Accepted for IEEE Transactions on Multimedia

Estimation of Quality Scores from Subjective Tests - beyond Subjects' MOS

Sergio Pezzulli, Maria G. Martini, *Senior Member, IEEE,* Nabajeet Barman, *Member, IEEE*

***Abstract*—Subjective tests for the assessment of the quality of experience (QoE) are typically run with a pool of subjects providing their opinion scores using a 5-point scale. The subjects' mean opinion score (MOS) is generally assumed as the best estimation of the average score in the target population. Indeed, for a large enough sample, we may assume that the mean of the variations across the subjects approaches zero, but this is not the case for the limited number of subjects typically considered in subjective tests. In this paper, we propose an approach based on generalized linear models (GLMs) for estimation of the population average QoE. The motivating dataset is composed of the individual scores assigned by 25 subjects to a set of gaming videos evaluated under different resolutions and compression ratios. The approach recognizes the multinomial nature of the data and allows for correlation between scores of the same subject. The resulting estimated average QoE is shown to follow more credible patterns than the MOS, particularly for higher bitrates, for which the model estimates present more coherent behavior. Similar convincing results are found on a second dataset, showing the validity of the approach.**

I. Introduction

While the design of multimedia services in the past was performed by relying only on quality of service (QoS) criteria, delivering an appropriate quality of experience (QoE) is increasingly important, and the capability of measuring it accurately is crucial in order to select the best transmission system technologies and parameters. The most appropriate way to measure QoE is by collecting the opinions of users via subjective tests. Subjective tests for quality of experience are typically run with a pool of subjects providing their opinion scores using a 5-point scale. The mean opinion score (MOS) of subjects is generally assumed to be the best estimation of the quality [1] [2]. Subjects present variations in perceiving and assessing the quality, and it is known that if the sample of subjects is large enough, the mean of the collected opinion scores approaches the population mean. More precisely, the mean is a consistent estimator of the population mean; i.e., it converges in probability to the population mean when the sample size tends to infinity. [1]

However, performing subjective tests with a large number of subjects is expensive in terms of time and resources. Thus, for practical reasons, only a small number of subjects are involved (e.g., 15 is the minimum recommended number according to ITU [1] [2], and this number is often used in actual tests). On the other hand, subjective tests are often performed on several videos presenting limited variations in terms of technical features and content. Therefore, the use of an appropriate modeling technique may help in distinguishing the individual variability from the relative merit of each video for estimating the population average QoE. The MOS, in fact, can be considered as the population mean estimate according to a model characterized by the maximum number of linearly independent parameters, which can be compared to simpler alternatives by using standard model selection techniques.

In this paper, we demonstrate this approach on a dataset composed of the individual scores assigned by 25 subjects to a set of gaming videos evaluated under different resolutions and bitrates [3]. We apply a model that recognizes the ordinal multinomial nature of the data and allows for correlation between scores of the same subject. The resulting estimated average QoE is shown to follow more credible patterns than the MOS, particularly for higher bitrates, for which the model estimates present more coherent behavior.

The main contributions of this paper are the following:

- A detailed analysis of the subjective scores from the dataset in [3] is performed in terms of subject consistency and dependence of the subjects' opinion scores on the content. Such analysis can benefit the research on quality of experience of gaming videos and further studies on statistics and models for quality assessment in general. The dataset in [3] is publicly available (processed video sequences (PVSs) and associated MOS scores). The per-subject scores will also be made available upon publication of this paper.
- A modeling technique for estimating the average QoE in the population (which we will refer to in the following as estimated population mean opinion score (EPMOS)) jointly exploits the information within the entire dataset. Such a model can be used as a replacement for MOS across subjects. We applied this modeling technique to the dataset in [3]. To prove the general validity of the approach, in Appendix 1, we also report the results on a second example regarding a dataset of videos of natural scenes [4].
- The software that implements the model is made publicly available to enable reproducible results and application of the model to different datasets.[2]

Sergio Pezzulli is with the University of Rome La Sapienza, Italy.

Maria Martini and Nabajeet Barman are with Kingston University London, UK.

Manuscript received April 3, 2019; revised August 28, 2019; April 10, 2020.

[1] Let M_n denote the sample mean over a sample of size n and let μ denote the population mean. Then, for all $\varepsilon > 0$, $P(|M_n - \mu| > \varepsilon) \to 0$ when $n \to \infty$.

[2] The link for downloading the per-subject scores and the code will appear here.

The remainder of this paper is structured as follows. Section II presents the related work. Section III introduces the dataset considered in this study. A detailed data analysis is presented in Section IV. Section V introduces the proposed model, the results of which are presented in Section VI. A final discussion and conclusions are presented in Section VII.

II. Related work

The Likert scale [5] was developed in 1932 as a five-point scale used for responses in surveys of opinions, with the labels of the original five categories of response ranging from strongly disagree, corresponding to 1, to strongly agree, corresponding to 5. Such scales fall within the ordinal level of measurement since the response categories have a rank order, but the intervals between values are not necessarily equal. However, it is common in research to assume that such intervals are equal [6]. We will find confirmation of this assumption in our study.

For quality assessment tests based on the Likert scale, a number of statistical tests are typically used to analyze the data. In [7], three reasons are listed to explain why the use of various parametric methods, such as analysis of variance and regression, is not appropriate: (a) the sample size is too small, (b) the data are not normally distributed, and (c) the data are from Likert scales, which are ordinal, so that parametric statistics cannot be used. In the same paper, however, the author states that many studies consistently show that parametric statistics are robust with respect to violations of the underlying assumptions.

Similar considerations are made in [8], where the authors aim at fixing a common practice of improper use of statistical tests.

Recently, statistical quality of experience analysis was also discussed in [9], focusing in particular on planning the sample size based on the requested accuracy and statistical significance testing.

The advantages of considering quality of experience distributions rather than mean opinion scores are highlighted in [10], where the author proposes to consider the full QoE distribution over the ordinal rating categories for evaluating and reporting QoE results instead of using MOS-based metrics.

In the following, we discuss the two main elements influencing the results of a subjective test: the reliability of subjects and the type of content used in the tests.

A. Reliability of subjects

As recommended in [1], the reliability of the subjects can be qualitatively evaluated by checking their behavior when "reference/reference" pairs are shown. In this case, we expect the score to be the maximum one (5 if a 5-point ordinal scale is used), and we can assume that the subject has low reliability if the score provided is far from this value.

In addition, the reliability of the subjects can be checked by using procedures described in [2] for the single stimulus continuous quality evaluation (SSCQE) method. In this method, the reliability of the votes depends on the following two parameters: systematic shifts and local inversions. During a test, a viewer may be too optimistic or too pessimistic or may have misunderstood the voting procedures (for instance, the voting scale). This case can lead to a series of votes being systematically shifted from the average series. On the other hand, observers can sometimes vote without devoting sufficient attention. In this case, local inversions can be observed.

The use of a tool allowing detection and, if necessary, discarding of inconsistent observers is recommended in [1].

In [2], a methodology for screening observers is provided, with a first step based on the mean, standard deviation, and kurtosis of the data to discard observers who have produced votes that are significantly distant from the average scores. A second step is proposed for the detection of local vote inversions, where the scores are preliminarily centered around the overall mean to minimize the shift effect that has already been treated during the first process stage.

A new method of data filtration is presented in [11]. The method proposes the use, for subjects' scores, of Mandel's k and h statistics that were developed for the comparison of inter-laboratory experiments [12], considering "the subject as a laboratory." The method results in a decrease in the MOS standard deviation, which is exemplified via SSCQE data of compressed video results.

To deal with the inevitable variations between each subject's use of the quality scale, and possibly also across sessions, Z-scores are typically computed [13].

B. "Criticality" of the content and PVS

In [14], the authors studied the relationship between MOS and standard deviation of opinion scores (SOS). They included a factor depending on the artifacts/use case (e.g., image coding artifacts, video streaming, and cloud gaming), measuring the difficulty that subjects encountered in assessing the quality of a particular dataset. ITU recommendations [2] also highlight that the scores obtained for different test sequences are dependent on the criticality of the test material used. For this reason, presenting results for different test sequences separately, rather than only as aggregated averages across all test sequences, is recommended. The "picture content failure characteristic" of the system under testing can be observed by arranging the results for individual test sequences in a rank order of test sequence criticality on an abscissa [2]. However, the ITU recommendation highlights that this form of presentation only describes the performance of the codec and does not provide an indication of the likelihood of the occurrence of sequences with a given degree of criticality. Further studies of test sequence criticality and the probability of occurrence of sequences of a given level of criticality are hence recommended.

C. Modeling subject bias and influence of content/PVS

Some recent works have proposed theoretical models for the characterization of subjects performing subjective tests [15] [16]. These models postulate that the obtained subjective score of each PVS can be considered a combination of a true quality score associated with the PVS and two additional terms, typically depending on content, associated with the

subject bias and inconsistency. Subject bias refers to the fact that some viewers tend to be biased toward lower scores and vice versa. Subject inconsistency refers to the fact that viewers may not assign the same score to the same PVS in a second visualization. Some subjects tend to rate more consistently than others.

In [15], the authors study subject bias and scoring error as functions of both PVS and subject. They propose normalizing the opinion scores with subject bias, mentioning that this appears to improve the ability of datasets to distinguish between PVS and MOS.

A maximum likelihood estimation (MLE)-based quality recovery model was presented in [16], enabling the estimation of subject bias and subject inconsistency, which could be used to reduce the number of subjects in a test to reach a given discriminability. The authors in [17] conducted a discriminability vs. numbers of subjects analysis, employing the score recovery model from [16], highlighting the potential savings in terms of number of subjects required.

Our work also addresses the impact of subject bias and PVS on quality assessment. Unlike [16], we abandon the normality assumption in favor of the multinomial distribution, which is more appropriate for data on the Likert scale. Additionally, our model is more parsimonious in terms of number of parameters in order to avoid the danger of overfitting. We report in the results section a comparison of our proposed model with this model, where we observe a more coherent pattern for increasing bitrate for our model.

Subject bias is also considered in the model developed in [18]. We highlight, however, that our goal is different since our model only considers the results of the subjective tests and does not attempt to establish a relationship between QoS and QoE as in [18].

III. Dataset

In this work, we use GamingVideoSET [19], an open source dataset of gaming video sequences containing subjective and objective quality assessment ratings. The dataset consists of twenty-four uncompressed raw gaming video sequences from twelve different games, each with a 30 second duration, 1080p resolution, and 30 fps frame rate. The games are selected to cover a wide range of genres and content complexity representative of real-world streaming applications such as Twitch.tv. Subjective assessment ratings in terms of mean opinion score (MOS) are available for 90 distorted sequences (stimuli) obtained by encoding six reference gaming video sequences as 15 multiple resolution-bitrate pairs, i.e., considering three resolutions (1080p, 720p and 480p) and five different bitrates per resolution, as shown in Table I. Subjective tests were conducted in line with the ITU-R BT.500 recommendation using the ACR methodology with a scale of 1 to 5 and a total of 25 valid test subjects.

To summarize, the dataset consists of $M = 2250$ scores given by $N = 25$ subjects for evaluating the quality of $K = 90$ distorted video sequences. The scores correspond are within the ordinal scale from 1 to 5. The six games considered are Counter Strike: Global Offensive (CSGO), FIFA 2017 (FIFA), H1Z1: Just Survive (H1Z1), Hearthstone (HSTO), League of Legends (LOL) and Project Cars (PCAR).

TABLE I: Compression levels by resolution

Resolution	Bitrate (Mbps)				
A: 1920 x 1080	0.6	0.75	1.2	2	4
B: 1280 x 720	0.5	0.6	1.2	2	4
C: 640 x 480	0.3	0.6	1.2	2	4

To show that the approach can be generalized to different datasets, we report results for the dataset in the first Appendix [4].

IV. Data analysis

The available observations constitute a completely balanced dataset of repeated measures, as each video has been evaluated by all individuals, with no missing data, and the data can be partitioned into N clusters of observations from the same individual.

For the graphical representation, we found the spaghetti plot and the image plot useful. The first plot allows studying the individual choices, while the second plot represents the distribution of the observed scores for each video. The aim of the spaghetti plot is to study the scoring behavior of individual subjects by representing the scores as piecewise linear functions of a quantitative variable. In our case, the opinion score as a function of the bitrate after compression gives us 25 piecewise lines for each game and resolution (one for each subject). To distinguish the lines produced by each subject, we added random noise into the data of the spaghetti plots and used colored lines. We plotted the curves by treating the bitrate with two alternative approaches, namely, as a quantitative variable and as an ordered qualitative variable, so that the rate levels are plotted as if they were equidistant. Since the individual trajectories are clearer in the latter approach, in particular for small rates of compression, we use the qualitative scale in the following two figures.

The scores for the full dataset are reported in Fig. 1, where each observed subject contributes one piecewise line with the same color for each plot. We notice several cases of locally decreasing trajectories. These patterns are inconsistent with the fact that improved compression cannot deteriorate the video quality. Before further exploring those erratic cases, notice from Fig. 1 that this behavior is common to all games and resolutions, and it also appears to be widespread among individuals.

Fig. 2 focuses on the score distributions by showing the image plots of the counts of the individual scores for each level of score and compression. The distributions are represented by the color intensity on the five vertical cells in each plot, whose frequencies sum to the number of individuals, N=25.

We analyzed spaghetti and image plots for hints regarding both the general pattern and the individual scoring behavior. It can be seen, for example, that the best resolutions A and B appear to be closer and better than C. In fact, the scores on the effect of compression span the full range in both A and B, while in resolution C, the impact produced by the highest bitrates is reduced, with none or few of the highest scores.

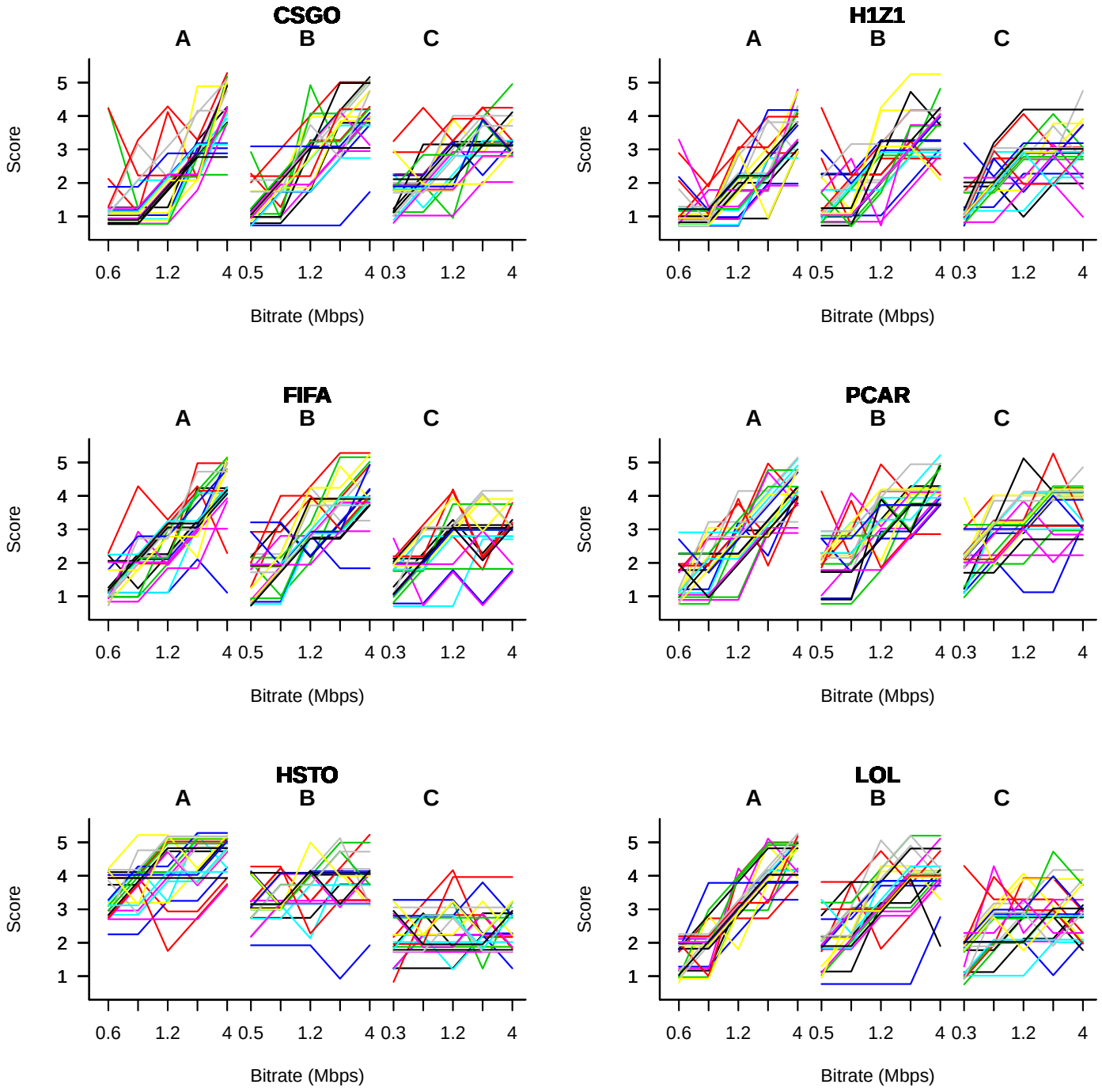


Fig. 1: Spaghetti plots of opinion scores vs. bitrate.

From both Fig. 1 and Fig. 2 it is clear that the game HSTO is different from the other games. With both resolutions A and B, the empirical frequencies are generally more stable and centered on higher scores than in the other games, with a mode of Y=3 at the lowest bitrate. In contrast, the scores are lower for HSTO when reproduced at resolution C. Finally, for HSTO, resolution A seems better than B, and B seems better than C (e.g., in terms of mode) for all compression levels, while this is not the case for the other games. Those patterns are confirmed in Fig. 3, showing the MOS trajectories when varying compression. For each game, the mean scores in cases of resolutions A, B and C are shown as connected with a piecewise line (blue, red and black, respectively). The difference between HSTO and the other five games is striking. In all cases except HSTO, the three lines produce some intersection, while the resolutions appear strictly ordered for HSTO.

A possible explanation for this result is that the content participates in the definition of the QoE itself so that the enjoyment in playing each game depends in different ways on the accurate reproduction of images, sounds, background details, dynamics, etc. Those elements serve different roles when the measured QoE is too different. In our case, while games such as FIFA and PCAR reproduce fullscreen rapid action, HSTO is a fast game in terms of occurring events. However, rather than movement, the action consists of localized objects including cards, balloons, arrows, etc., which appear or explode, flash or tremble. While the other games rely more on the synchronicity between reproduced events and a player's reactions, HSTO is based to a greater extent on turns of events and heavily pictorial backgrounds representing important, changing "elements" to be kept in mind during gameplay. We believe that those crucial components are highly impaired under resolution C, while in cases of resolutions A and B, those graphical objects are clearly distinguishable so that the QoE is low in C while remaining acceptable along the entire tested range of compression in cases of A and B.

A. Subject Consistency

Subject consistency is studied in prior work via repetition of tests with the same stimulus for the same subject [15]. Since we do not have repeated measurements in our dataset, we adopt another approach.

When the bitrate after compression increases, we expect an improved (or unchanged, in the case of bitrate increase below the JND) QoE so that the decreasing "steps" observed in Fig. 1 show inconsistent behaviors of the subjects. From Fig. 2, we can observe that the mode of the opinion scores (OS) often remains constant but never decreases. We also verified that the observed median is a nondecreasing function of the bitrate. This does not always occur for the MOS; instead, as shown in Fig. 3, we can observe several cases in which it is locally decreasing. More precisely, this occurs in six cases.

Fig. 3 also shows the 95% confidence intervals around the MOS. To distinguish the intervals under different resolutions, we slightly offset the points on the horizontal axis.

The question is now whether a locally decreasing MOS is due to a peculiar misbehavior of the mean operator or an actual incoherent performance of the panel of subjects. It is known, in fact, that the mean is not resistant to extreme observations because even a single outlier may unduly affect its value. To address this question, we employ a statistic that we call compression advantage. Assume that we sample two scores, Y_0 and Y_1, from their empirical distributions, where both scores refer to the same video but the latter is reproduced at a higher bitrate. We then expect that $P(Y_0 < Y_1) \geq P(Y_0 > Y_1)$, with the equality sign used only if the difference is imperceptible; that is, the *compression advantage* is:

$$A = P(Y_0 < Y_1) - P(Y_0 > Y_1) \geq 0. \tag{1}$$

Assuming that the two samples are independent, the probabilities in (1) are simply calculated. For example, denoting the empirical probability mass function of Y_0 and Y_1 by $p_1, p_2, \ldots, p_5$ and $q_1, q_2, \ldots, q_5$, respectively, then

$$P(Y_0 < Y_1) = \sum_{i=1}^{4} p_i \sum_{j=i+1}^{5} q_j.$$

Unlike the mean, A is based on the ordering of the data; hence, it is robust with respect to outliers. Thus, a negative compression advantage is strong evidence that the sample of subjects has expressed a preference for the PVS reproduced at the poorer compression rate. Out of the 72 comparisons between consecutive compression levels, we found seven cases of those distributional inconsistencies, as shown in Table II. In all but the PCAR case, where it is constant, the MOS is signaling this occurrence. Thus, on the one hand, we have confirmation that the observed cases of locally decreasing MOS are not due to single outliers. On the other hand, this result shows that erratic scoring is a nontrivial occurrence even in cases of moderately large samples.

Finally, the number of inconsistent scores per individual was found to be less than 20% in all cases, except for one subject (28%) who was consistent only 52 times out of 72. In Fig. 4, we show the performance of each subject in terms of consistency. We also plotted the same data in the form of distribution of inconsistent evaluations (not reported here due to space limitations). Based on the results, we believe that even the less accurate individual is not "bad enough" to be considered an outlier, as this seems to "correctly prolong" the performance of the less talented individuals in the population.

In conclusion, given the previous analysis in terms of both collective and individual behaviors, it appears that inconsistent scoring is not an isolated fact, but an inherent consequence of the variability of the evaluation process. For a deeper analysis, it seems simplistic to remove or correct the data since any full or partial omission might overlook population variability. Rather, we will show that the model presented in the next section is able to correct most of those inconsistencies without posing any constraint while acknowledging both the ordinal multinomial nature of the data and the existence of subject error in eliciting the scores.

V. The Common Slope Model

The ordinal multinomial regression model is a generalized linear model (GLM) used for regressing a multinomial variable

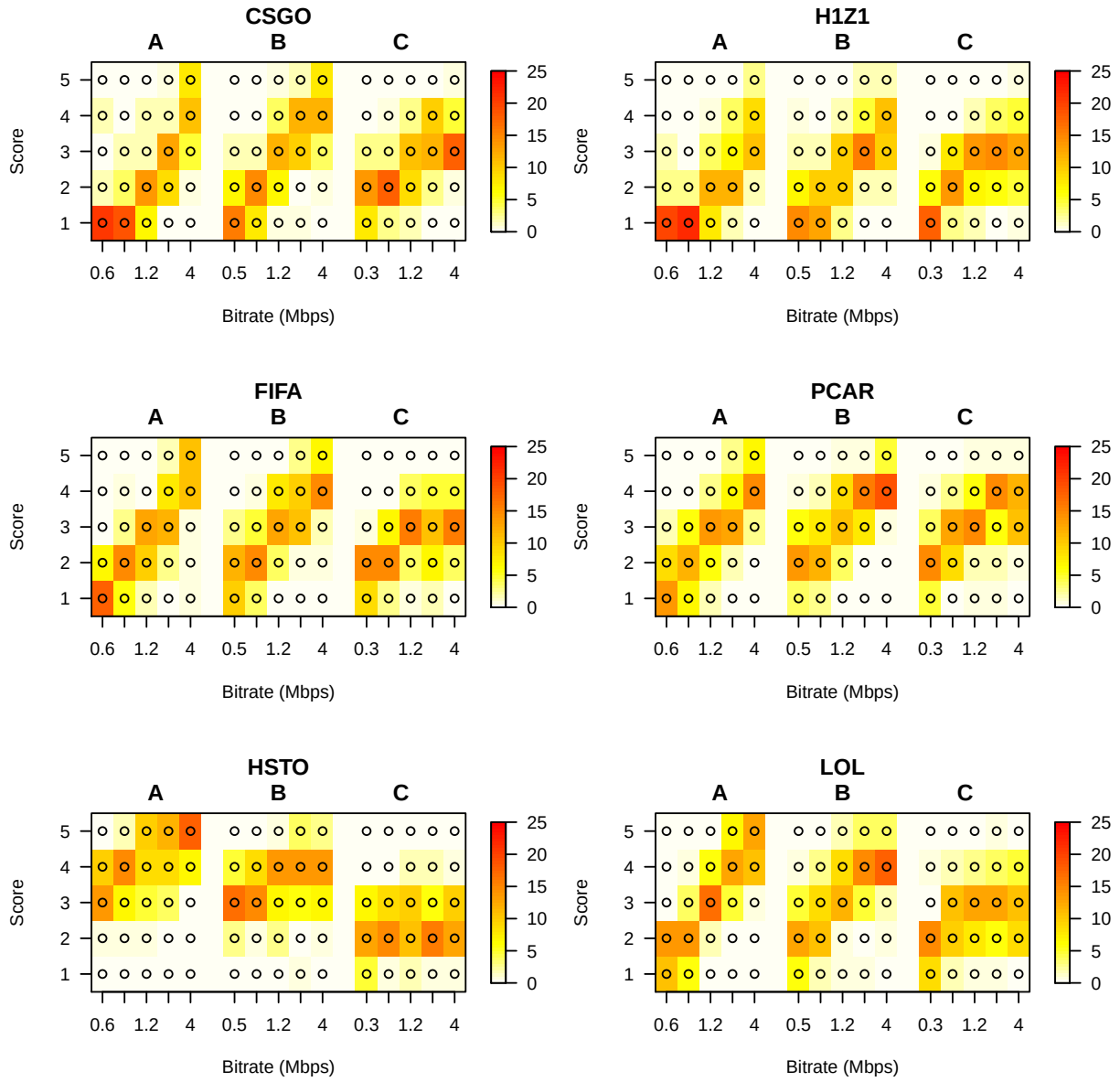


Fig. 2: Image plots of opinion score frequency distributions.

TABLE II: CASES OF OBSERVED DISTRIBUTIONAL INCONSISTENCIES: OBSERVED SCORE COUNTS, MODE, MEDIAN, MEAN (MOS) AND COMPRESSION ADVANTAGE (A)

Video	Resolution	Bitrate (Mbps)	OS Frequencies					Mo OS	Me OS	MOS	A
			1	2	3	4	5				
CSGO	C	2	0	3	12	10	0	3	3	3.28	-7.7%
		4	0	1	18	5	1	3	3	3.24	
H1Z1	A	0.6	20	3	2	0	0	1	1	1.28	-9.0%
		0.75	22	3	0	0	0	1	1	1.12	
FIFA	C	1.2	1	4	16	4	0	3	3	2.92	-10.4%
		2	2	7	11	5	0	3	3	2.76	
PCAR	C	2	1	2	6	15	1	4	3	3.52	-7.0%
		4	0	1	11	12	1	4	3	3.52	
HSTO	B	2	1	0	6	14	4	4	4	3.80	-5.9%
		4	0	1	7	14	3	4	4	3.76	
HSTO	C	1.2	2	11	10	2	0	3	2	2.48	-11.4%
		2	1	16	6	2	0	3	2	2.36	
LOL	C	2	1	6	13	4	1	3	3	2.92	-5.8%
		4	0	9	11	5	0	3	3	2.84	

$Y = 1, 2, .., k$ against a set of predictors $\boldsymbol{x} = (x_1, .., x_p)$. As in ordinary linear regression, the vector $\boldsymbol{x}$ may be formed by

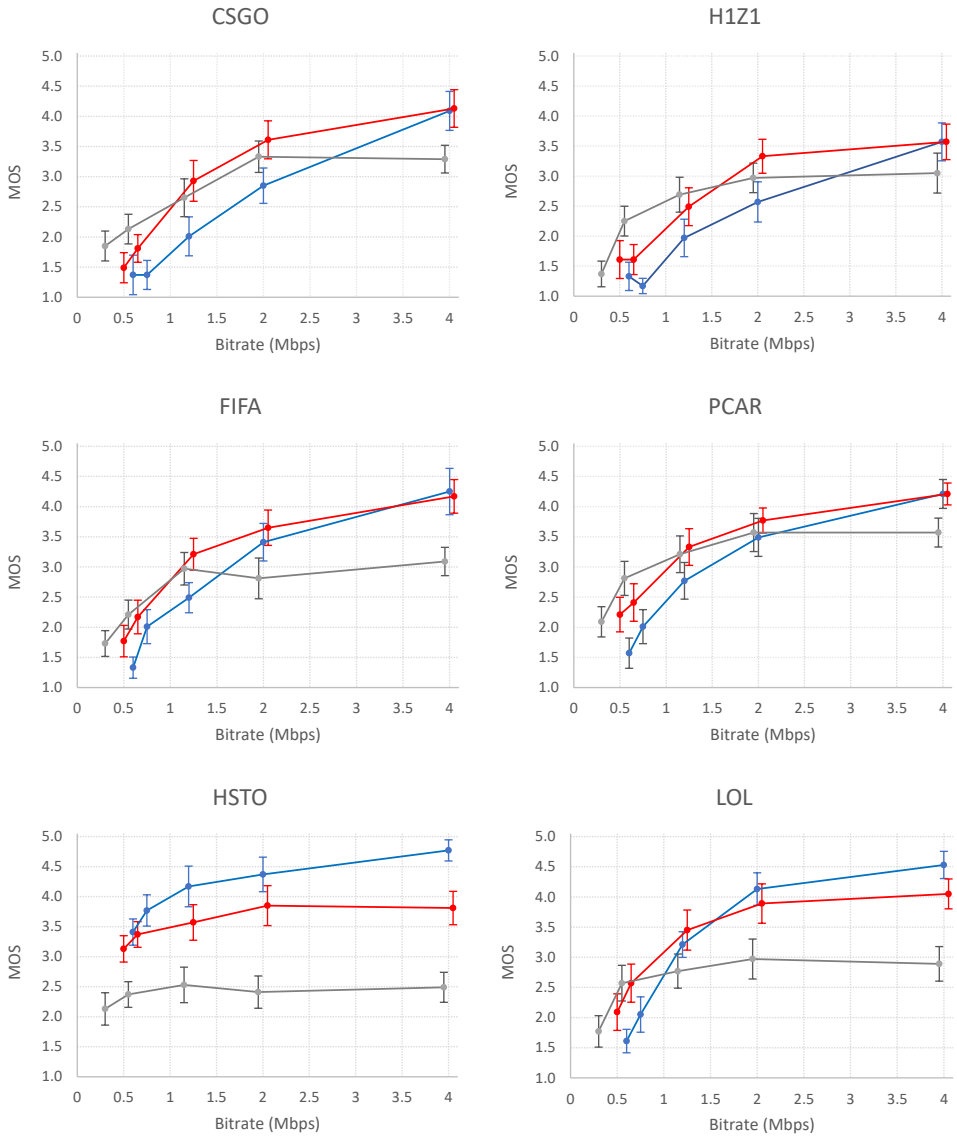


Fig. 3: MOS under resolutions A (blue line), B (red line) and C (gray line) with 95% confidence intervals.

either categorical "dummy" variables or quantitative variables (see, e.g., [20] and [21]). In the most general formulation, which we call the general ordinal multinomial (GOM) model, the distribution function (DF) of the score $F_j = P(Y \leq j)$ is modeled as

$$h(F_j) = \alpha_j - \mu_j \tag{2}$$

(with $j = 1, 2, ..., k-1$ and $F_k = 1$), where h is a nondecreasing function and μ_j is a linear combination of the predictors $\mu_j = \boldsymbol{\beta}_j' \boldsymbol{x}$. Note that the negative sign of μ_j makes it an increasing measure of the quality. In fact, since (2) grows with F_j, μ_j grows with $1 - F_j = P(Y > j)$.

The GOM is very flexible because it includes an intercept parameter α_j and a parameter vector of slopes $\boldsymbol{\beta}_j$ for each $j = 1, 2, .., k-1$. A more parsimonious model is the common slope model (CSM), which assumes a constant slope vector $\boldsymbol{\beta}$. This assumption is based on the *latent variable interpretation* of the scoring process. The idea is that the QoE of a video is actually perceived on a continuous scale so that the observed score is a version of an unobserved (latent) continuous variable that is discretized into k ordered classes.

Note that a thought experiment, such as this one, does not entail performance, or even the belief that the experiment actually occurred, but only the assumption that such an experiment is sufficiently realistic. We can imagine, for example, that the subject's latent quality evaluation is between 0 and 100 and that he/she must choose among "bad", "medium" and "good", or among $k = 5$ scores, as in our case. Hence, the interval 0-100 must be partitioned into k ordered intervals defined by $(k-1)$ cut-off points $\alpha_1, ..., \alpha_{k-1}$.

Thus, we assume that the subject's perceived quality is an unobservable variable

$$Z = \mu + \epsilon$$

, where μ is the average QoE in the latent scale and ϵ is a measurement error with distribution function $G(\epsilon)$. Note that no error is assumed in the discretization step, as it would be confounding with the measurement error ϵ. It follows that the

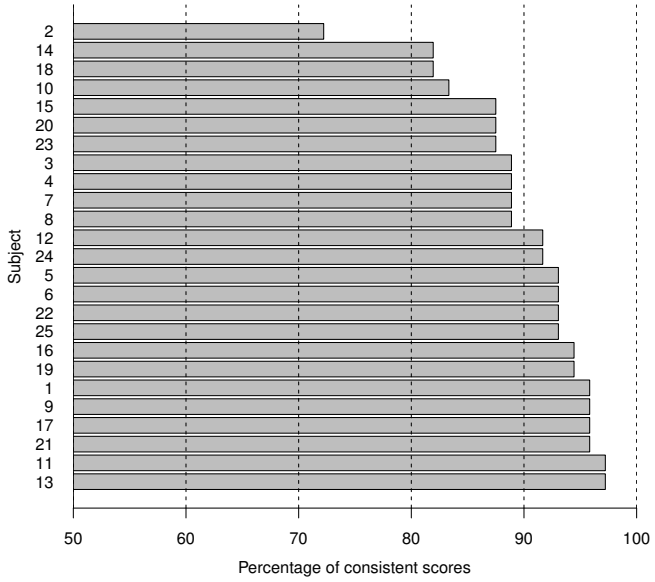


Fig. 4: Analysis of local inversions (inconsistencies). Bar diagram of subject id vs. percentage of consistent evaluations.

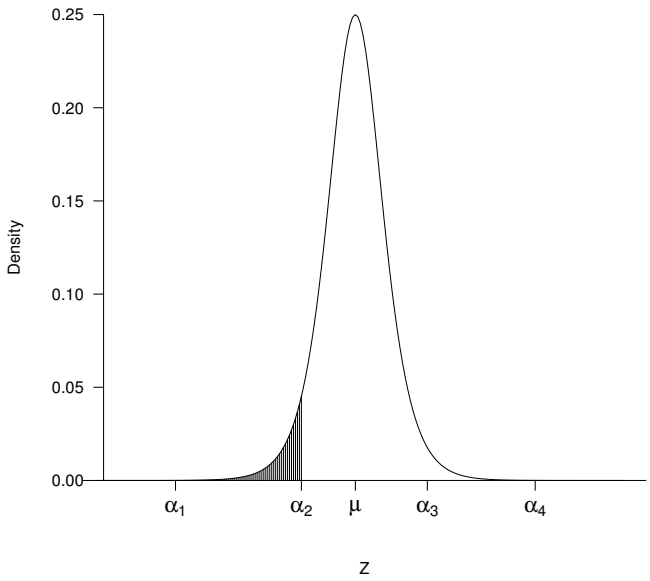


Fig. 5: Example of the latent variable interpretation. The probability density function of the latent score is centered on the average QoE (μ); the cut-off points (α's) identify five categories. The shaded area equals F_2.

DF of Z is $G(Z-\mu)$ and that the DF of Y is:

$$F_j = P(Z \leq \alpha_j) = G(\alpha_j - \mu), \qquad j = 1, 2, ...k-1. \quad (3)$$

Fig. 5 exemplifies the latent variable interpretation for a 5-level score Y. The cut-off points $\alpha_1, \ldots, \alpha_4$ are envisioned as unknown points on the latent scale. In this case, the average QoE, μ, is in the third interval, delimited by α_2 and α_3; however, since the individual's evaluations Z will distribute around that value, the observable score will be $Y = 3$ most of the time (but not always). For example, the probability of a lower score $Y \leq 2$ is the shaded area in the figure. Notice how easily this model may produce misplaced assessments, especially when μ is close to a cut-off point.

As a result, the DF of the OS Y is identified by the DF of the evaluation error G, the average QoE in the latent scale μ and the cut-off values α_j. In fact, from (3), and by setting $h = G^{-1}$, we obtain

$$h(F_j) = \alpha_j - \mu. \quad (4)$$

By comparing (4) with (2), we see that this apparently innocuous set of assumptions is enough to eliminate a large number of candidate models. Since the population average QoE on a continuous scale is a unique value, it cannot depend on j. Thus, the effect of the predictors will be measured by a single vector $\boldsymbol{\beta}$ of common slopes:

$$\mu = \boldsymbol{\beta}'\boldsymbol{x}. \quad (5)$$

A. Generalized Estimating Equations

Finally, we arrive at the fact that the $N = 2250$ observations have been repeatedly taken on the same 25 subjects, so the independence assumption is hardly valid. In other words, we expect to find some correlation between scores given by a subject on different videos, which may be inherited, for example, from the existence of subjective bias and other sources of systematic shifts, as discussed in the introduction.

We also notice that the subjects represent a random sample from the population of players and gaming viewers. However, rather than the QoE of the videos *conditional* to those particular subjects, we are interested in the average QoE with respect to the full population of subjects.

It is well known that the ordinary approaches are not efficient in this case. Thus, we follow the approach based on generalized estimating equations (GEEs) [22]. This is a semiparametric method that is appropriate when the focus of the analysis is on estimating population-averaged parameters, as in our case. Moreover, the semiparametric approach ensures the consistency of the estimates towards the population parameters, even when the covariance structure is misspecified.

Let Y_h, $h = 1 \ldots n$ denote the h-th score on a particular cluster of n observations that are supposedly correlated. Each response Y_h may be represented by a vector $D_h = (D_{h1}, D_{h2} \ldots, D_{hk-1})$ of indicator variables

$$D_{hj} = 1(Y_h \leq j)$$

, where $1(E) = 1$ if E is true and $1(E) = 0$ otherwise. In the ordinary approach, we model the average of those indicator variables: $F_j = E(D_{hj})$ $(j = 1, 2, \ldots k-1)$. The GEE approach introduces a further equation for modeling the correlation between scores that are inside the same cluster, while the no correlation assumption remains valid when they belong to different clusters.

The case of the ordinal multinomial distribution was developed by [23] via the so-called alternating logistic regression algorithm. Instead of modeling the correlation directly, which for the multinomial is constrained within limits that depend in a complicated way on the means of the data, the association between responses is modeled by the generalized log odds ratio (LOGOR). Consider two observations Y_h and Y_i belonging to the same cluster; then, for any pair of score values j and c, the LOGOR is defined as

$$LOGOR(D_{hj}, D_{ic}) = log\left(\frac{P(D_{hj}=1, D_{ic}=1)P(D_{hj}=0, D_{ic}=0)}{P(D_{hj}=1, D_{ic}=0)P(D_{hj}=0, D_{ic}=1)}\right). \quad (6)$$

Positive values of (6) indicate the tendency to associate similar scores and vice-versa; a negative $LOGOR$ indicates negative correlation. We thus followed the GEE approach by using the alternating logistic regression algorithm as implemented in SAS 9.4, where the $LOGOR$ is assumed to be constant for each pair $j, c = 1, 2, ..., k-1$.

B. The general approach to model selection

The model selection process is a sequence of trials and errors that can only be summarily described here. It requires not only the diligent evaluation of significance levels and other goodness-of-fit statistics but also the careful comparison of model results with prior knowledge. The components of a CSM with intragroup correlation are the following:

- the elements of the vector $\boldsymbol{x}$ in (5);
- the link function $h(F)$ or, equivalently, the probability distribution function of the evaluation error $G = h^{-1}$;
- the clusters of correlated observations.

The most critical step is to identify the best set of predictors forming the vector $\boldsymbol{x}$ in (5). In fact, $\boldsymbol{x}$ defines the analytic structure of the mean in the latent scale and therefore represents a fundamental step of model identification. On the other hand, the choice of the link function typically involves few alternatives, e.g., the inverse of symmetric distributions such as the logistic or the normal (called the *logit* link and the *probit* link, respectively). Moreover, as we found in our case studies and as is often noticed in the literature (e.g., [24], [20]), the final results are not sensibly affected by the link function. Similarly, we do not have many choices for the correlation structure. As we will show later, in the case of the game scores, we checked four alternative groupings, while for our second example, we had only two.

For the components of $\boldsymbol{x}$, on the other hand, the choice is among numerous options, it is highly critical and, unlike the other two steps, it cannot be performed in terms of goodness of fit only. In fact, this vector must represent both quantitative effects and *group effects*. For the latter, several alternatives are usually possible, and for the former, there are many more choices. In fact, the quantitative effects may be described by the original, untransformed predictor but also by one or more transformations of the original variable.

In the game data, for example, we used categorical variables for testing alternative groupings of games, resolutions and compression levels. We also tested the use of the bitrate as a continuous variable R (for example). It is obvious that the choice of which and how many transforms of R are needed is potentially unlimited. A flexible and parsimonious approach is to consider fractional polynomials (see, e.g., [25] and [26] for recent applications). More precisely, we evaluated the opportunity to introduce the powers R^v, for $v = \pm 0.5, \pm 1, \ldots, \pm 5$. We also tested the natural logarithm $log(R)$ and the exponential transforms $exp(-R)$ and $exp(R)$. A convenient algorithm is the "stepwise" procedure implemented in the SAS "proc logistic" routine, based on significance testing for inclusion and exclusion of the variables, which consents to identify the most valuable predictors.

Once a few candidate formulations of (5) have been selected, the resulting models can be compared by trading off a model's parsimony with goodness of fit. In fact, the goodness of the fit can always be improved by increasing the number of parameters, which may cause overfitting. To avoid this issue, the classical GLM literature offers two main criteria that are valid in the case of noncorrelated observations. The Akaike information criterion (AIC), proposed by [27], is an estimate of the Kullback-Leibler divergence between the current model and the true model. On the other hand, the Bayesian information criterion (BIC), developed by [28], aims to evaluate the posterior probability of the model. Both criteria aim to minimize the negative twice likelihood, plus a penalty that increases with the number of parameters. Since the penalty increases with the model's complexity, both criteria realize, as required, a compromise between simplicity and goodness of fit.

For a candidate model to be definitely accepted, we must check the common slope assumption, the link function and the correlation structure. Although a rigid procedure is not advisable, we give a schematic representation of the full approach in Fig. 6. The *select predictors* box represents the process of identifying the best model for the mean in the latent scale by first identifying a few candidate models (i.e., by using the stepwise procedure) and then comparing them via the AIC and BIC criteria. This process may be done under the independence assumption and the default (logit) link. Alternative links can then be used to determine whether there are noticeable differences; the common slope assumption can be tested, and finally, the best correlation structure can be chosen. For the latter, the only available criterion that corresponds to the AIC is the so-called quasi-likelihood information criterion (QIC) proposed by [29].

All of those analyses and comparisons are essentially iterative and characterize the *model identification* step. Finally, the selected model can be checked by using tests of fit and subjective judgments against independent knowledge, such as, e.g., the expectation that the population average score is a smooth and nondecreasing function of the bitrate.

VI. GAME DATA: RESULTS

By using the stepwise procedure, we eventually found that no grouping of games and resolution is worthwhile, while the effect of compression can be based on the log transform

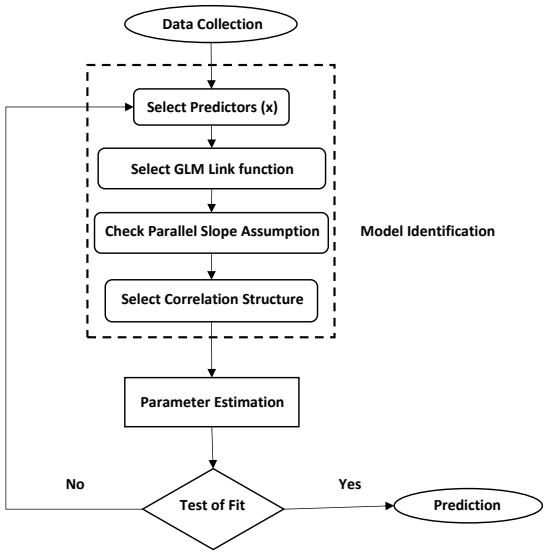


Fig. 6: Block diagram of the adopted approach.

and the square root of the bitrate. As a result, the number of linearly independent slope parameters is 26, so by including the intercepts, the proposed model reaches a total of 30 parameters.

Table III compares the selected quantitative compression model (labeled as QCM) against three noteworthy alternatives under the independence assumption. The general ordinal multinomial (GOM) model in its full parametrization (indicated as FGOM) is the unequal slopes model (2) with the maximum number of parameters. More specifically, we use four parameters for each of the 90 PVSs. As is well known, the maximum likelihood method in the case of the multinomial distribution is such that the probabilities of each score are then estimated by the observed relative frequencies. Thus, the estimated distributions exactly replicate the empirical distributions and, with it, any other observed statistic, including the MOS.

The other models are all based on the common slope assumption. In the first CSM (Cat 7), the compression level is treated as a categorical variable so that each one of the seven adopted bitrate levels ($0.3, 0.5, 0.6, 0.75, 1.2, 2$ and 4 Mbps) corresponds to a different parameter. Similarly, in the second CSM (Cat 5), the compression is also categorical but uses the ranks (from 1st to 5th), i.e., the ordinal values of the five compression levels observed for each resolution. Both categorical models include all two-way interactions between game, resolution and compression, while the three-way interaction was found to be insignificant.

A further test provided by SAS regards the common slope assumption. It is known that the test is rather liberal (see [24]) because it tends to reject the assumption too frequently. In our case, the test rejects the hypothesis for Cat 7 but accepts the CS assumption for both Cat 5 and QCM.

All asymptotic chi-square tests based on the likelihood ratio against the full model favor the simpler model, as can be easily determined from Table III by noticing that the loss in the twice log-likelihood is lower than its expected value, which is the difference between the number of parameters. Since the QCM can be considered as nested into Cat 7, the same test gives a chi-square value of 40 with 33 degrees of freedom (p-value 18.7%). Thus, the QCM can be preferred in terms of significance testing.

TABLE III: PROPOSED MODEL WITH QUANTITATIVE COMPRESSION SPECIFICATION (QCM) COMPARED TO THREE ALTERNATIVES IN TERMS OF GOODNESS-OF-FIT STATISTICS

Model	No Parameters	-2 log L	AIC	BIC
FGOM	360	4434	5154	7213
Cat 7	63	4696	4822	5182
Cat 5	53	4713	4819	5122
QCM	30	4736	4796	4968

From Table III, it is also clear that both the AIC and BIC criteria suggest selecting the model with quantitative compression.

We then investigated the candidate models by using the GEE approach for allowing correlated observations (and using the SAS proc GEE algorithm). For each model, the QIC can be used to verify that the correlated structure realizes an improvement with respect to the uncorrelated structure. With the same criterion, candidate models can be compared under the assumption of correlated observations. In our experience (e.g., in both of our case studies), the analytical form for μ chosen under the independence assumption appears to be the best model, even in these cases.

In particular, for the game score data, we considered the following structures:

(a) 25 clusters: one for each subject, where each cluster is formed by 90 observations;
(b) 75 clusters: one for each combination of subject and resolution, where each cluster is formed by 30 observations;
(c) 150 clusters: one for each combination of subject and game, where each cluster is formed by 15 observations;
(d) 450 clusters: one for each combination of subject, game and resolution, where each cluster is formed by 5 observations.

Table IV shows the values of the QIC statistic for the proposed model by assuming either independence or correlation inside each clustering structure. We can observe that in all cases the QIC improved by assuming the correlation structure. Moreover, since the minimum is reached in case (c), we present the corresponding results in the following subsection.

As noticed before, for the independent case and in results not reported here for the correlated case, the QIC statistic of the other candidate models (FGOM, Cat 7, Cat 5) is always higher than the corresponding value in Table IV. In particular, for the FGOM, which corresponds to adopting the observed frequencies and the MOS, the best structure is (d), with assumed independence and a QIC equal to 4844.

It is also worth mentioning that the estimated LOGOR indicates a moderate positive association in all instances. The appropriateness of the QCM was finally validated by the

Hosmer-Lemeshow goodness-of-fit test for the multinomial distribution (70% p-value).

TABLE IV: QIC STATISTIC VALUES OF THE QCM UNDER EACH CLUSTERING STRUCTURE, ASSUMING INDEPENDENCE AND CORRELATION INSIDE CLUSTERS

Structure	Independent	Correlated	LOGOR
(a)	4865	4755	1.09
(b)	4849	4748	1.01
(c)	4844	4745	0.94
(d)	4834	4750	0.87

A. The Estimated Population Mean Opinion Score (EPMOS)

In summary, with $l = 1, 2, \ldots, 6$ and $m = 1, 2, 3$ denoting the subscripts for each game and resolution, respectively, the selected model has the form (3), with G equal to the (standardized) logistic DF: $G(z) = 1/(1 + exp(-z))$ and:

$$\mu = \eta_l + \eta_m + \eta_{lm} + \gamma\sqrt{R} + (\delta + \delta_l + \delta_m)\, ln(R) \quad (7)$$

where R is the bitrate after compression in Mbps. In the above, for simplicity, we adopted a slight abuse of notation by indicating different parameters with the same letter but varying subscripts. Hence, e.g., η_l is the effect of game l, η_m is the effect of resolution m, and η_{lm} is the effect of the interaction between the two.

Table V shows the parameters' estimates. The parameterization follows the standard notation in linear and generalized linear models, with the game PCAR and the resolution C as the *base levels*, so that the corresponding coefficients are set to zero. Thus, for each fixed game, resolution and compression level, the formula (7) becomes simply

$$\mu = a_{lm} + b\sqrt{R} + c_{lm}\, ln(R) \quad (8)$$

where $a_{lm} = \eta_l + \eta_m + \eta_{lm}$, $b = \gamma$ and $c_{lm} = \delta + \delta_l + \delta_m$.

For example, in the case of the CSGO sequence ($l = 1$) at resolution A ($m = 1$), we obtain $a_{1,1} = -1.12 - 1.88 - 0.66 = -3.66$, $b = -4.23$ and $c_{1,1} = 3.67 + 0.53 + 2.81 = 7.01$, while the same PVS at resolution C ($m = 3$) has $a_{1,3} = -1.12$, $b = -4.23$ and $c_{1,3} = 3.67 + 0.53 = 4.2$.

We reiterate that μ is the average QoE in the latent scale, so the probability that the observed score falls into each of the five categories depends on its position with respect to the cut-off points α_j. As shown in Table V, those limits are almost perfectly equispaced. Thus, the latent variable interpretation suggests that the subjects appear to divide the range of their perceived quality into equal intervals before eliciting their ordinal score. This finding confirms a common view in research as, e.g., discussed in [6].

In Fig. 7, we present a comprehensive comparison between the MOS and EPMOS. Here, the EPMOS is continuously calculated over the interval $0 - 4$ Mbps owing to the quantitative specification of the bitrate.

In the figure, we also report the results obtained when applying the model in [16], also tested in [17], to our dataset.

From Fig. 7, we notice that the differences are not severe. In fact, the MOS appears to require only relatively small corrections, but they are very interesting. The model's estimates outline a behavior that is typical of technological improvements, with an initial burst of QoE followed by a concave shape that indicates a pattern of diminishing returns. For all three resolutions, the concavity is rather dominant since the quality improvements start to decrease from approximately 0.5 Mbps onward. Moreover, the diminishing return pattern is clearly stronger, and somewhat anticipated, when the resolution is lower. For the HSTO video under resolution C, the model identifies a decreasing pattern from 1200 Mpbs onwards. However, we may cast some doubts upon this solution. Note that from Table III, this video exhibits the strongest inconsistency (e.g., the lowest A), which refers to the comparison between 1200 Mpbs and 2000 Mbps. Moreover, when comparing the OS at 1200 Mpbs vs. 4000 Mbps, we have 5 subjects scoring higher for 1200 Mbps than for 4000 Mbps against 4 subjects scoring higher for a higher bitrate, while the remaining 16 give the same score to the two videos. Thus, these repeated inconsistencies at 2000 Mbps and 4000 Mbps deceived the model by triggering the identification of a decreasing quality pattern, but in fact, the hypothesis that the quality is relatively constant after 1.2 cannot be refuted with these data.

Apart from the mentioned HSTO case, the other six MOS

TABLE V: PARAMETER ESTIMATES AND STANDARD ERRORS OF THE PROPOSED MODEL

Parameter			Estimate	Standard Error
α_1			-8.26	0.72
α_2			-5.76	0.64
α_3			-3.37	0.61
α_4			-0.71	0.63
Game	CSGO		-1.12	0.37
Game	FIFA		-1.22	0.36
Game	H1Z1		-1.55	0.37
Game	HSTO		-1.52	0.37
Game	LOL		-1.18	0.39
Resolution	A		-1.88	0.29
Resolution	B		-0.24	0.22
Game*Resolution	CSGO	A	-0.66	0.35
Game*Resolution	CSGO	B	0.00	0.31
Game*Resolution	FIFA	A	0.94	0.35
Game*Resolution	FIFA	B	0.71	0.28
Game*Resolution	H1Z1	A	-0.74	0.40
Game*Resolution	H1Z1	B	-0.30	0.29
Game*Resolution	HSTO	A	5.43	0.50
Game*Resolution	HSTO	B	2.78	0.38
Game*Resolution	LOL	A	2.00	0.39
Game*Resolution	LOL	B	1.30	0.34
sqrtComp			-4.23	0.49
logComp			3.67	0.33
logComp*Game	CSGO		0.53	0.24
logComp*Game	FIFA		0.19	0.23
logComp*Game	H1Z1		0.25	0.25
logComp*Game	HSTO		-1.43	0.20
logComp*Game	LOL		-0.06	0.23
logComp*Resolution	A		2.81	0.17
logComp*Resolution	B		1.48	0.12

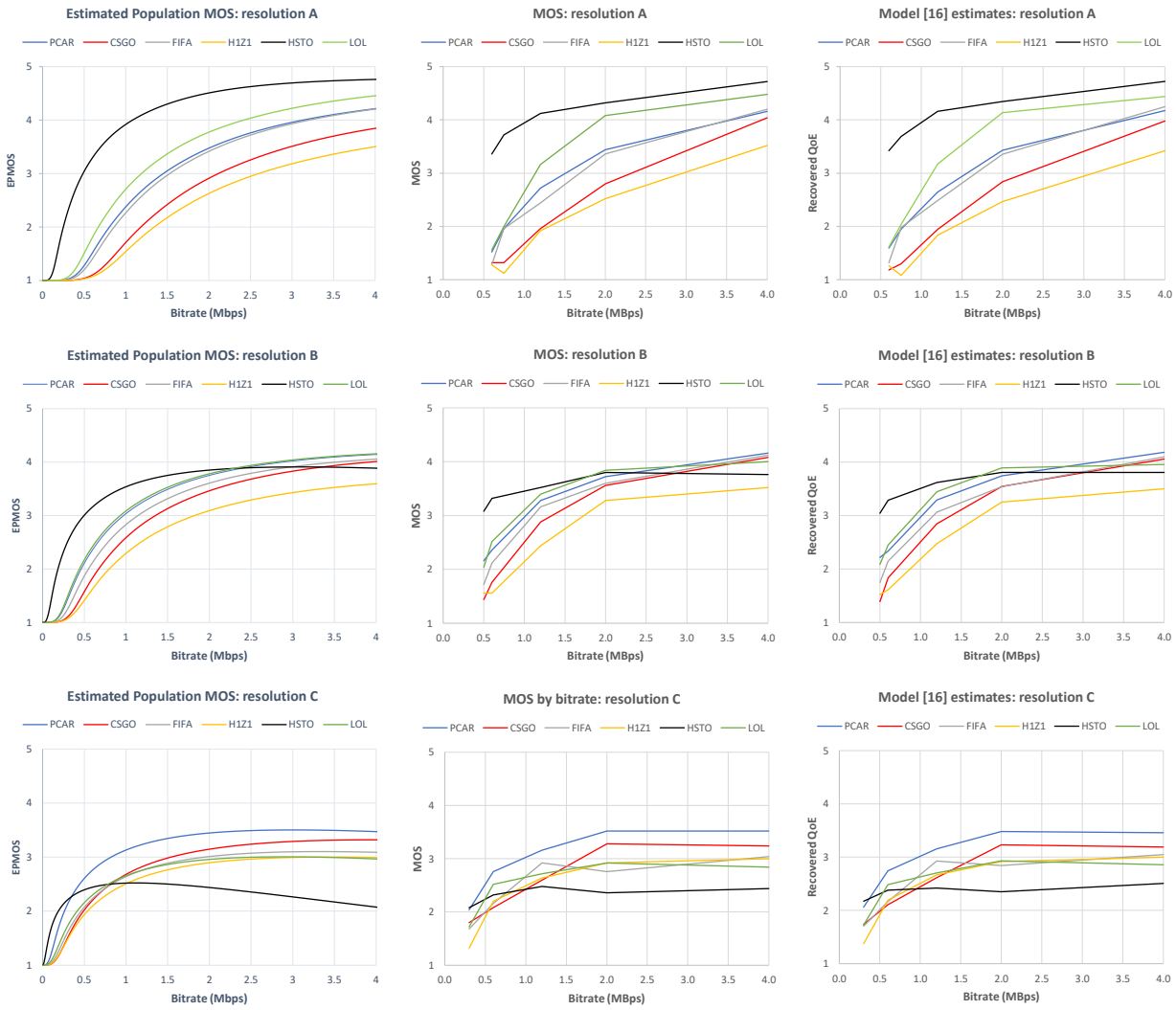


Fig. 7: Left: EPMOS obtained via our model; Center: observed MOS; Right: quality estimated with the model in [16]. Each row reports the results for a different resolution.

inconsistencies are all corrected into an increasing shape by the EPMOS, and no further inconsistencies are introduced. This result shows the effect of the borrowing strength between sample patterns.

For the model in [16], the recovered quality scores are very similar to the MOS values, and we can observe the same inconsistencies present in the MOS scores.

Finally, from Fig. 3 and Fig. 8, we can compare the 95% confidence intervals for the MOS and EPMOS, respectively. For the construction of the intervals of the EPMOS, we followed the delta method by using the estimated covariance matrix of the regression parameters obtained from the SAS output. Since to the best of our knowledge there is no software that implements this calculation for the multinomial case, we give the theoretical details in the Appendix.

The confidence intervals turn out to be nearly always smaller than the intervals corresponding to the MOS. The MOS, in fact, uses only the 25 observations pertaining to each configuration of game, resolution and compression, while the selected model also uses the rest of the data. In other words, assuming that the model is correct, the data cooperate beyond each particular configuration; in doing so, the uncertainty is reduced.

VII. Conclusion

In this paper, we argue that when scores are collected for a set of PVSs with similar contents, then the MOS can be improved by multinomial modeling. An appropriate GLM can, in fact, spot the common patterns to more efficiently estimate the multinomial distributions involved and the corresponding average score in the target population. The MOS itself corresponds to a particular GLM in which the data for each PVS are treated separately. This is accomplished by using the maximum number of parameters, so in a sense, the MOS corresponds to the most complicated model. However, this model can be

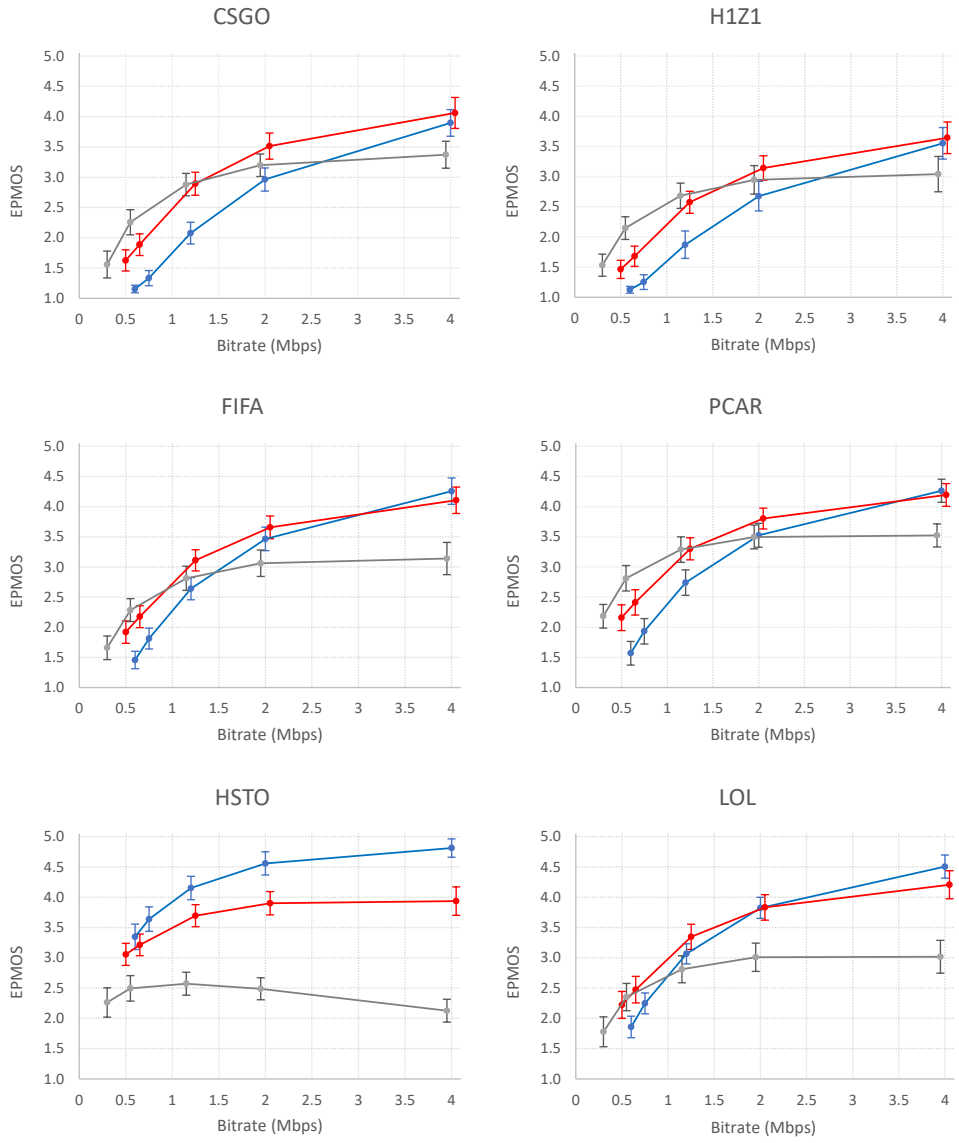


Fig. 8: EPMOS and 95% confidence intervals at resolutions A (blue), B (red) and C (gray). The corresponding subjects' MOS results are reported in Fig. 3.

compared and replaced by alternative models according to standard criteria for model selection.

We demonstrated this approach on a dataset of six gaming videos with various resolutions and compression ratios. To show that the approach can be generalized, we report the results for another dataset in Appendix 1. In both cases, a multinomial common slope model was ultimately selected, which is proven to be preferable to the "fully parameterized" model corresponding to the MOS estimate. A sign of this improvement is the reduction in the number of inconsistencies with respect to compression, and in the second dataset, a "saturation" effect, often assumed in literature, was also clearly identified.

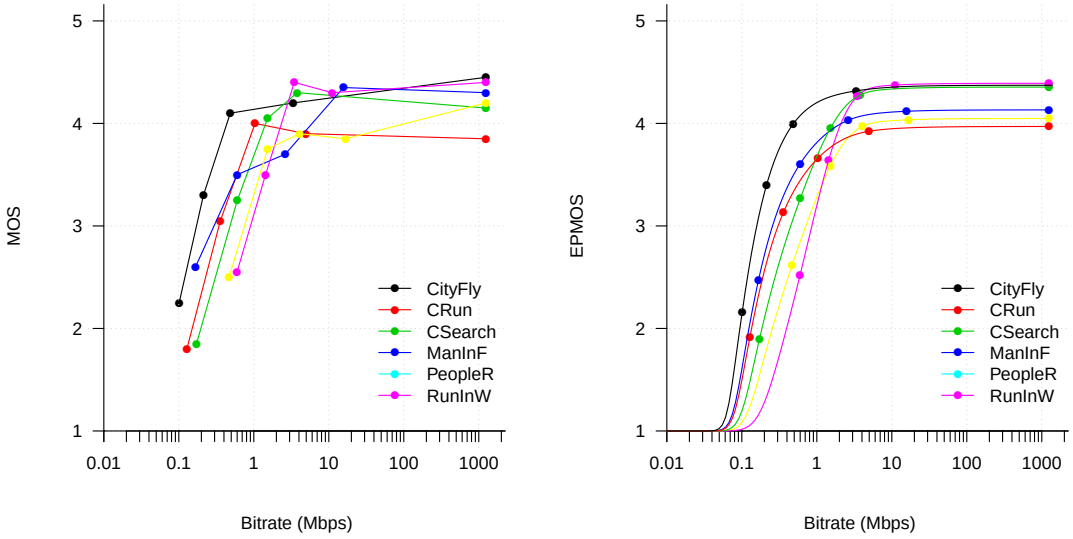
For the first dataset, it is worth noting that the other two alternative models discussed above (Cat 7 and Cat 5) also achieve a more coherent pattern than the MOS and give similar results. Moreover, and apart from the HSTO(C) case, the model average QoE exhibits a reasonable behavior of quality gain per bps increase. As another effect of the borrowing strength between sample patterns, we see a regular law of decreasing returns over most of the compression range and a diminishing uncertainty around our estimates.

Appendix 1. Second Case Study: Results

We report here the results obtained for a different dataset, composed of videos with natural content. The dataset [4] is formed by six video sequences, denominated "City Fly", "Costumes Run", "Costumes Searching", "Man In Fountain", "People Run" and "People In Woods". These videos were evaluated by 20 subjects in the uncompressed format and with four levels of compression, for a total of $N = 600$ observations.

As before, let $l = 1, 2, \ldots, 6$ denote the subscripts of the PVSs and R denote the bitrate in Mbps. After the identification step, we selected a model of the form (3) with G equal (again) to the standardized logistic DF and:

$$\mu = \eta_l + \gamma \frac{1}{R} + \delta_l \ \ exp\{-R\}. \tag{9}$$



(a) (b)

Fig. 9: MOS and EPMOS for the natural video dataset [4].

TABLE VI: QIC STATISTIC VALUES OF THE SELECTED QUANTITATIVE COMPRESSION MODEL FOR THE SECOND DATASET (QCM2) VERSUS THE MODEL BASED ON THE PREDICTORS SELECTED FOR THE FIRST DATASET (QCM) UNDER EACH CLUSTERING STRUCTURE, ASSUMING INDEPENDENCE AND CORRELATION INSIDE CLUSTERS

Model	Structure	Independent	Correlated	LOGOR
QCM2	(a)	1450	1413	0.60
	(b)	1447	1414	0.58
QCM	(a)	1458	1420	0.58
	(b)	1455	1419	0.55

Again, we assured that the model performs better than the fully parameterized model (i.e., the MOS) either in terms of AIC (and BIC) and QIC, and the parallel slope assumption is acceptable. In this case, the videos have the same resolution, and therefore the analytical form (9) is simpler. Moreover, we only have two possible correlation structures:

(a) 20 clusters: one for each subject, where each cluster is formed by 30 observations;
(b) 120 clusters: one for each combination of subject and PVS, where each cluster is formed by 5 observations.

Note that, instead of the square root and the logarithm of the bitrate, the stepwise selection procedure identified the reciprocal and the negative exponential as the best predictors for these data. However, as shown in Table VI, it is worth noting that, by using the former pair of predictors, the resulting model (e.g., QCM) is only marginally worse than the selected model (QCM2). Table VI also demonstrates that the correlation assumption is preferable for both models.

On the other hand, a subtle but significant difference between QCM2 and QCM is the saturation effect for a sufficiently high bitrate. This is shown in Fig. 9. The figure compares, with the bitrate on the logarithmic scale, the MOS with EPMOS as estimated by our best model (QCM2). From Fig. 9 (b), it is evident that the estimated population average is essentially constant in all six PVSs within the range of approximately $4-5$ Mbps. The same plot for the QCM (not shown here) shows that the points at observed bitrates were fitted almost identically as the QCM2 with the price of a slight concavity on the right side. It follows that the form based on the reciprocal and negative exponential of the bitrate seems more appropriate to reproduce a saturation effect of the QoE vs. compression.

Interestingly, the presence of the exponential form to represent the saturation effect of the QoE on a continuous scale was also suggested in the literature ([18] and references therein).

Finally, note that all inconsistent concavities of the MOS are eliminated by the model estimates so that the EPMOS of each PVS always appears as a smooth nondecreasing function of the bitrate.

APPENDIX 2. SAMPLE VARIANCE CALCULATION OF THE POPULATION MEAN ESTIMATOR

Let $\boldsymbol{b}$ denote a vector of random variables whose mean vector and variance matrix are estimated by $\hat{\boldsymbol{b}}$ and V, respectively. The delta method (see, e.g., [30], Chapter 10.5) uses Taylor's approximation for calculating the variance of a nonlinear transformation $\mu = g(\boldsymbol{b})$. Let $g'(\boldsymbol{b})$ be the column vector of the partial derivatives of μ with respect to $\boldsymbol{b}$. The delta method provides the formula

$$Var(g(\boldsymbol{b})) \approx g'(\hat{\boldsymbol{b}})^T V g'(\hat{\boldsymbol{b}})$$

where T indicates the transpose. In our case, $\boldsymbol{b}$ is the regression parameter estimator formed by 4 intercepts and 26 common slopes.

On the other hand, any given mean M is a linear function of the four values of the distribution function

$$M = F_1 + 2(F_2 - F_1) + 3(F_3 - F_2) + 4(F_4 - F_3) + 5(1 - F_4) = 5 - S$$

where $S = F_1 + \ldots + F_4$, so that $Var(M) = Var(S)$. Finally, since $G(z) = 1/(1 + exp(-z))$, we have

$$S = g(\boldsymbol{b}) = \sum_{j=1}^{4} \frac{1}{1 + e^{-\alpha_j + \boldsymbol{\beta}^T \boldsymbol{x}}}.$$

Therefore,

$$g'(\boldsymbol{b})^T = \left(\left(\frac{\partial g(\boldsymbol{b})}{\partial \boldsymbol{\alpha}} \right)^T, \left(\frac{\partial g(\boldsymbol{b})}{\partial \boldsymbol{\beta}} \right)^T \right)$$

has components

$$\frac{\partial g(\boldsymbol{b})}{\partial \alpha_h} = \frac{e^{-\alpha_h + \beta^T x}}{(1 + e^{-\alpha_h + \beta^T x})^2}$$

for $h = 1, 2, 3, 4$, followed by

$$\frac{\partial g(\boldsymbol{b})}{\partial \beta_l} = -x_l \sum_{j=i}^{4} \frac{e^{-\alpha_j + \boldsymbol{\beta}^T \boldsymbol{x}}}{(1 + e^{-\alpha_j + \boldsymbol{\beta}^T \boldsymbol{x}})^2}$$

for $l = 1, 2, .., 26$.

ACKNOWLEDGMENT

This work was partially supported by the European Union's Horizon 2020 research and innovation program, under grant agreement No. 643072 (QoE-Net), and by EPSRC Grant EP/P022715/1 (IoSiRe). We would like to thank Zhi Li and Christos Bampis for having published the code associated with their model (which we used for the comparison with our model in Fig. 7).

REFERENCES

[1] ITU-T Rec. P.910, "Subjective video quality assessment methods for multimedia applications," April 2008.

[2] ITU-T Rec. BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," Jan 2012.

[3] N. Barman, S. Zadtootaghaj, S. Schmidt, M. G. Martini, and S. Möller, "An objective and subjective quality assessment study of passive gaming video streaming," *International Journal of Network Management*, vol. e2054, 2018.

[4] S. Bosse, K. Brunnström, S. Arndt, M. G. Martini, N. Ramzan, and U. Engelke, "A common framework for the evaluation of psychophysiological visual quality assessment," *Quality and User Experience*, vol. 4, no. 3, 2019.

[5] R. Likert, "A Technique for the Measurement of Attitudes," *Archives of Psychology*, vol. 140, no. 55, 1932.

[6] S. Jamieson, "Likert scales: how to (ab)use them," *Medical Education*, vol. 38, no. 12, pp. 1217–1218, 2004.

[7] G. Norman, "Likert scales, levels of measurement and the "laws" of statistics," *Advances in health sciences education*, vol. 15, no. 5, pp. 625–632, 2010.

[8] M. Narwaria, L. Krasula, and P. Le Callet, "Data analysis in multimedia quality assessment: Revisiting the statistical tests," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2063–2072, 2018.

[9] K. Brunnström and M. Barkowsky, "Statistical quality of experience analysis: on planning the sample size and statistical significance testing," *Journal of Electronic Imaging*, vol. 27, no. 5, pp. 1–11, 2018.

[10] M. Seufert, "Fundamental advantages of considering quality of experience distributions over mean opinion scores," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, (Berlin, Germany), IEEE, June 2019.

[11] A. Ostaszewska and S. Żebrowska-Łucyk, "The method of increasing the accuracy of mean opinion score estimation in subjective quality evaluation," in *Wearable and Autonomous Biomedical Devices and Systems for Smart Environment*, pp. 315–329, Springer, 2010.

[12] J. Mandel, "The validation of measurement through interlaboratory studies," *Chemometrics and intelligent laboratory systems*, vol. 11, no. 2, pp. 109–119, 1991.

[13] A. M. Van Dijk, J.-B. Martens, and A. B. Watson, "Quality assessment of coded images using numerical category scaling," in *Advanced Image and Video Communications and Storage Technologies*, vol. 2451, pp. 90–102, International Society for Optics and Photonics, 1995.

[14] T. Hoβfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!," in *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on*, (Mechelen, Belgium), pp. 131–136, IEEE, 2011.

[15] L. Janowski and M. Pinson, "The accuracy of subjects in a quality experiment: A theoretical subject model," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2210–2224, 2015.

[16] Z. Li and C. G. Bampis, "Recover subjective quality scores from noisy measurements," in *Data Compression Conference (DCC), 2017*, (Snowbird, UT, USA), pp. 52–61, IEEE, 2017.

[17] J. Li and P. Le Callet, "Improving the discriminability of standard subjective quality assessment methods: a case study," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, (Sardinia, Italy), pp. 1–3, May 2018.

[18] H.-S. Chang, C.-F. Hsu, T. Hoßfeld, and K.-T. Chen, "Active learning for crowdsourced QoE modeling," *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3337–3352, 2018.

[19] N. Barman, S. Zadtootaghaj, S. Schmidt, M. G. Martini, and S. Möller, "GamingVideoSET: A Dataset for Gaming Video Streaming Applications," in *2018 16th Annual Workshop on Network and Systems Support for Games (NetGames)*, (Amsterdam, Netherlands), pp. 1–6, June 2018.

[20] P. McCullagh and J. A. Nelder, *Generalized linear models*. London Chapman and Hall, 2^{nd} ed., 1989.

[21] A. Agresti, *An introduction to categorical data analysis*. Wiley, 3^{rd} ed., Nov 2018.

[22] K.-Y. Laing and S. L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, vol. 73, no. 1, pp. 13–22, 1986.

[23] P. Heagerty and S. Zeger, "Marginal regression models for clustered ordinal measurements," *Journal of the American Statistical Association*, vol. 91, pp. 1024–1036, Sept 1996.

[24] M. Stokes, C. Davis, and G. Koch, *Categorical Data Analysis Using the Sas®System*. SAS Publishing, 2^{nd} ed., 2000.

[25] P. Royston and D. Altman, "Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling," *Insurance Mathematics and Economics*, vol. 16, no. 2, pp. 165–166, 1995.

[26] J. Cui, N. de Klerk, M. Abramson, A. Del Monaco, G. Benke, M. Dennekamp, A. W. Musk, and M. Sim, "Fractional Polynomials and Model Selection in Generalized Estimating Equations Analysis, with an Application to a Longitudinal Epidemiologic Study in Australia," *American Journal of Epidemiology*, vol. 169, no. 1, pp. 113–121, 2009.

[27] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Second International Symposium on Information Theory*, (Budapest, Hungary), pp. 267–281, Akadémiai Kiado, 1973.

[28] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[29] W. Pan, "Akaike's Information Criterion in Generalized Estimating Equations," *Biometrics*, vol. 57, no. 1, pp. 120–125, 2001.

[30] M. G. Kendall, A. Stuart, and J. K. Ord, *Kendall's Advanced Theory of Statistics*. New York, USA: Oxford University Press, 5^{th} ed., 1987.