A non-parametric approach to subjective media quality recovery in the presence of spammer annotators

Lohic Fotio Tiotsop, Andrés Altieri, Giuseppe Valenzise

VQEG JEG-Hybrid

May 6, 2025

Outline

1. A Non-parametric approach to measure the reliability of a subject
2. The Non-Parametric subjective Quality Recovery (NPQR) algorithm
3. Computational results
4. Conclusions

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

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 Overview

- $\mathcal{S}$: set of stimuli, $\mathcal{I}$: set of subjects.
- $O_{i,s}$: subject i's score on stimulus s.
- Q_s: true quality, RV with pmf p_{Q_s}.
- $p_{O_{i,s}}$: distribution of of the opinion scores of subject i for stimulus s.
- Noise: Sometimes $p_{O_{i,s}}$ is different from p_{Q_s}.

Non-parametric Measure of Subject Reliability

Inter-Stimulus Consistency (Inter-SC):

- Measures how well a subject ranks different stimuli.
- Computed via Spearman correlation c_i between:
  - Modes of subject scores $\{\mathrm{mode}(O_{i,s})\}$,
  - Modes of ground truth $\{\mathrm{mode}(Q_s)\}$.
- Truncated: $\max(0, c_i)$.

Intra-Stimulus Consistency (Intra-SC):

- Measures repeatability and accuracy.
- Based on divergence between $p_{O_{i,s}}$ and p_{Q_s}.

Intra-Stimulus Consistency Metric

Intra-SC for subject i:

$$T_i = \left(\frac{1}{|\mathcal{S}_i|} \sum_{s \in \mathcal{S}_i} d(p_{O_{i,s}}, p_{Q_s}) \right)^{-1}$$

- In practice, $p_{O_{i,s}}$, cannot be easily estimated

proposition

Assume *$d(p_{O_{i,s}}, p_{Q_s}) = \mathbb{E}[f(p_{Q_s}, O_{i,s})]$ and $\mathrm{Var}[f(p_{Q_s}, O_{i,s})] < c$.*

Then, as $|\mathcal{S}_i| \to \infty$:

$$T_i^{-1} \approx \frac{1}{|\mathcal{S}_i|} \sum_{s \in \mathcal{S}_i} f(p_{Q_s}, O_{i,s})$$

- Implication: T_i can be estimated using only $O_{i,s}$ and p_{Q_s}.

Getting a suitable $f()$

Components of $d(p_{O_{i,s}}, p_{Q_s})$:

- **Entropy** (Repeatability):

$$H(p_{O_{i,s}}) = -\mathbb{E}[\log p_{O_{i,s}}(O_{i,s})]$$

- **KL Divergence** (Accuracy):

$$D_{\mathrm{KL}}(p_{O_{i,s}} \| p_{Q_s}) = \mathbb{E}\left[\log\left(\frac{p_{O_{i,s}}(O_{i,s})}{p_{Q_s}(O_{i,s})}\right)\right]$$

We define:

$$d(p_{O_{i,s}}, p_{Q_s}) = H(p_{O_{i,s}}) + D_{\mathrm{KL}}(p_{O_{i,s}} \| p_{Q_s})$$

$$= -\mathbb{E}[\log(p_{Q_s}(O_{i,s}))]$$

$$\Rightarrow f(p_{Q_s}, O_{i,s}) = -\log(p_{Q_s}(O_{i,s}))$$

Final Reliability Expression

Reliability of subject i:

$$R_i \approx \max(0, c_i) \cdot \left(\frac{1}{|\mathcal{S}_i|} \sum_{s \in \mathcal{S}_i} - \log p_{Q_s}(O_{i,s}) \right)^{-1}$$

Advantages:

- Purely non-parametric.
- Works with one rating per subject-stimulus pair.

NPQR Algorithm

1. Estimate p_{Q_s} as the empirical distribution (histogram) from $\{O_{i,s}\}_{i\in\mathcal{I}_s}$.
2. Compute Spearman correlation c_i for each subject.
3. Compute:

$$R_i \approx \max(0, c_i) \cdot \left(\frac{1}{|\mathcal{S}_i|} \sum_s - \log p_{Q_s}(O_{i,s}) \right)^{-1}$$

4. Estimate quality:

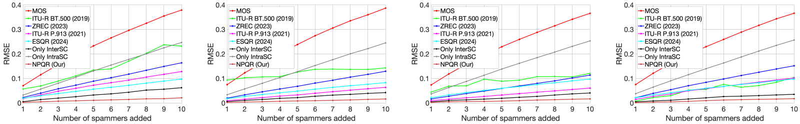
$$q_s = \frac{\sum_{i\in\mathcal{I}_s} R_i \cdot O_{i,s}}{\sum_{i\in\mathcal{I}_s} R_i}$$

Experimental Setup

- Compared NPQR with 5 state-of-the-art methods: MOS, ZREC, ITU-R P.913, BT.500, ESQR.
- Datasets:
  - **Controlled**: VQEG-HD1, VQEG-HD3, VQEG-HD5, Netflix Public.
  - **Crowdsourced**: KoNViD-1k, NIVD, MovieLens-1M.
- Evaluation metric: RMSE between quality estimates before and after introducing spammer annotators.

Robustness to Simulated Spammers

- Spammers simulated as users giving uniformly random scores in [1–5].
- NPQR maintains low RMSE even with increasing spammer count.
- Ablation: using only inter- or intra-SC worsens performance.

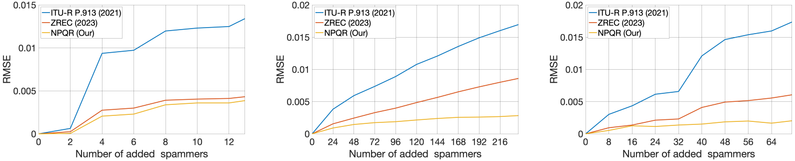


(a) NETF PUB (b) VQ-HD1 (c) VQ-HD3 (d) VQ-HD5

Figure: **Robustness to simulated spammers**. RMSE between the quality recovered on the original dataset and under noisy conditions. The noise was generated by adding simulated subjects (see the *x*-axis) that score the quality of each stimulus with an integer number sampled at random between 1 and 5.

Robustness to Real Spammers (Crowdsourcing)

- We consider the 3 approaches that measure subject reliability.
- Spammers identified as bottom 5% of reliability by 3 methods.
- Removed first, then gradually reintegrated.
- NPQR shows superior robustness to noise across all datasets.



(a) KoNViD-1k (b) NIVD (c) MoviesLens-1M

Figure: **Crowdsourcing experiments: robustness to common spammer annotators.** Pre-identified spammer annotators are progressively reintroduced into the data, and the robustness of each method is assessed via the RMSE between the reference quality (without spammers) and the quality recovered after reintegration. Lower RMSE indicates better robustness.

Key Takeaways

- NPQR outperforms other methods in terms of robustness to spammers.
- Both **inter**- and **intra-stimulus consistency** are crucial for robustness.
- Non-parametric nature allows generalization without model assumptions.

Thanks for your attention