Overview of SAM Activities

Lucjan Janowski (AGH)
Zhi Li (Netflix)

VQEG Meeting, Shenzhen, China
Oct. 15, 2019

Talk Outline

- What is SAM?
- Early work
- Recent development
- Future plans

SAM

- SAM - Statistical Analysis Methods
- Mission:
  - The SAM group addresses problems related to how to better analyze and improve data quality coming from subjective experiments and how to consider uncertainty in objective media quality predictors/models development

Goals – Long Term

- Improve methods used to draw conclusions from subjective experiments
- Understand the process of expressing opinion in a subjective experiment
- Improve subjective experiment design to facilitate analysis and applications
- Improve the analysis of objective model performances

Goals – Mid Term

- Popularize the analysis related to the subject model by publishing a white paper and ITU recommendation modification
- Revisit standardized methods for the assessment of the performance of objective model performances

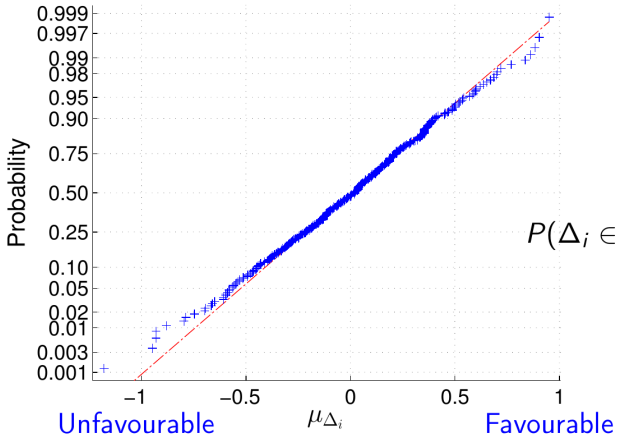
Goals – Short Term

- Unify notation used for analysis
- Create a common subjective data input format
- Fix a stability problem of parameter estimation for the subject model based on Maximum Likelihood Estimation (MLE) method proposed by Li et al.

Early work (Janowski&Pinson'15)

$$U_{ij} = \psi_j + \Delta_i + \upsilon_i X + \phi_j Y$$

- U_{ij} – r.v. describing raw opinion scores
- Ψ_j – true quality of PVS j
- Δ_i – voting bias of subject i
- υ_i – voting inconsistency (std) of subject i
- Φ_j – ambiguity (std) of PVS j
- Empirical result to validate that additive model is better than multiplicative model in fitting real-world data



$E(\Delta_i) = 0$

$\sigma(\Delta_i) = 0.34$

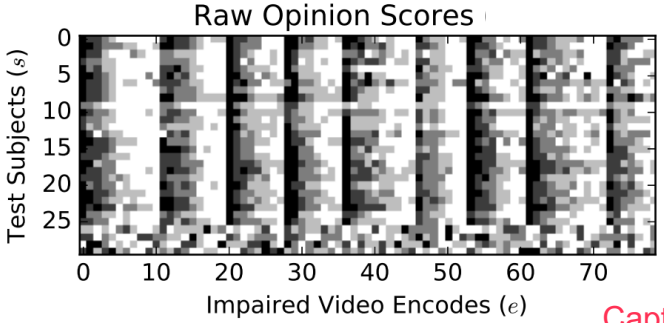
$P(\Delta_i \in \underbrace{(-0.5, 0.5)}_{25\% \text{ of scale}}) \approx 0.85$

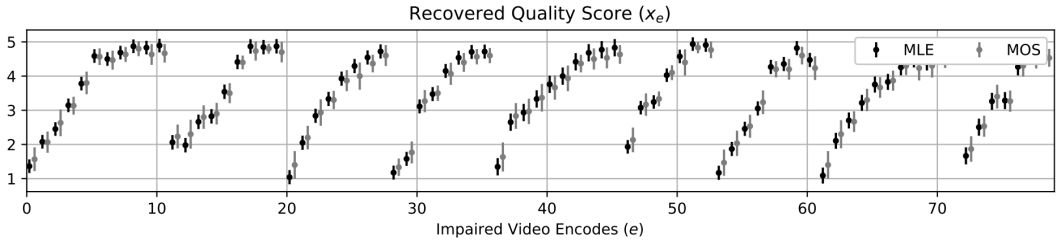
Early work (Li&Bampis'17)

$$U_{ij} = \psi_j + \Delta_i + \upsilon_i X + \rho_{k:\boldsymbol{k}(j)=k} Y$$

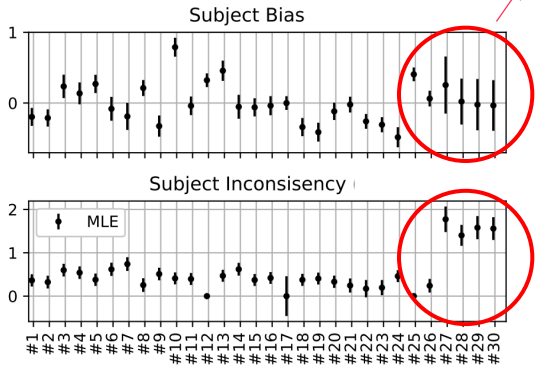
- U_{ij} – r.v. describing raw opinion scores
- Ψ_j – true quality of PVS j
- Δ_i – voting bias of subject i
- υ_i – voting inconsistency (std) of subject i
- $\rho_{k:\boldsymbol{k}(j)=k}$ – ambiguity (std) of SRC k
- Model outlier subjects as having large bias and inconsistency
- Maximum likelihood estimation (MLE) and belief propagation (BP) method to solve model parameters

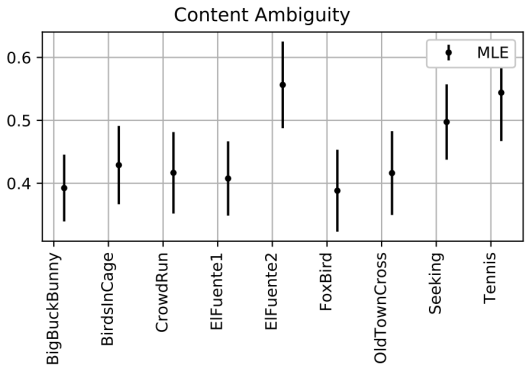
Example Result – Li&Bampis'17

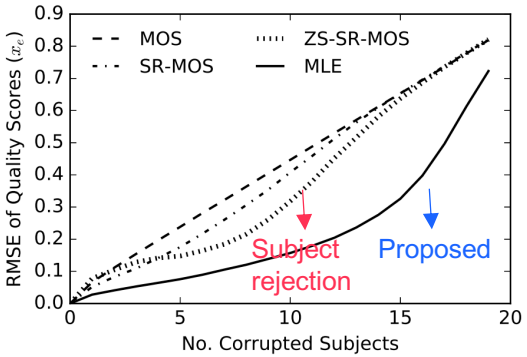




Capturing outliers by large variance and loose confidence interval







Recent Development

- Unified notations used for analysis
- SuJSON – a common subjective data input format
- A simplified discrete model
- Bayesian methods to address stability issue of MLE solutions
- Application to adaptive media playout [Pérez, García et al.]
- Error origin of SRC or HRC?
- Generalized score distribution (GSD)
- Paired comparison and active learning
- Planning the number of subjects [Kjell et al.]

Unified Notations Used for Analysis

- By unifying the notations, we hope to create a common language between different subjective model algorithms, details: https://arxiv.org/abs/1903.05940

- i for a subject
- j for a PVS,
- k for an SRC,
- r for a repetition,
- o for an order, and
- h for an HRC

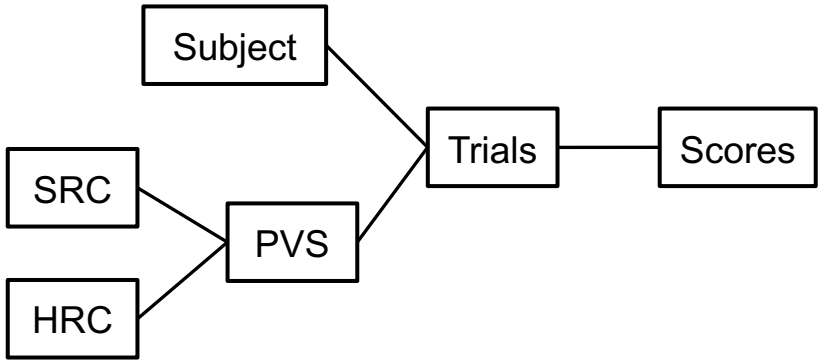
- u as a single subject answer,
- ψ (psi) as a true quality,
- Δ (Delta) as a subject bias,
- υ (upsilon) as a standard deviation related with a given subject,
- ϕ (phi) as a standard deviation related with a given PVS
- ρ (rho) as a standard deviation related with a given SRC

$$U_{ij} = \psi_j + \Delta_i + \upsilon_i X + \phi_j Y$$ [Janowski&Pinson'15]

$$U_{ij} = \psi_j + \Delta_i + \upsilon_i X + \rho_{k:\boldsymbol{k}(j)=k} Y$$ [Li&Bampis'17]

SuJSON – A Common Subjective Data Input Format


```
{
  "dataset_name": "vqeghd1",
  "sujson_version": "1.1-in_progress",
  ...
  "src": [
    {
      "id": 1,
      "name": "NTIA Red Kayak",
      "path": "/vqeghd1_src01_hrc00.v1.yuv"
    },
    ...
  ],

  "hrc": [
    {
      "id": 1,
      "characteristics": {
        "codec": "MPEG-2",
        "bit_rate": 6,
        "plr": "133c",
        "fps": 29.97,
        "interlace": false
      }
    },
    ...
  ],

  "pvs": [
    {
      "id": 1,
      "hrc_id": 1,
      "src_id": 1,
      "path": "/vqeghd1_src01_hrc01.v1.avi"
    },
    ...
  ],

  "subjects": [
    {
      "id": 1
    },
    ...
  ],

  "trials": [
    {
      "id": 1,
      "subject_id": 1,
      "task_id": 1,
      "pvs_id": 1,
      "score_id": 1
    },
    ...
  ],

  "scores": [
    {
      "id": 1,
      "question_id": 1,
      "pvs_id": 1,
      "score": 2
    },
    ...
  ],
}
```

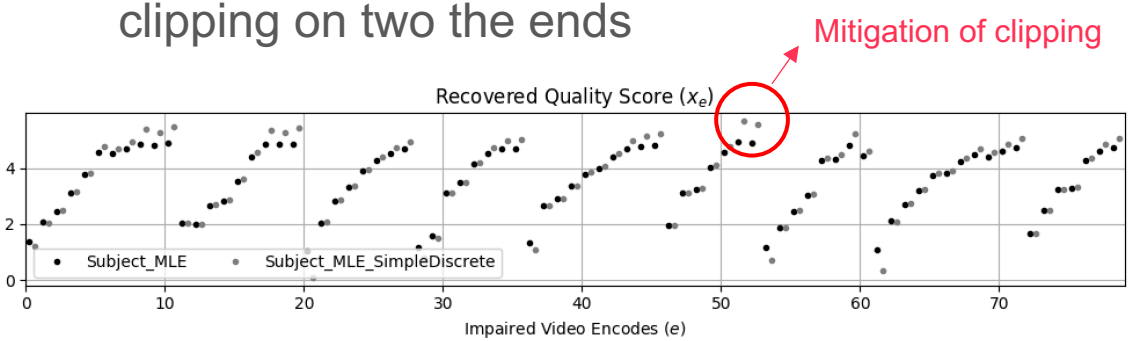
https://github.com/LucjanJanowski/translator-to-suJSON

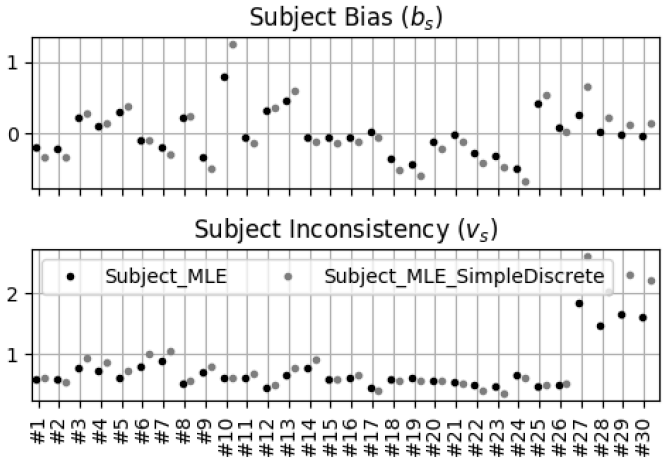
A Simplified Discrete Model

- Simplified discretized model $U_{ij} = Q(\psi_j + \Delta_i + \upsilon_i X)$

$$P(U_{ij} = u) = \begin{cases} \int_{-\infty}^{1.5} \frac{1}{\sqrt{2\pi}\upsilon_i} e^{-\frac{(u-\psi_j-\Delta_i)^2}{2\upsilon_i}} & u = 1 \\ \int_{u-0.5}^{u+0.5} \frac{1}{\sqrt{2\pi}\upsilon_i} e^{-\frac{(u-\psi_j-\Delta_i)^2}{2\upsilon_i}} & u \in \{2,3,4\} \\ \int_{4.5}^{\infty} \frac{1}{\sqrt{2\pi}\upsilon_i} e^{-\frac{(u-\psi_j-\Delta_i)^2}{2\upsilon_i}} & u = 5 \end{cases}$$

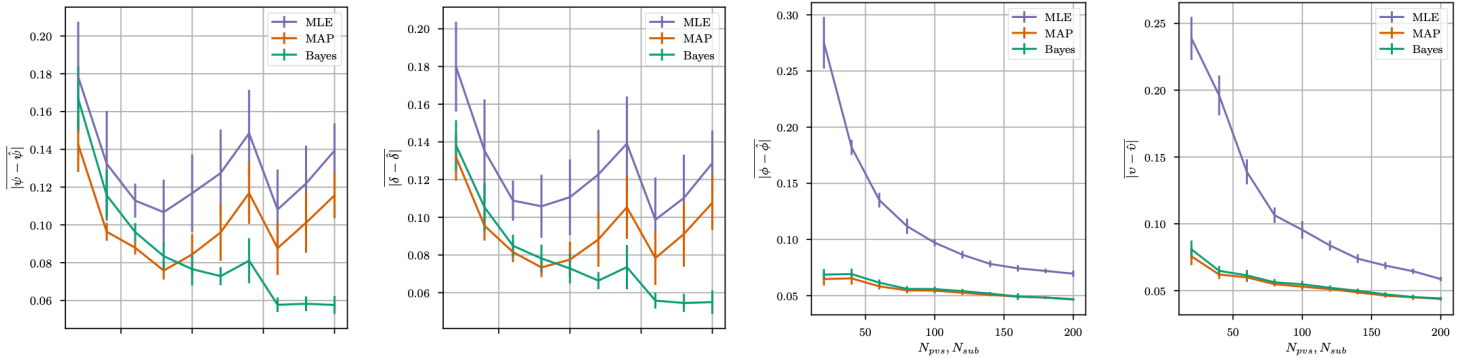
- Taking into account: effect of discrete scale and clipping on two the ends





Bayesian Methods to Address Stability Issue of MLE Solutions (Rusek et al.)

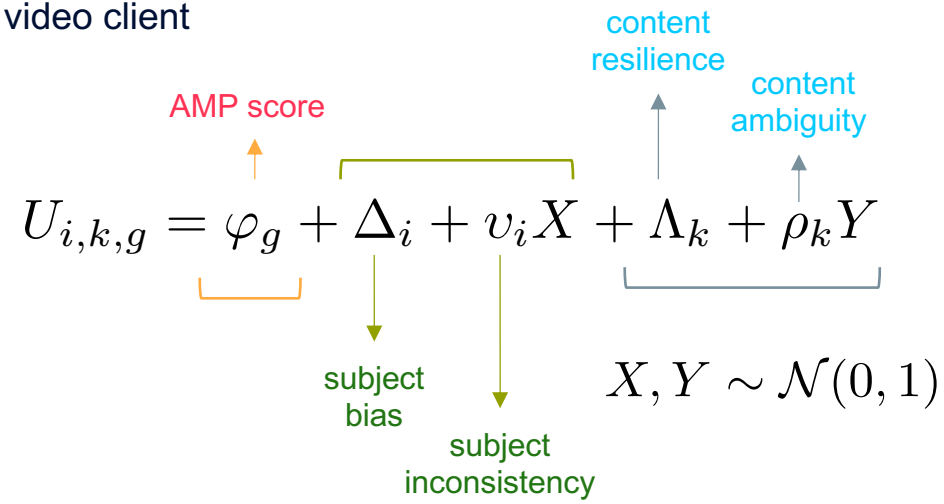
- MLE solution is a special case of more general Bayesian methods such as MAP (maximum a posteriori) estimation and full Bayesian
- Solution stability issue: $U_{ij} = Q\left((\psi_j - \alpha) + (\Delta_i + \alpha) + (\upsilon_i - \beta)X + (\psi_j + \beta)Y\right)$
- MAP: $\hat{\theta} = \arg\max_{\theta} \mathcal{L}(u|\theta) + \log P(\theta)$
- Full Bayesian: $\hat{\theta} = \mathbf{E}_{P(\theta|u)}\theta$



Subjective Assessment of Adaptive Media Playout for Video Streaming [Pablo Pérez, Narciso García, and Álvaro Villegas – QoMEX 2019]

- Experiment on subjective assessment of Adaptive Media Playout (AMP)
  - Dynamically changing playout speed at the video client
- Application of modified subject model

Score for
- Subject *i*
- SRC *k*
- HRC *g*



- Insights on
  - AMP quality itself
  - Subject behavior / response characterization

NOKIA Bell Labs

Error Origin of SRC or HRC?

- Compare two models: SRC-only vs. HRC-only

$$M_{src}: \quad U_j = \psi_j + \rho_{k:k(j)=k} X,$$

$$M_{hrc}: \quad U_j = \psi_j + \xi_{h:h(j)=h} X,$$

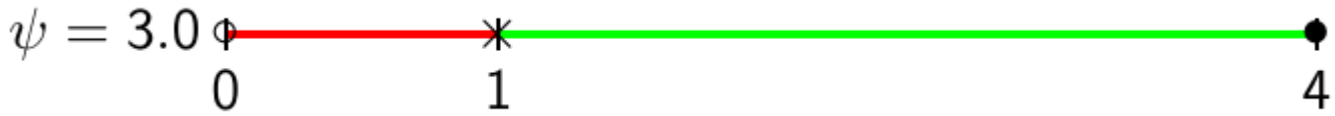
- Which one fits real data better?
- Observation: neither model fits real data well, with SRC-only worse than HRC-only

Generalized Score Distribution (GSD)

$$P(X = 1) = F_1(\psi), P(X = 2) = F_2(\psi)$$

Example: If $\psi < 2$: $P(X = 1) = 2 - \psi$, $P(X = 2) = \psi - 1$

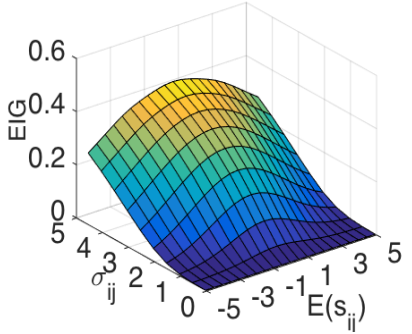
$$U \sim GSD(\psi, \rho)$$



Paired Comparison (PC) and Active Learning

Boosting pair comparison

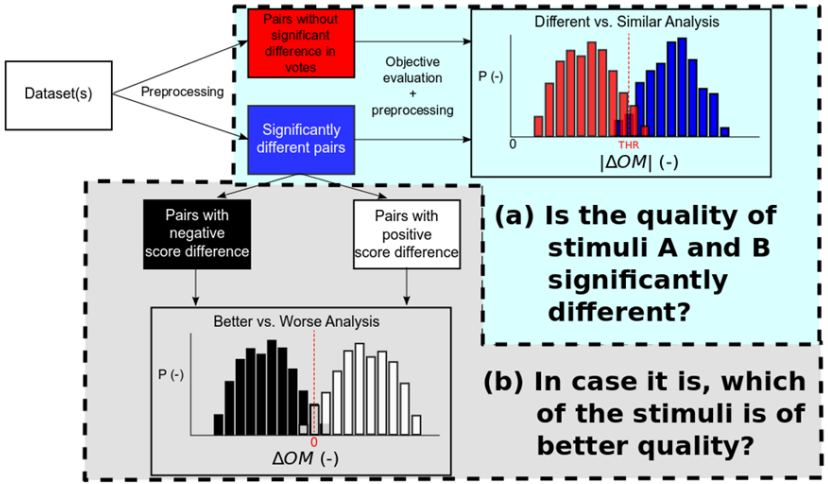
- Learn which pair could generate the maximum information gain (EIG)
- Bayesian theory (prior and posterior)



$$U_{ij} = \int \sum_{y_{ij}} log \left\{ \frac{p(s_{ij}|y_{ij})}{p(s_{ij})} \right\} p(s_{ij}|y_{ij}) p(y_{ij}) ds_{ij}$$

Objective metrics evaluation using PC data

- Can the metric determine if quality of stimuli is significantly different?
- Can the metric determine which stimulus is preferred in any different pair?



Multiple Comparisons and Planning Number of Test Subjects

- Planning and design a subjective test based on the expected power in the statistical analysis, the estimated variance and the number of comparisons, the needed number to test subjects can be estimated.
  - Journal paper: Brunnström, K. and M. Barkowsky, *Statistical quality of experience analysis: on planning the sample size and statistical significance testing.* Journal of Electronic Imaging, 2018. **27**(5): p. 11. PDF http://www.diva-portal.org/smash/get/diva2:1252987/FULLTEXT01.pdf
  - ITU-T contribution: P.1401 will be updated, P910, P.913 and BT.500 is still under discussion
  - R-code: https://github.com/VQEG/number-of-subjects
  - GUI: https://slhck.shinyapps.io/number-of-subjects/ (by Werner Robitza)

Future Plans

- Prepare a document on ITU standard modification (ITU-T P.1401, ITU-R BT.500)
- Continued development on Generalized Score Distribution
- Temporal behavior analysis of subjective experiments
- Continued paired comparison (PC) methodologies investigation
  - Apply subject/SRC/HRC-based MLE analysis to PC
  - Active learning