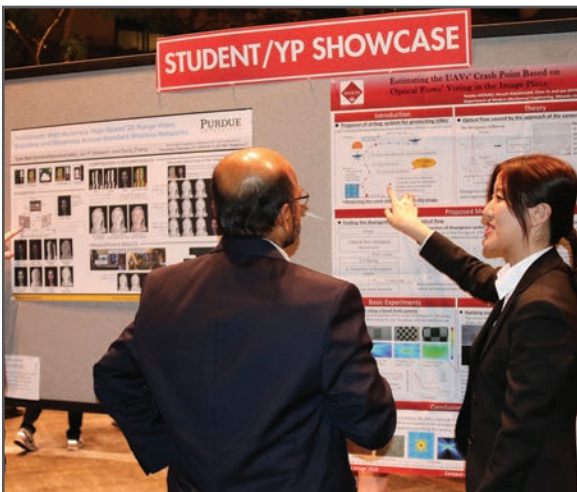Calculation and Table CI% for more discussions).