Recover Subjective Quality Scores from Noisy Measurements

Zhi Li, Ioannis Katsavounidis
Netflix

Christos Bampis
University of Texas at Austin

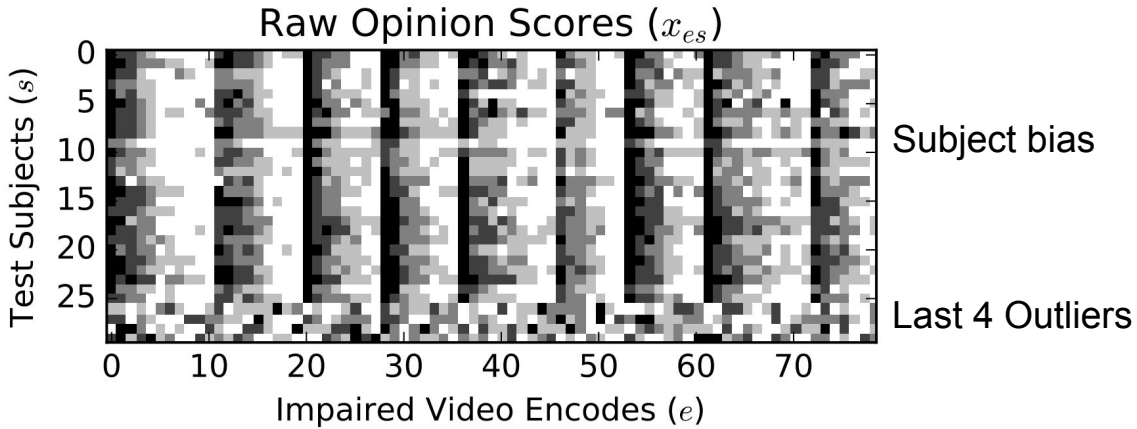
Patrick LeCallet
University of Nantes

VQEG Kraków 2017

Background - Acknowledgements

- Netflix has invested significant resources in video quality
  - VMAF (algorithmic development, subjective testing, OSS)
- Collaboration with research universities to address open problems
  - University of Nantes (Patrick Le Callet)
  - University of Southern California (C.-C. Jay Kuo)
  - University of Texas at Austin (Al Bovik)
  - More to come

Raw opinion scores are noisy and unreliable



Would MOS or DMOS be good enough?

Partial remedies

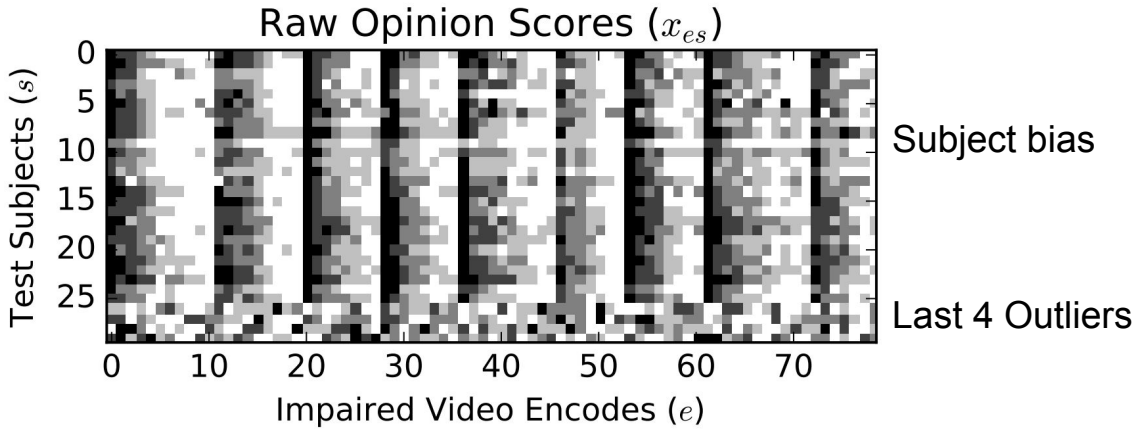
- Z-scoring - can only partially compensate subject bias
- Subject rejection

Subject rejection (ITU-R BT.500)

Algorithm 1 Subject rejection [8]

- Input: $x_{e,s}$ for $s = 1, ..., S$ and $e = 1, ..., E$.
- Initialize $p(s) \leftarrow 0$ and $q(s) \leftarrow 0$ for $s = 1, ..., S$.
- For $e = 1, ..., E$:
  - Let $Kurtosis_e = \frac{m_{4,e}}{m_{2,e}^2}$.
  - If $2 \leq Kurtosis_e \leq 4$, then $\epsilon_e = 2$; otherwise $\epsilon_e = \sqrt{20}$.
  - For $s = 1, ..., S$:
    - If $x_{e,s} \geq \mu_e + \epsilon_e \sigma_e$, then $p(s) \leftarrow p(s) + 1$.
    - If $x_{e,s} \leq \mu_e - \epsilon_e \sigma_e$, then $q(s) \leftarrow q(s) + 1$.
- Initialize $Set_{rej} = \emptyset$.
- For $s = 1, ..., S$:
  - If $\frac{p(s)+q(s)}{E} \geq 0.05$ and $\left|\frac{p(s)-q(s)}{p(s)+q(s)}\right| < 0.3$, then $Set_{rej} \leftarrow Set_{rej} \cup \{s\}$.
- Output: Set_{rej}.

BT.500 limitations



- All scores corresponding to rejected subjects are discarded -- an overkill
- Often only identifies a subset of outliers
  - In the example above, only subjects #26, #28 and #29 were identified
- Does not generalize well for selective sampling (i.e. missing data)

Can we do better?

Take into account subject characteristics

- Subject bias
  - Picky viewers tend to be biased toward lower scores
  - Not every subject has “golden eyes” - their sensitivity to impairment varies
  - Different sessions
- Subject inconsistency
  - Subjects may not rate consistently throughout a viewing session
  - Outliers - a special case with very large inconsistency

First need a model to capture these factors !!

Modeling raw opinion score

Raw Score: $X_{e,s}$ = True Score: $x_e + B_{e,s}$, for $e = 1, \ldots, E$ (Impairment (PVS)), $s = 1, \ldots, S$ (Subject)

where:

$$B_{e,s} \sim N(b_s, v_s^2)$$

Subject Bias: b_s; Subject Inconsistency: v_s^2

Independently validated by:
L. Janowski and M. Pinson, “The accuracy of subjects in a quality experiment: A theoretical subject model,” IEEE Transactions on Multimedia, Dec 2015.

Proposed approach

Main idea: find unknown parameters to maximize likelihood function of the observations - maximum likelihood estimation (MLE)

Example problem size

- # observations: 300 (PVS) * 30 (Subject) = 9000
- # unknowns:
  - True quality scores (300)
  - Subject Bias (30)
  - Subject inconsistency (30)

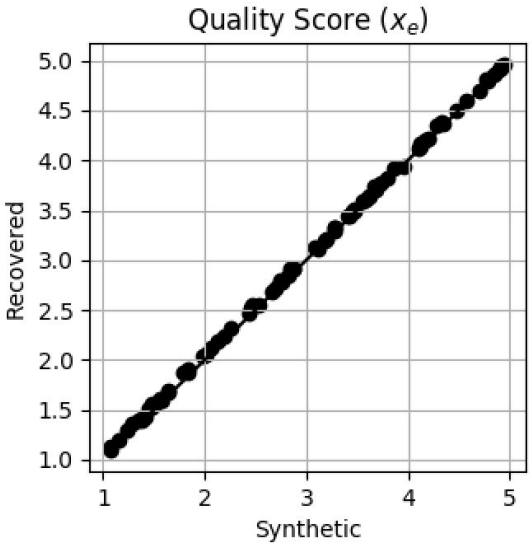
A solution based on Belief Propagation

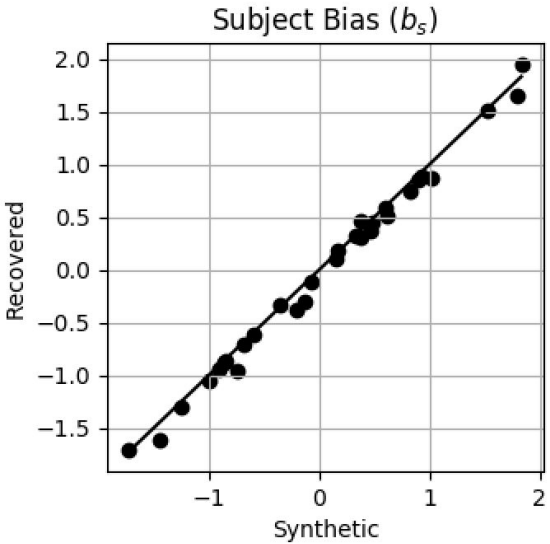
Algorithm 2 BP solution for the proposed MLE formulation

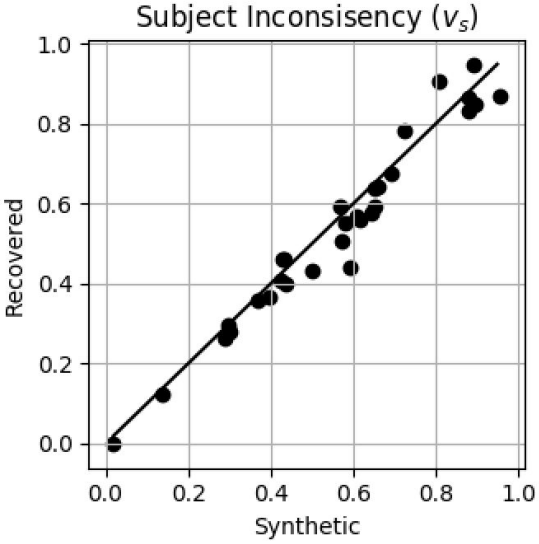
- Input:
  - $x_{e,s}$ for $s = 1, ..., S$ and $e = 1, ..., E$.
  - Refresh rate α.
  - Stop threshold Δx^{thr}.
- Initialize $\{x_e\} \leftarrow \{\mu_e\}$, $\{b_s\} \leftarrow \{0\}$, $\{v_s\} \leftarrow \{\sigma_s\}$
- Loop:
  - $\{x_e^{prev}\} \leftarrow \{x_e\}$.
  - $b_s \leftarrow (1-\alpha) \cdot b_s + \alpha \cdot b_s^{new}$ where $b_s^{new} = b_s - \frac{\partial L(\theta)/\partial b_s}{\partial^2 L(\theta)/\partial b_s^2}$ for $s = 1, ..., S$.
  - $v_s \leftarrow (1-\alpha) \cdot v_s + \alpha \cdot v_s^{new}$ where $v_s^{new} = v_s - \frac{\partial L(\theta)/\partial v_s}{\partial^2 L(\theta)/\partial v_s^2}$ for $s = 1, ..., S$.
  - $x_e \leftarrow (1-\alpha) \cdot x_e + \alpha \cdot x_e^{new}$ where $x_e^{new} = x_e - \frac{\partial L(\theta)/\partial x_e}{\partial^2 L(\theta)/\partial x_e^2}$ for $e = 1, ..., E$.
  - If $\left(\sum_{e=1}^{E}(x_e - x_e^{prev})^2\right)^{\frac{1}{2}} < \Delta x^{thr}$, break.
- Output: $\{x_e\}$, $\{b_s\}$, $\{v_s\}$

Implementation at: github.com/Netflix/vmaf/tree/master/python/src/vmaf/mos

Algorithm Validation: Synthetic Data



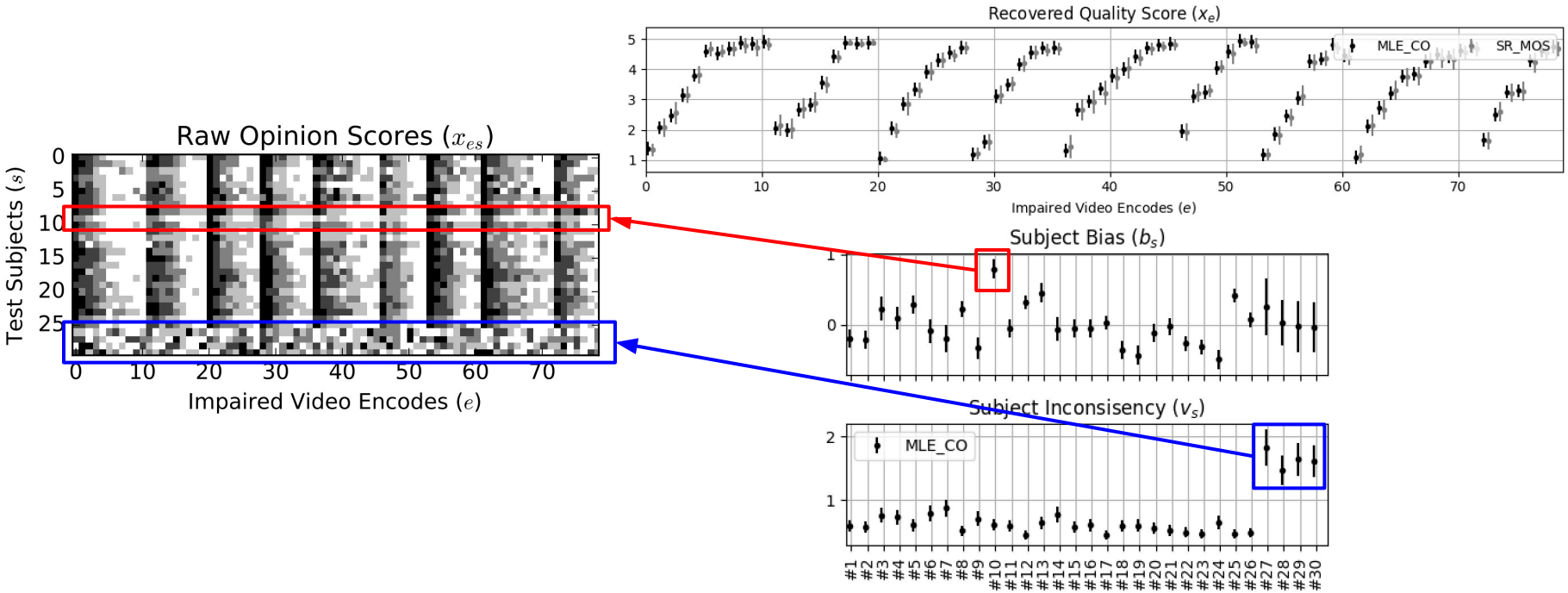




Synthetic data generation

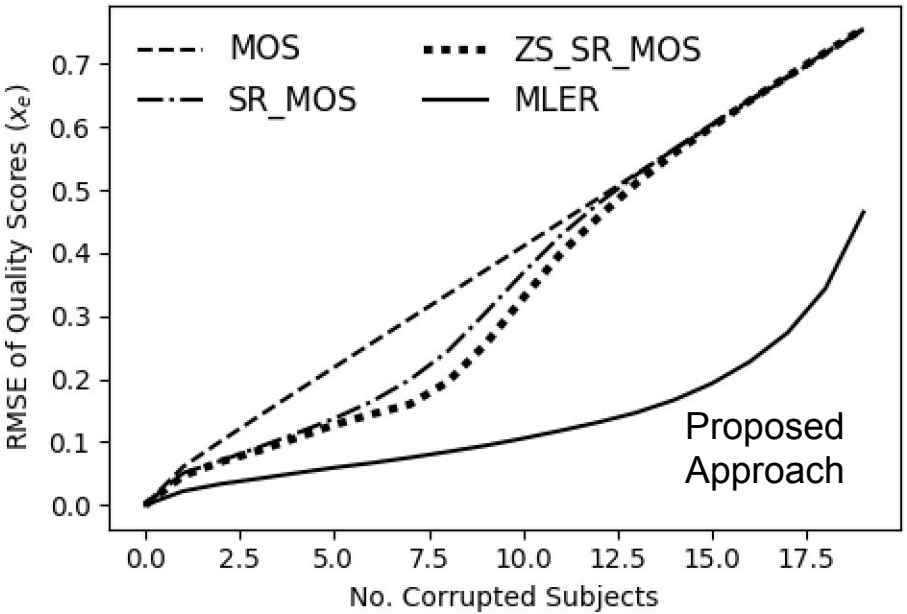
- Randomly generate parameters according to $x_e \sim U[1, 5]$, $b_s \sim N(0, 1)$, $v_s \sim U[0, 1)$
- Randomly generate observations according to parameters and model

Sample recover results



Resistance to outliers

Worse



Better

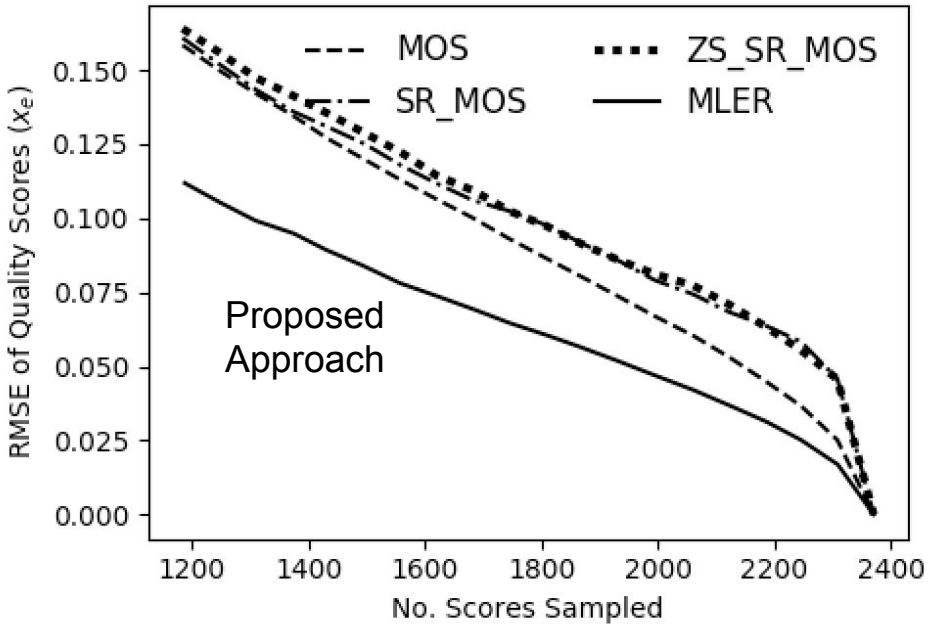
ZS - Z-scoring
SR - Subject rejection

Y-axis: RMSE w.r.t. clean case

Selective sampling in the presence of outliers

Worse

Better



ZS - Z-scoring
SR - Subject rejection

Y-axis: RMSE w.r.t. clean case

Conclusions

Jointly estimating quality scores with subject characteristics yields more robust recovery against outliers than the BT.500 recommendation, and tighter confidence intervals.

Recovered side information provides additional insight on subjects' bias and inconsistency.

Work in progress: content ambiguity

$$X_{e,s} = x_e + B_{e,s} + A_{e,s} \text{ for } e = 1, \ldots, E, s = 1, \ldots, S$$

Video Encode

Subject

$$B_{e,s} \sim N(b_s, v_s^2)$$

$$A_{e,s} \sim N(0, a_{c:c(e)=c}^2)$$

Content Ambiguity

Thank you

Source code at:
github.com/Netflix/vmaf/tree/master/python/src/vmaf/mos

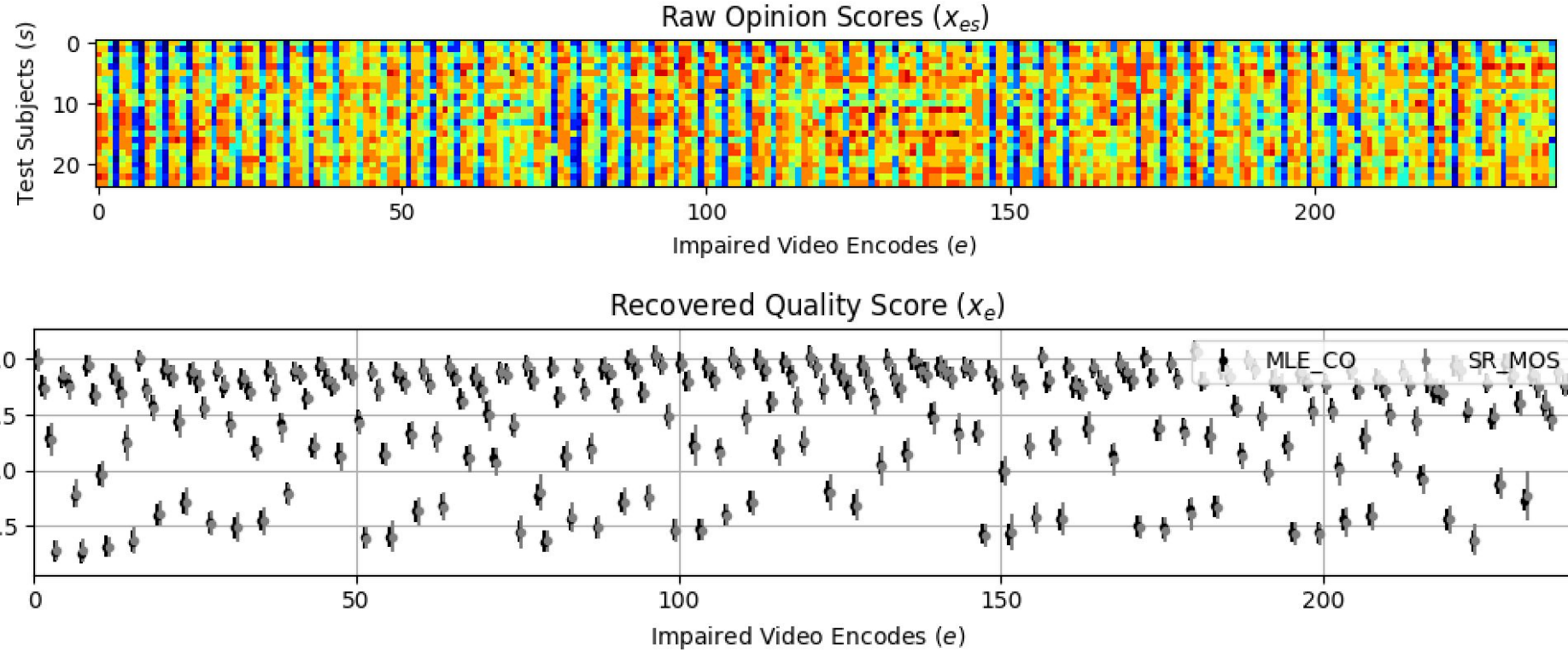
Will also release a stand-alone version soon.

NETFLIX

More datasets

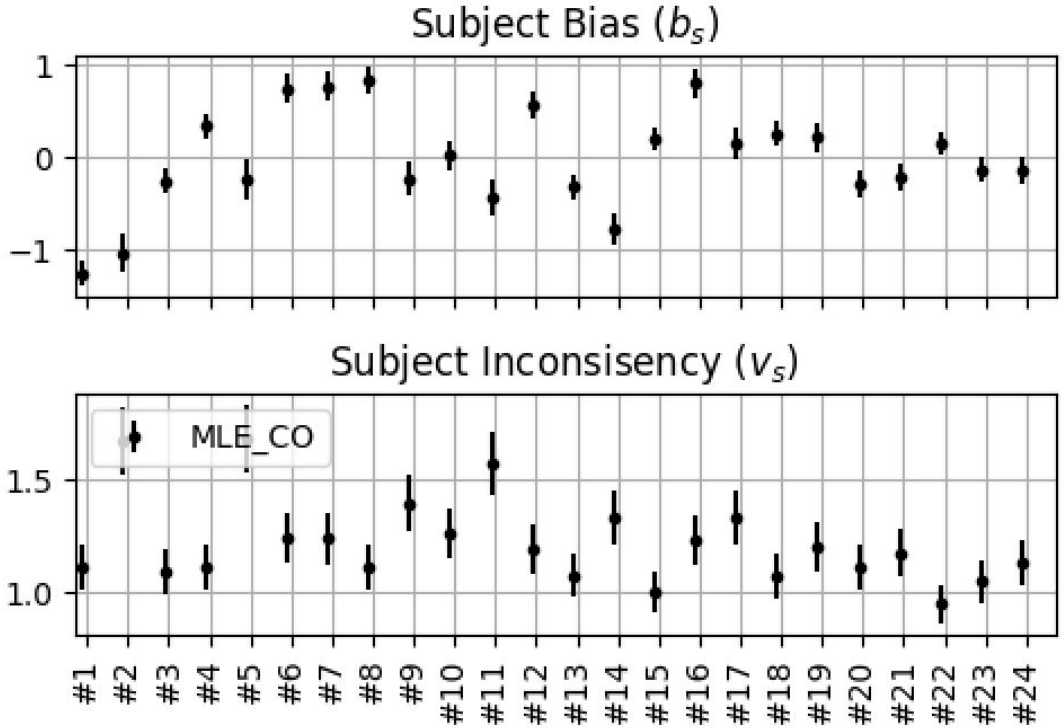
NETFLIX

Yonsei UHD ACR Dataset

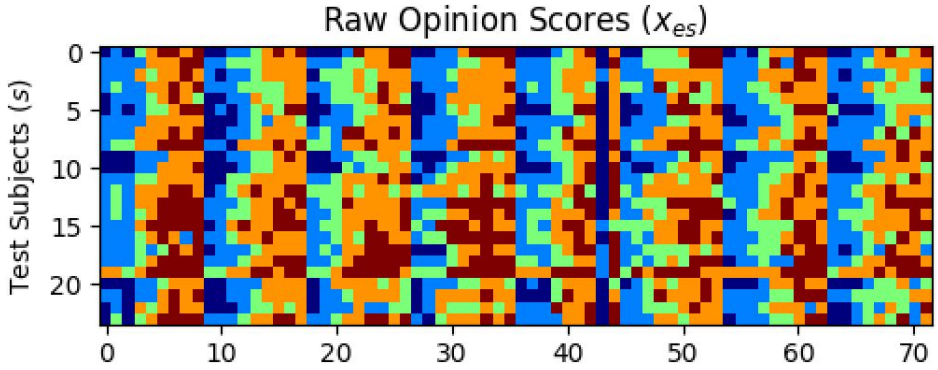


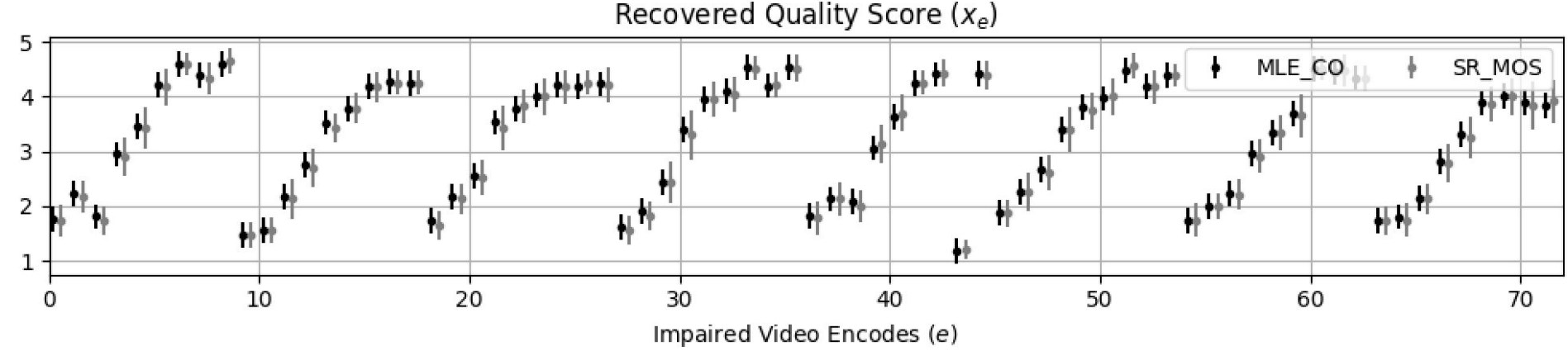
NETFLIX

Yonsei UHD ACR Dataset (Cont'd)



VQEG HD3 Dataset





VQEG HD3 Dataset (Cont'd)