Subjective Media Quality Recovery from Noisy Raw Opinion Scores: A Non-Parametric Perspective

Andrés Altieri, Lohic Fotio Tiotsop, Giuseppe Valenzise

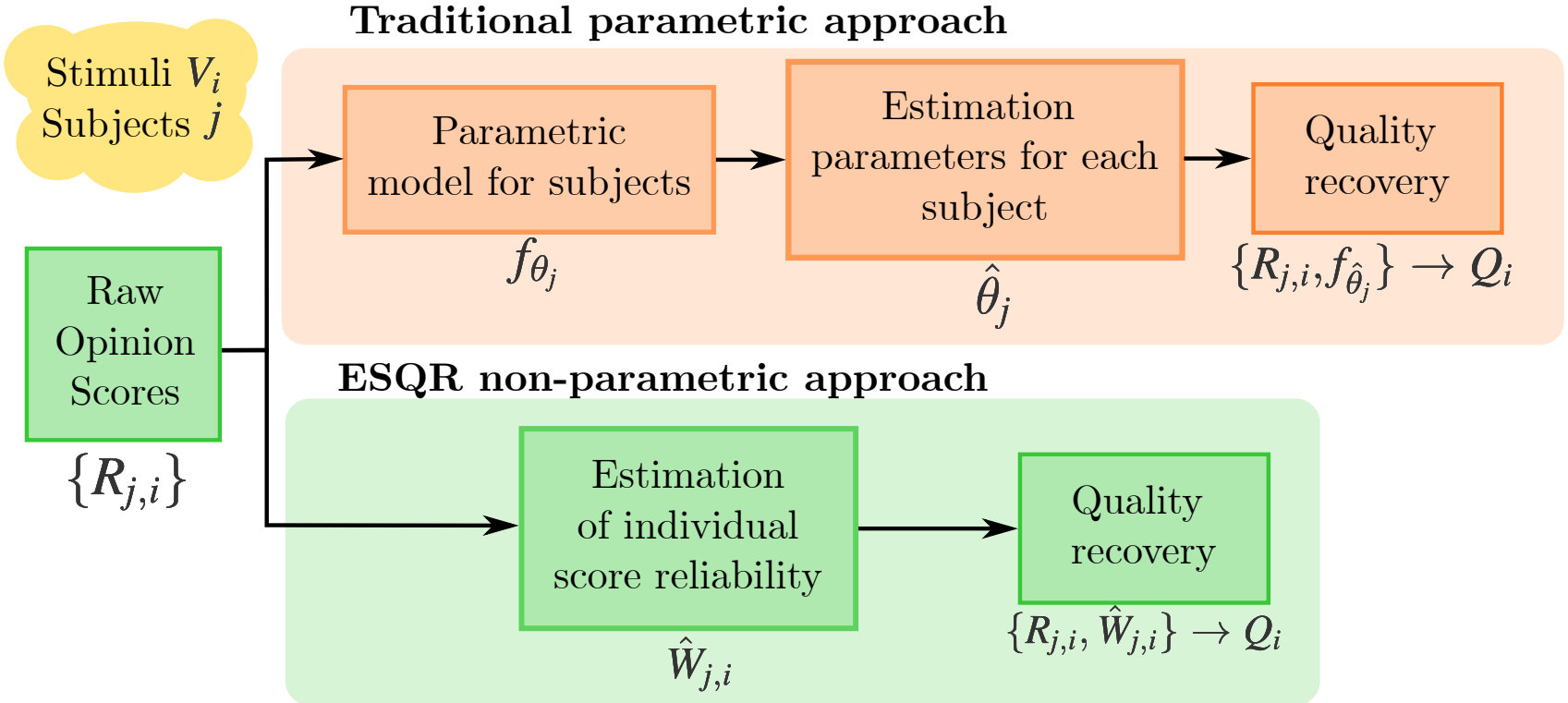
VQEG JEG-Hybrid

Dec 18, 2023

Outline

1. A Non-parametric approach to measure the reliability of a rating given by a subject
2. The Entropy-based Subjective Quality Recovery (ESQR) algorithm
3. Computational results
4. Conclusions

Parametric vs Non-Parametric Approach



- Parametric approach: assume a scoring model f_θ, estimate parameters θ, and determine stimuli quality
- Our non-parametric approach: assess reliability of each rating, prioritizing reliable opinions to determine quality

Parametric vs Non-Parametric Approach

Parametric approaches

- Try to explain the subject scoring behavior
- Make potentially restrictive assumption for stability
- Suffer under/over-fitting issues
- The parameter estimation process is usually computationally demanding

Our Non-parametric approach

- Greater robustness as no assumption is made
- No risk of under/over-fitting the data
- Efficiency, there is no optimization problem to solve
- Do not explain the subject scoring process

Notation

- $\mathcal{I}$, a set of rated stimuli
- $\mathcal{J}$, a set of subjects
- $\mathcal{I}_j \subset \mathcal{I}$ the subset of the stimuli rated by the subject $j \in \mathcal{J}$
- $\mathcal{J}_i \subset \mathcal{J}$, the subset of subjects that rated the stimulus $i \in \mathcal{I}$ using a discrete scale in the range $\{1, ..., K\}$
- V_i the discrete random variable that describes noiseless opinion scores, on the quality of the stimulus $i \in \mathcal{I}$ in the range $\{1, ..., K\}$
- p_{V_i}, the probability mass function of V_i
- $R_{j,i}$, the discrete random variable modeling the score of the subject j for the stimulus i on a quality scale in the range $\{1, .., K\}$.
- $p_{R_{j,i}}$ the probability mass function of $R_{j,i}$.

Measure of Reliability

Definition

The reliability $W_{j,i}$ of the rating $R_{j,i}$ of subject j on the quality of stimulus i is the following ratio:

$$W_{j,i} = -\frac{1}{\log\left(p_{V_i}(R_{j,i})\right)}. \tag{1}$$

- Note that an estimate of p_{V_i} is needed to compute $W_{j,i}$
- A non-parametric estimation of p_{V_i} will be discussed later
- Let's first motivate why $W_{j,i}$ measures how reliable is $R_{j,i}$

Motivation

Let us denote by

- $H(p_{R_{j,i}})$ the entropy of the distribution $p_{R_{j,i}}$
- $D_{\mathrm{KL}}(p_{R_{j,i}}||p_{V_i})$ denotes the Kullback–Leibler (KL) divergence between $p_{R_{j,i}}$ and p_{V_i}

The average inability of subject j to provide repeated ratings on the quality of stimulus i can be measured as:

$$A_J = \frac{1}{|\mathcal{I}_j|} \sum_{i \in \mathcal{I}_j} H(p_{R_{j,i}})$$

The average inability of subject j to rate stimulus i according to p_{V_i} can be measured as:

$$B_J = \frac{1}{|\mathcal{I}_j|} \sum_{i \in \mathcal{I}_j} D_{\mathrm{KL}}(p_{R_{j,i}}||p_{V_i}).$$

Clearly, $A_J + B_J$ measures the overall unreliability of subject j

Motivation

- Let us introduce the following statistic:

$$S_j(\mathcal{I}_j) = \frac{1}{|\mathcal{I}_j|} \sum_{i \in \mathcal{I}_j} W_{j,i}^{-1}. \tag{2}$$

- The following proposition links $W_{j,i}^{-1}$ to the overall subject unreliability

proposition

For each subject j, if there is a constant c such that $\text{var}\left[W_{j,i}^{-1}\right] < c \ \forall i \in \mathcal{I}_j$, *then, as* $|\mathcal{I}_j| \to \infty$,

$$S_j(\mathcal{I}_j) \to (A_J + B_J) \tag{3}$$

Motivation

- The average $\frac{1}{|\mathcal{I}_j|} \sum_{i \in \mathcal{I}_j} W_{j,i}^{-1}$ converges to the overall unreliability, but what about each single term?
- $W_{j,i}$ finds its theoretical explanation in information theory
- $\log\left(p_{V_i}(R_{j,i})\right)$ is the "self-information" contained in the event of choosing $R_{j,i}$ as opinion score for the quality of stimulus i
- The less information the event brings, the more likely it was
- Hence $W_{j,i}$ measures how unlikely and thus potentially noisy is the opinion scores $R_{j,i}$ if the distribution of accurate ratings is p_{V_i}

P_{V_i} Estimation

- P_{V_i} is not directly observable
- A set of noisy ratings that provides an estimate of each $P_{R_{j,i}}$ is observed
- We performed a **non-parametric** estimation of P_{V_i}

$$\hat{p}_{V_i} = \sum_{j \in \mathcal{J}_i} \epsilon_{ij} \hat{p}_{R_{j,i}} \quad i \in \mathcal{I}, \tag{4}$$

where

$$\epsilon_{ij} = \frac{|\hat{C}_j|}{\sum_{k \in \mathcal{J}_i} |\hat{C}_k|} \quad i \in \mathcal{I}, \quad j \in 1, 2, \ldots, |\mathcal{J}_i|. \tag{5}$$

and $\hat{C}_j$ is the overall Spearman Rank Order Correlation Coefficient (SROCC) between the ratings of subject j and those of all the other subjects

Subjective Quality Recovery

- If the exact P_{V_i} could be known, then the quality q_i of stimulus i would have been: $q_i = f(P_{V_i}) = \mathbb{E}_{P_{V_i}}$
- But we only have a "not-sophisticated" estimate $\hat{p}_{V_i}$ of P_{V_i}
- From our point of view, $\mathbb{E}_{\hat{p}_{V_i}}$ is not a very robust estimator of q_i
- We argue that a suitable estimator Q_i of the quality q_i can be obtained by taking into consideration also the reliability of each rating
- In particular, we define:

$$Q_i = g(\hat{p}_{V_i})$$

where $g()$ depends on the introduced measure of reliability and will be defined on the next slides

Entropy-based Subjective Quality Recovery (ESQR)

- Our approach is said to be "entropy-based" because $W_{j,i}$ is linked to the subject reliability trough the concept of entropy
- Our idea: **giving more importance to reliable ratings in the estimator of the ground-truth quality estimator**
- An estimate $\hat{W}_{j,i}$ of $W_{j,i}$ can be computed using $\hat{p}_{V_i}$

Definition

The weight/importance of the rating $R_{j,i}$ of subject j for stimulus i is defined as:

$$\omega_{ij}(\hat{p}_{V_i}) = \frac{\hat{W}_{j,i}}{\sum_{k \in \mathcal{J}_i} \hat{W}_{k,i}}. \tag{6}$$

Entropy-based Subjective Quality Recovery (ESQR)

Definition

The ESQR estimator of the quality of the stimulus i is:

$$Q_i = g(\hat{p}_{V_i}) = \sum_{j \in \mathcal{J}_i} \omega_{ij}(\hat{p}_{V_i}) R_{j,i}$$

ESQR Algorithm

Algorithm 1: Entropy based Subjective Quality Recovery (ESQR)

Data: $R_{j,i},\ i \in \mathcal{I}_j;\ j \in \mathcal{J}$ // stimuli i, subjects j

1 $C_{jk} \leftarrow \text{SROCC}(R_{j,.}, R_{k,.}) \quad j,k \in \mathcal{J}$ // pairwise subject scores correlation

2 $\hat{C}_j \leftarrow \text{FZT}^{-1}\left(\frac{\sum_{k\in\mathcal{J}} \text{FZT}(C_{jk})}{|\mathcal{J}|}\right) \quad j \in \mathcal{J}$ // overall subject-to-subject correlation

3 $\epsilon_{ij} \leftarrow \frac{|\hat{C}_j|}{\sum_{k\in\mathcal{J}_i} |\hat{C}_k|} \quad i \in \mathcal{I};\ \ j \in \mathcal{J}_i$ // importance of the ratings of subject j in the P_{V_i} estimation

4 $\hat{p}_{V_i} \leftarrow \sum_{j\in\mathcal{J}_i} \epsilon_{ij} p_{R_{j,i}} \quad i \in \mathcal{I}$ // estimate the distribution P_{V_i}

5 $\hat{W}_{j,i} \leftarrow \frac{1}{-\log(\hat{p}_{V_i}(R_{j,i}))} \quad i \in \mathcal{I}_j;\ \ j \in \mathcal{J}$ // estimate each opinion score reliability

6 $Q_i = \frac{\sum_{j\in\mathcal{J}_i} \hat{W}_{j,i} R_{j,i}}{\sum_{k\in\mathcal{J}_i} \hat{W}_{k,i}} \quad i \in \mathcal{I}$ // estimate the quality

Result: $Q_i, \quad i \in \mathcal{I}$

FZT stands for Fisher Z-Transform

Confidence Interval of the Recovered Quality

- We observed through simulation that the distribution of Q_i is well approximated by a Gaussian one if $|\mathcal{J}_i| >= 20$
- Note that this does not mean that we are making assumption on the subject scoring behavior
- An unbiased estimate of the standard deviation of the estimator Q_i is:

$$\sigma_{Q_i} = \sqrt{\frac{|\mathcal{J}_i|}{|\mathcal{J}_i| - 1} \sum_{j \in \mathcal{J}_i} \omega_{ij} (R_{j,i} - Q_i)^2}. \tag{7}$$

- The 95% CI of the recovered quality of stimulus i is:

$$\mathrm{CI}_{Q_i} = Q_i \pm 1.96 \frac{\sigma_{Q_i}}{\sqrt{|\mathcal{J}_i|}}. \tag{8}$$

Results: Uncertainty on the Recovered Quality

Table: **Uncertainty of quality estimates**: Comparison of the size of CIs estimated by the different quality recovery approaches. Percentages indicate relative size of the CIs with respect to MOS CIs.

Methods	AVG CI SIZE			
	NETF PUB	**VQ-HD1**	**VQ-HD3**	**VQ-HD5**
MOS	0.509 (——)	0.493 (——)	0.565 (——)	0.575 (——)
BT500	0.515 (+1.18%)	0.613 (+24.34%)	0.586 (+3.72%)	0.575 (+0.00%)
ZREC	0.417 (-18.07%)	0.437 (-11.36%)	0.458 (-18.94%)	0.475 (-17.39%)
SUREAL	0.445 (-12.57%)	0.459 (-6.90%)	0.481 (-14.87%)	0.489 (-14.96%)
RMLE	0.453 (-11.00%)	0.417 (-15.42%)	0.472 (-16.46%)	0.483 (-16.00%)
ESQR	**0.355 (-30.26%)**	**0.361 (-26.77%)**	**0.436 (-22.83%)**	**0.439 (-23.65%)**

Results: Accuracy in Predicting Uncertainty

- We simulated ratings in a way that the ground-truth CI ($gtCI$) of each stimulus is known
- We then used each method m to estimate the ($gtCI$), yielding $\hat{CI}_m$
- Two indexes to measure CI prediction accuracy
  - Δ^m the average distance between the centers of $gtCI$ and $\hat{CI}_m$
  - ρ^m the average ratio between the sizes of $gtCI$ and $\hat{CI}_m$
- Clearly one wants Δ^m close to 0 and ρ^m close to 1

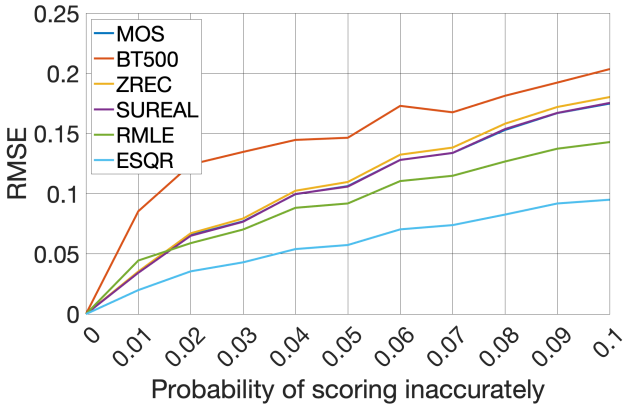
Table: CI prediction accuracy

Method	MOS	BT500	ZREC	SUREAL	RMLE	ESQR
Δ^m	0.127	0.062	0.058	0.051	0.087	0.056
ρ^m	1.470	1.263	1.242	1.242	1.256	0.979

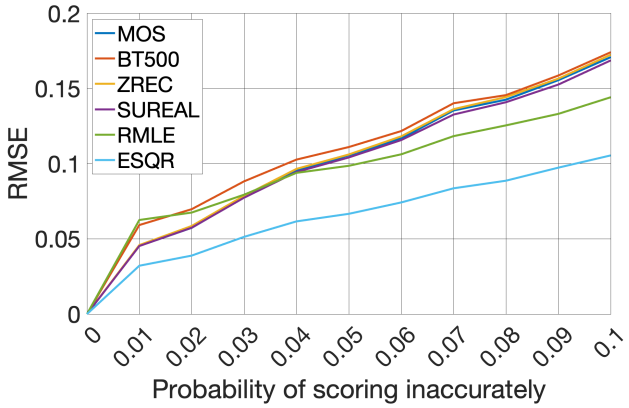
- ZREC, SUREAL and ESQR better predict the CI center
- ESQR better predicts the CI size

Results: Robustness to Noise

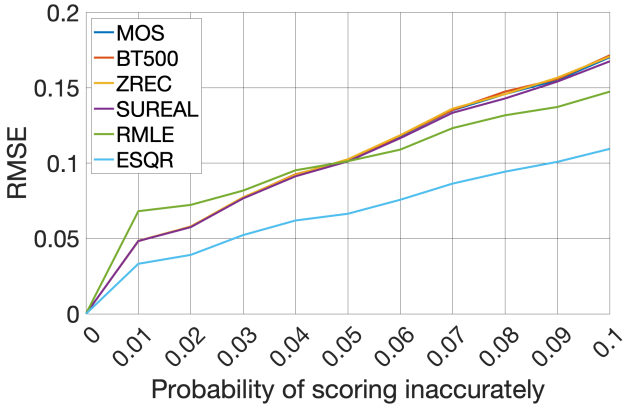
- All subjects have a small probability to score inaccurately
- Simulating noise caused for instance by fatigue or distraction



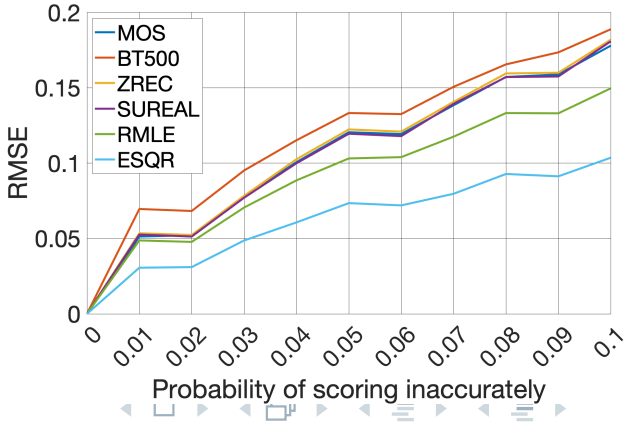
(a) VQEG-HD1



(b) VQEG-HD3



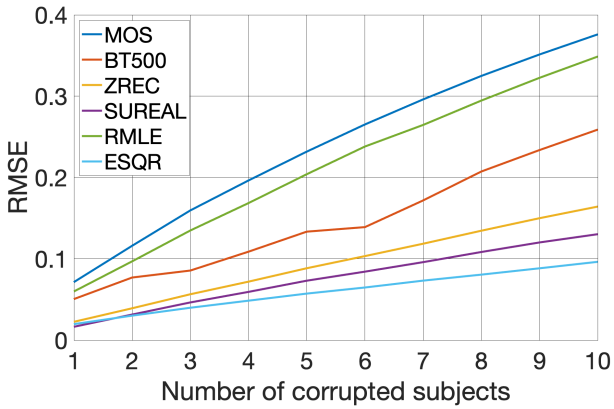
(c) VQEG-HD5



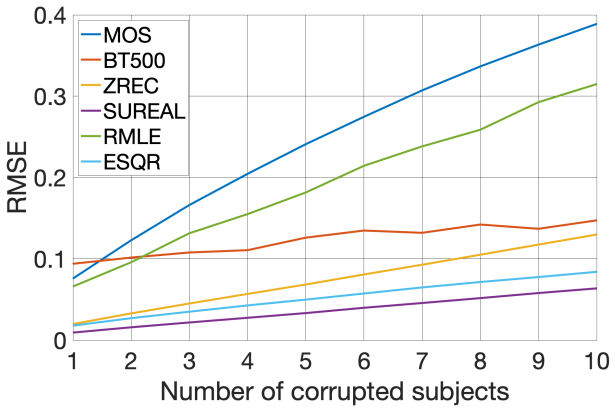
(d) Netflix Public dataset

Results: Robustness to Noise

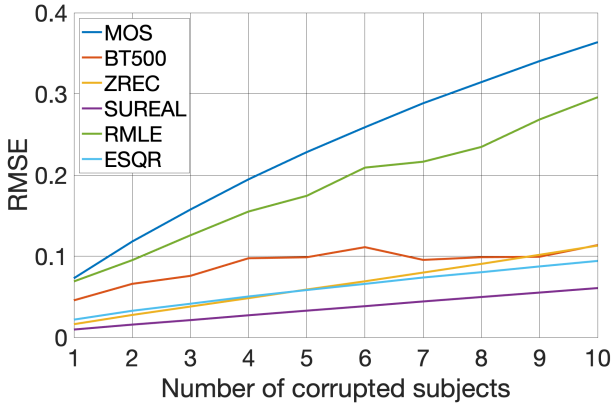
- Adding spammer annotators to the dataset



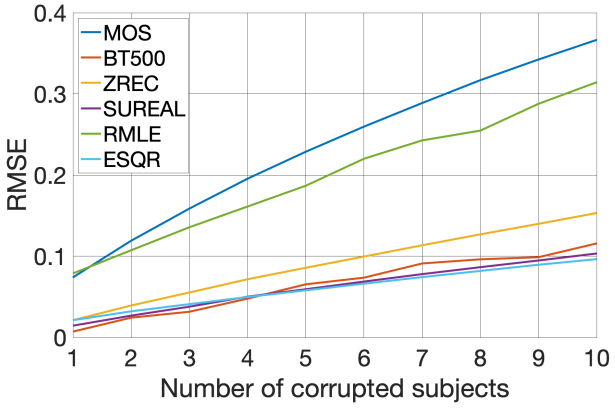
(e) NETF PUB



(f) VQ-HD1

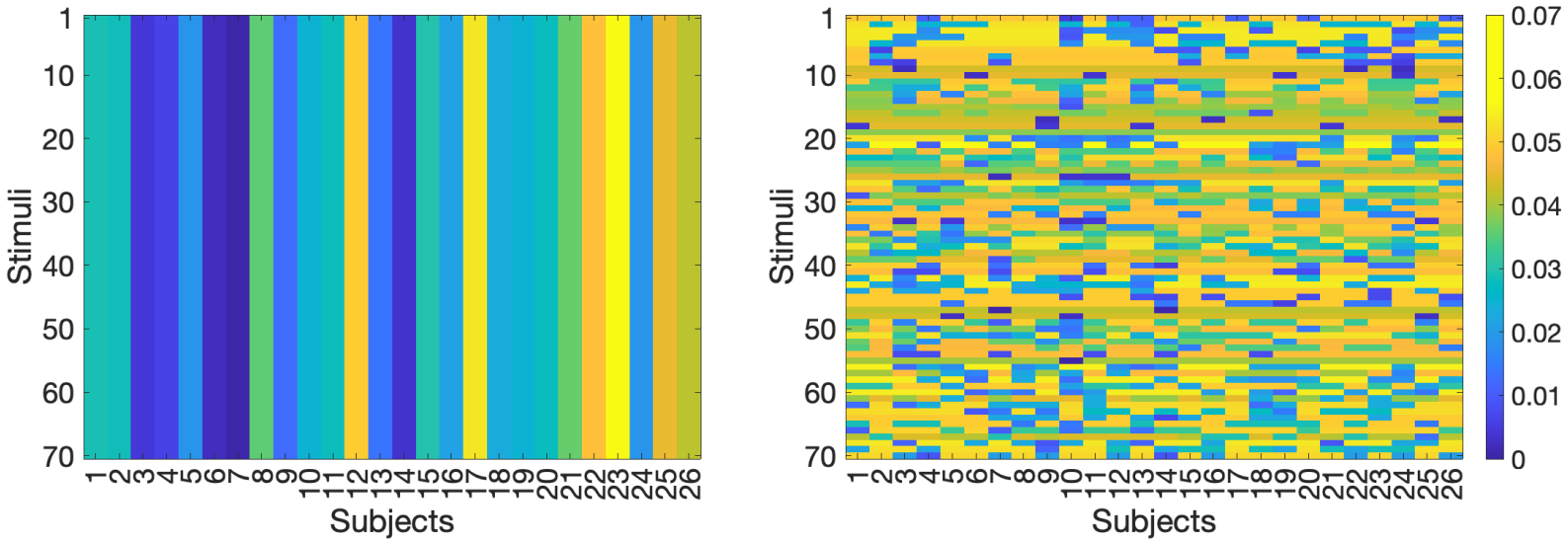


(g) VQ-HD3



(h) VQ-HD5

Results: Effectively Measuring the Importance of a Rating



- Importance of each rating as computed by SUREAL (left) and ESQR (right) for the PVSs in the Netflix Pub dataset
- The ESQR measure of importance can:
  - highlight cases where subjects #7 is still reliable
  - highlight cases where subjects #17 or #23 are unreliable
  - treat all subjects equally when they choose the same opinion as in the case of stimulus #19

Results: ESQR & Prior Art

- We noticed as expected that, the ESQR output is strongly aligned to that of existing quality recovery algorithms in general
- There are however cases where assumptions made by a specific method might be violated yielding a significant difference with ESQR

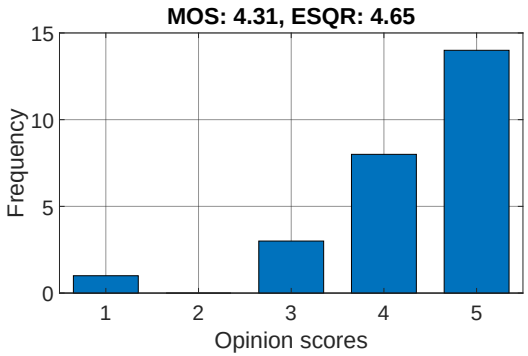
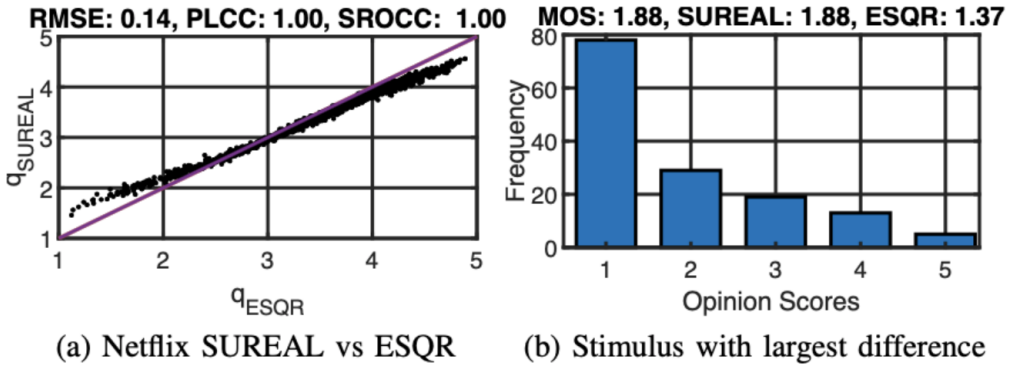


Figure: Example of stimulus from the Netflix Pub dataset on which ESQR and the MOS differ significantly



(a) Netflix SUREAL vs ESQR (b) Stimulus with largest difference

Figure: Comparing the output of SUREAL to that of ESQR on the MovieLens 1M dataset (Crowdsourcing)

Conclusions

Results synthesis

- ESQR recovers a subjective quality prone to lower uncertainty
- ESQR is robust to a noise uniformly distributed among all subjects
- ESQR competes with SUREAL and ZREC in terms of robustness to the introduction of spanner annotators
- ESQR effectively weights the importance of individual ratings
- ESQR differs from prior approaches mainly at the quality scale extremes

Open questions

- How to generalize ESQR to cases where pairwise correlations cannot be computed?
- How to account for the ordinal nature of the quality scale when measuring reliability?

Thanks for your attention