Robust Evaluation for Quality of Experience in Crowdsourcing*

Qianqian Xu
BICMR, Peking University & University of Chinese Academy of Sciences, Beijing 100871, China
qqxu@jdl.ac.cn

Jiechao Xiong
School of Mathematical Sciences & BICMR, Peking University, Beijing 100871, China
xiongjiechao@pku.edu.cn

Qingming Huang*
University of Chinese Academy of Sciences & Institute of Computing Technology of Chinese Academy of Sciences, Beijing 100049, China
qmhuang@jdl.ac.cn

Yuan Yao*
School of Mathematical Sciences, LMAM-LMEQF-LMP, Peking University, Beijing 100871, China
yuany@math.pku.edu.cn

ABSTRACT

Strategies exploiting crowdsourcing are increasingly being applied in the area of Quality of Experience (QoE) for multimedia. They enable researchers to conduct experiments with a more diverse set of participants and at a lower economic cost than conventional laboratory studies. However, a major challenge for crowdsourcing tests is the detection and control of outliers, which may arise due to different test conditions, human errors or abnormal variations in context. For this purpose, it is desired to develop a robust evaluation methodology to deal with crowdsourceable data, which are possibly incomplete, imbalanced, and distributed on a graph. In this paper, we propose a robust rating scheme based on robust regression and Hodge Decomposition on graphs, to assess QoE using crowdsourcing. The scheme shows that the removal of outliers in crowdsourcing experiments would be helpful for purifying data and could provide us with more reliable results. The effectiveness of the proposed scheme is further confirmed by experimental studies on both simulated examples and real-world data.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Evaluation/methodology*; H.1.2 [**Models and Principles**]: User/Machine Systems—*Human factors*

*Area Chair: Martha Larson.
⋆Corresponding author.

MM '13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2502083 .

Keywords

Quality of Experience (QoE); Crowdsourcing; Paired Comparison; Robust Evaluation; Outlier Detection; LASSO; Hodge Decomposition; Random Graph

1. INTRODUCTION

Quality of Experience (QoE), which reflects the degree of a user's subjective satisfaction, has drawn increasing attention from multimedia researchers during recent years. The ultimate goal is to provide a satisfying end-user experience. Reaching this goal requires a technique that can measure the quality of multimedia content efficiently, reliably, and that is easy to implement in reality.

Among various subjective approaches for multimedia QoE evaluation, paired comparison is expected to yield more reliable results. This enables an easy and scalable implementation suitable for the purpose of calling on an Internet crowd [6] to participate in experiments using their personal computers (i.e., crowdsourcing [15]). Such a scheme uses mass collaboration and the wisdom of the crowd, and is more economical compared with conventional laboratory studies. A general framework for crowdsourcing scheme, called HodgeRank on Random Graphs (HRRG), is proposed in our previous work [33, 31]. The framework exploits various randomized paired comparison methods based on random graph theory to infer a global ranking from incomplete and imbalanced samples. It can be used to control inconsistency in paired comparison data, derive the constraints on sampling complexity to which the random selection must adhere, and allow an extension to online sampling settings [32].

Crowdsourcing subjective multimedia assessment, however, is not without pitfalls — The Crowd Is Not All Trustworthy [6]. In other words, since participants perform experiments without supervision, when the testing time for a single participant lasts too long, the participant may become impatient and may input random decisions. Such ran-

dom decisions are useless and may deviate significantly from other raters' decisions. Such outliers have to be identified for QoE evaluation.

In [6], Transitivity Satisfaction Rate (TSR) is proposed for outlier detection, which checks all the intransitive triangles, e.g., $A \succ B \succ C \succ A$. In this way, we can identify and discard inconsistent (noisy) data provided by unreliable assessors. However, TSR can only be applied for complete and balanced paired comparison data. When the paired data are incomplete, i.e., have missing edges, the question remains open of how to detect the noisy pairs.

In this paper, we fill in this gap by presenting a robust outlier detection method for crowdsourcing QoE evaluation with incomplete and imbalanced data. Our algorithm is based on a linear model that has been exploited in HodgeRank [31]. Equipped with Hodge Decomposition theory, outlier detection can be solved via sparse approximation of cyclic rankings, which consists of both harmonic and triangular cyclic rankings. Such a scheme enables us to detect outliers by solving an L1-norm regularized least squares problem and recover the global ratings simultaneously.

We demonstrate the effectiveness and generality of the proposed method on both simulated examples and real-world data. Two real datasets are considered: PC-VQA (Paired Comparison based Video Quality Assessment, complete and balanced data) and PC-IQA (Paired Comparison based Image Quality Assessment, incomplete and imbalanced data), which include 38,400 and 23,097 paired comparisons, respectively. Experimental results show that the proposed outlier detection algorithm is a promising and robust assessment method suitable for crowdsourcing QoE evaluation.

Our contributions in this work are threefold:

1. A novel method for robust evaluation of QoE is proposed to deal with incomplete and imbalanced data in crowdsourcing experiments. As in robust regression, the framework provides the possibility of carrying out the assessment procedure with automatic detection of sparse outliers.

2. In the core of the framework lies the outlier detection, formulated as a LASSO problem based on sparse approximations of cyclic ranking projection of paired comparison data. Regularization paths of LASSO provide us an order on samples suspected to be outliers.

3. Global ranking after successful outlier removal provides us a robust evaluation score with more reliable results. Through experiments on both simulated and real-world data, we show that our algorithm works effectively in practice.

The remainder of this paper is organized as follows. Section 2 contains a review of related work. Then we describe the proposed framework in Section 3, which establishes the outlier detection model based on statistical linear model. Detailed experiments are presented in Section 4, followed by the conclusions in Section 5.

2. RELATED WORK

2.1 Crowdsourcing QoE

Existing methods of QoE evaluation can be divided into two categories: subjective assessment and objective assessment. *Objective* assessment builds objective quality measurement models (see [21], a survey paper, and its references) to predict perceived quality automatically and intelligently. It may or may not reflect humans' perceptual experiences. On the other hand, *subjective* assessment can provide ground-truth and verification for objective models. It is, however, labor-intensive and time-consuming.

In subjective viewing tests, stimuli are shown to a group of viewers, and then their opinions are recorded and averaged to evaluate the quality of the stimuli. Among various approaches to conducting subjective tests, Mean Opinion Score (MOS) [1] and paired comparison are the two most popular ones. In the MOS test, individuals are asked to specify a rating from Bad to Excellent (e.g., Bad-1, Poor-2, Fair-3, Good-4, and Excellent-5) to grade the quality of a stimulus. However, such a test may suffer from various problems such as ambiguity in definition of scales and dissimilar interpretations of the scale among users [6]. For this reason, the paired comparison method is currently gaining growing attention. In this approach, raters are asked to compare two stimuli simultaneously and vote on which one has the better quality based on their perceptions. The paired comparison method is an easier, less demanding task for raters, and yields more reliable data with less personal scale bias in practice. A shortcoming of paired comparison is its more expensive sampling complexity compared to the MOS test.

To tackle the cost problem, with the growth of crowdsourcing platforms, e.g., Amazon Mechanical Turk (MTurk) [20], more and more researchers tend to seek help from the Internet crowd to conduct user studies for QoE evaluation [6, 33, 31, 32, 11, 18]. However, a major challenge of crowdsourcing QoE evaluation is that not every Internet user is trustworthy. Therefore, it is necessary to detect unreliable input and remove them since they may cause inaccuracy in the estimation of QoE scores. For example, with complete and balanced data, the method in [6] proposes TSR to measure the consistency of participants' judgments. In contrast, the outlier detection method proposed in this paper provides a general framework for outlier detection when the paired comparison data are incomplete and imbalanced.

2.2 Statistical Ranking

QoE based on paired comparisons can be recast as a statistical ranking or rating problem with paired comparison data. This problem has been widely studied in various fields including decision science [24], machine learning [13], social choice [4], and statistics [19].

In particular, recent work in [17] takes a graph theoretic view, which maps paired comparison data to edge flows on a graph, possibly imbalanced (where different pairs may receive different number of comparisons) and incomplete (where each participant may only provide partial comparisons). It then applies the combinatorial Hodge Theory to achieve an orthogonal decomposition of such edge flows into three components: gradient flow for global rating (optimal in the L2-norm sense), triangular curl flow for local inconsistency, and harmonic flow for global inconsistency. Such a perspective provides us with a universal geometric description of the structure of paired comparison data, which may help understand various models, in particular the linear models with sparse outliers used in this paper.

2.3 Outlier Detection and Robust Statistics

Outliers, also called anomalies, are typically defined to be data samples that have an unusual deviation from the most common or expected pattern. Outliers are rare events, but once they have occurred, they may lead to a large instability

of models estimated from the data. Outlier detection is a critical task in many fields and has been explored for a long time, especially in statistics [16]. In subjective quality evaluation in multimedia, outliers may arise due to different test conditions, human errors, or abnormal deviations in context factors such as those involving raters or systems. Various methods have been developed in literature for outlier detection and robust statistics. Among these studies, perhaps the most well-known one is robust regression with Huber's loss [16], which combines the least squares and the least absolute deviation problems. Recently, [10] discovered that robust regression with Huber's loss is equivalent to a LASSO problem, which leads to a new understanding of outlier detection based on modern variable selection techniques, e.g., [25]. There are also clustering-based, supervised learning-based, and semi-supervised learning-based procedures used in the artificial intelligence community [14]. However, there is no universal approach that is applicable in all settings. In this paper, we adopt a statistical linear model for the paired comparison data collected from an Internet-based crowd, namely HodgeRank model [31], and consider additive sparse outliers as they are defined in recent studies [30].

2.4 Random Graphs

Among various random graphs (i.e., Erdös-Rényi random graph [8], random regular graph [29], preferential attachment random graph [5], small world random graph [28], and geometric random graph [23]), Erdös-Rényi random graphs can be viewed as a random sampling process of pairs or edges independently and identically distributed (I.I.D.), hence they are well suited for the crowdsourcing scenario. In [33, 31], a random design principle based on Erdös-Rényi random graph theory is investigated to conduct crowdsourcing tests. Experimental results show that for a large Erdös-Rényi random graph $G(n,q)$ with n nodes and every edge sampled with probability q, it is necessary to have $q \gg n^{-1}\log n$ such that the graph is connected and global ranking is thus possible. To avoid global inconsistency from Hodge Decomposition, it suffices to have larger sampling rates at $q \gg n^{-1/2}$. In this paper, we also focus on this simple yet powerful random graph model particularly in the scenarios where outliers are present. We call it robust QoE evaluation in the setting of Erdös-Rényi random graph.

3. ROBUST HODGERANK

In this section, we propose a robust rating method based on Huber's robust regression and HodgeRank on graphs for multimedia quality assessment, thus is called robust HodgeRank here. Specifically, we first start from the linear model in HodgeRank and describe robust regression with Huber's loss. Then, we present how to detect outliers using Huber-LASSO, followed by an interpretation via Hodge Theory. Specific discussions are made with dichotomous choices. Finally, we discuss how to tune the regularization parameter in applications.

Let $\wedge = \{1, ..., m\}$ be a set of participants and $V = \{1, ..., n\}$ be the set of videos to be ranked. Paired comparison data is collected as a function on $\wedge \times V \times V$, which is *skew-symmetric* for each α, i.e., $Y_{ij}^{\alpha} = -Y_{ji}^{\alpha}$ representing the degree that α prefers i to j. Without loss of generality, one assumes that $Y_{ij}^{\alpha} > 0$ if α prefers i to j and $Y_{ij}^{\alpha} \leq 0$ otherwise. How to choose Y_{ij}^{α} can be seen in [33]. The strategy often used in QoE evaluation is dichotomous choice or a k-point Likert scale, $k = 3, 4, 5$. In this paper, we shall focus on the simplest case — dichotomous choice, in which Y_{ij}^{α} can be taken as $\{\pm 1\}$. However, the theory can be applied to more general case with multiple choices mentioned above.

In subjective multimedia assessment, it is natural to assume

$$Y_{ij}^{\alpha} = s_i^* - s_j^* + z_{ij}^{\alpha}, \tag{1}$$

where $s^* \in \mathbb{R}^V$ is some true scaling score on V and z_{ij}^{α} are noise. Define the gradient operator (finite difference operator) [17, 33] by $\delta_0 : \mathbb{R}^V \to \mathbb{R}^E$ such that $(\delta_0 s)(i,j) = s_i - s_j$, then one can rewrite (1) as

$$Y = Xs^* + z, \tag{2}$$

where the design matrix $X = \delta_0$.

If $z_{ij}^{\alpha} = \varepsilon_{ij}^{\alpha}$ represents independent noise with mean zero and fixed variance, the Gauss-Markov theorem tells us that the unbiased estimator with minimal variance is given by the following least squares problem (L2),

$$\min_{\sum_{i\in V} s_i = 0} \sum_{i,j,\alpha} (s_i - s_j - Y_{ij}^{\alpha})^2. \tag{3}$$

Such an algorithm has been used in [33, 31, 32] to derive scaling scores in subjective multimedia assessment.

However, not all comparisons are trustworthy and there may be sparse outliers due to different test conditions, human errors, or abnormal variations in context. Putting in a mathematical way, here we consider

$$z_{ij}^{\alpha} = \gamma_{ij}^{\alpha} + \varepsilon_{ij}^{\alpha}, \tag{4}$$

where outlier γ_{ij}^{α} has a much larger magnitude than $\varepsilon_{ij}^{\alpha}$ and is sparse as zero with probability $p \in (0, 1]$. When sparse outliers exist, (3) becomes unstable and may give bad estimation. How can one modify least squares problem to achieve a robust estimator against sparse outliers?

3.1 Robust Regression with Huber's Loss

Among various choices, Huber [16] proposes the following robust regression with Huber's loss function,

$$\min_{\sum_{i\in V} s_i = 0} \sum_{i,j,\alpha} \rho_{\lambda}(s_i - s_j - Y_{ij}^{\alpha}), \tag{5}$$

where Huber's loss function $\rho_{\lambda}(x)$ is defined by

$$\rho_{\lambda}(x) = \begin{cases} x^2/2, & \text{if } |x| \leq \lambda \\ \lambda|x| - \lambda^2/2, & \text{if } |x| > \lambda. \end{cases}$$

When $|s_i - s_j - Y_{ij}^{\alpha}| < \lambda$, the comparison is regarded as a "good" one with Gaussian noise and L2-norm penalty can be used on the residual. Otherwise, it is regarded as a "bad" one contaminated by outliers and L1-norm penalty should be used which is less sensitive to the amount of deviation. So when $\lambda = 0$, it reduces to a least absolute deviation (LAD) problem or L1-norm ranking [22].

A crucial question here is how to choose λ, which is equivalent to estimating the variance of $\varepsilon_{ij}^{\alpha}$ properly. For this purpose, Huber [16] proposes concomitant scale estimation, which jointly estimates s and λ as follows:

$$\min_{\sum_{i\in V} s_i = 0, \sigma} \sum_{i,j,\alpha} \rho_{\lambda_0}\left(\frac{s_i - s_j - Y_{ij}^{\alpha}}{\sigma}\right)\sigma + m\sigma, \tag{6}$$

where m is the total number of paired comparisons, $\sigma > 0$ is a scale parameter which estimates the standard deviation

of $\varepsilon_{ij}^{\alpha}$, and λ_0 controls the shape of Huber's loss where the transition from quadratic to linear takes place. A larger λ_0 implies the Huber's loss becomes more similar to least squares regression, more efficient for normally distributed data but less robust; while smaller λ_0 makes it closer to least absolute deviation regression, more robust against outliers but less efficient for normally distributed data. [16] suggests to fix $\lambda_0 = 1.35$ in order to be robust as much as possible while retaining 95% statistical efficiency for normally distributed data. Note that for fixed σ, minimization problem (6) is equivalent to minimize (5) with $\lambda = \lambda_0\sigma$. Problem (6) becomes a convex optimization problem jointly in s and σ, which can be solved efficiently.

However, we find that in our applications the concomitant scale estimation (6) only works when outliers are sparse enough. To avoid this issue, we turn to a LASSO formulation of (5).

3.2 Huber-LASSO

It's not hard to see [10] the robust regression with Huber's loss (5) is equivalent to the following optimization problem:

$$\min_{\sum_{i\in V} s_i=0,\gamma} \frac{1}{2}\|Y - Xs - \gamma\|_2^2 + \lambda\|\gamma\|_1 \quad (7)$$
$$:= \sum_{i,j,\alpha}[\frac{1}{2}(s_i - s_j + \gamma_{ij}^{\alpha} - Y_{ij}^{\alpha})^2 + \lambda|\gamma_{ij}^{\alpha}|].$$

where $X = \delta_0$. Assume (7) has solution $(\hat{s}^{lasso}, \hat{\gamma}^{lasso})$. Here we introduce a new variable γ_{ij}^{α} for each comparison Y_{ij}^{α} such that $|\gamma_{ij}^{\alpha}| > 0$ is equivalent to $|\hat{s}_i^{lasso} - \hat{s}_j^{lasso} - Y_{ij}^{\alpha}| > \lambda$, i.e., an outlier. To be less sensitive to outliers, an L1-norm penalty of $\gamma_{ij}^{\alpha} = \hat{s}_i^{lasso} - \hat{s}_j^{lasso} - Y_{ij}^{\alpha}$ is added as in Huber's loss. Otherwise, an L2-norm is used to attenuate the Gaussian noise. This optimization problem is a partially penalized LASSO [26], and is called Huber-LASSO (or HLASSO) in this paper.

A precise equivalence between (7) and (5) is given in the following proposition.

Proposition. Assume (5) has solution $\hat{s}^{huber}$. Then

$$\hat{s}^{lasso} = \hat{s}^{huber}$$

and

$$\hat{\gamma}^{lasso} = \Theta_{\lambda}(Y_{ij}^{\alpha} - (\hat{s}_i^{lasso} - \hat{s}_j^{lasso})), \quad (8)$$

where $\Theta_{\lambda}(t) = \text{sign}(t)(|t| - \lambda)_+$ is the soft-thresholding function.

HLASSO shares the same piecewise-linear regularization paths $\lambda \mapsto \hat{s}_{\lambda}$ as classical LASSO, and thus can be solved efficiently, e.g., by the LARS algorithm [7].

However, HLASSO still suffers the following issues.

- HLASSO gives a biased estimation [9], $\hat{\gamma}$ and $\hat{s}$.
- Cross-validation to find optimal λ, turns out to be highly unstable here. Since every sample is associated with an outlier variable, leaving out samples thus loses all information about the associated outlier variables.

These issues can be alleviated using the following method.

3.3 Outlier Detection

There are two groups of variables in HLASSO (7), the score s and outlier γ, and the L1-norm penalty is only applied to γ. Therefore, via orthogonal projections of data Y onto the column space of X and its complement, one can split HLASSO into two subproblems with the two groups of variables decoupled. In particular, the outlier γ is involved in a standard LASSO problem, whose design matrix comes from random projections onto the complement of the column space of X. Thanks to the exploitation of Erdös-Rényi random graphs in crowdsoucring experiments [33, 31], positions of outliers can be consistently identified with cross-validation. After locating the outliers, one can drop those comparisons contaminated by outliers and use the least squares estimation to achieve an unbiased estimation.

To see this, let X has a full SVD decomposition $X = U\Sigma V^T$ and $U = [U_1, U_2]$ where U_1 is an orthonormal basis of the column space $\text{col}(X)$ and U_2 becomes an orthonormal basis for $\ker(X^T)$. Then the following result gives a precise statement of the split of HLASSO.

Proposition. The HLASSO solution $(\hat{s}, \hat{\gamma})$ can be obtained by the following two problems

$$\min_{\gamma} \frac{1}{2}\|U_2^T Y - U_2^T\gamma\|_2^2 + \lambda\|\gamma\|_1 \quad (9)$$

$$\min_{\sum_{i\in V} s_i=0} \frac{1}{2}\|U_1^T Xs - U_1^T(Y - \hat{\gamma})\|_2^2. \quad (10)$$

It can be seen that the original HLASSO is split into two separate optimization problems: the first is a standard LASSO problem and the second is a least squares problem. Equation (9) detects outliers and (10) modifies Y using the result of (9) and calculates scores. This score is the solution of robust regression with Huber's loss.

To solve the two issues in the last subsection, we make the following notes.

- Even though the estimator of (9) is biased [9] in the estimation of the magnitudes of γ, it can consistently identify locations of outliers under mild conditions. Such conditions, roughly speaking, require that the projection matrix U_2 satisfies an incoherence (irrepresentable) condition and the sparse outliers have large enough magnitudes, which can be found precisely as in [30], and have been widely used in sign consistency of LASSO [27]. Therefore, to avoid a biased score estimation (10), we suggest to discard the outliers picked out in (9) and run L2 on the rest of data (see Algorithm 1).
- For Erdös-Rényi random graph $G(n, q)$ in crowdsourcing experiments [31], the dimension of $\text{col}(X)$ equals to $n - 1$ and the dimension of $\ker(X^T)$ thus equals to $m - n + 1$ with $m \sim n^2 q$. To ensure connectivity of $G(n, q)$, one needs $nq \gg \log n$ which implies that asymptotically $(m-n+1)/m \to 1$. Therefore, for large n, U_2^T is arbitrarily close to a random projection matrix, which satisfies the incoherence condition for the consistency of outlier detection. Moreover, for cross-validation with sparse outliers, one can use a subset of random projections as the training set and the remaining orthogonal random projections as the validation set. This increases the stability of cross-validation in applications.

These observations suggest the following algorithm for outlier detection and robust ranking, denoted by LASSO+L2 for short.

Algorithm 1: Outlier Detection and Robust Ranking.

1 **Initialization:** Compute the SVD of X and obtain U_2;
2 **Solve the spilt problem (9)**;
3 **Tuning parameter**. Determine an optimal λ^* by cross-validation with random projections;
4 **Rule out outliers and perform least squares (L2) to get an unbiased score estimation $\hat{s}$.**

3.4 Interpretation via Hodge Theory

The algorithm proposed above admits a neat interpretation from Hodge Decomposition for pairwise ranking on graphs [17]. Such a theory, referred to HodgeRank, was introduced by [33, 31] to multimedia QoE assessment. Roughly speaking, it says that all paired comparison data Y on graph G admits the following orthogonal decomposition:

$$\textit{aggregate paired comparisons} =$$

$$\text{global ranking} \oplus \text{harmonic cyclic} \oplus \text{triangular cyclic}.$$

In particular, the latter two subspaces, harmonic and triangular cyclic rankings, are both called cyclic ranking here (i.e., subspace $\ker(X^T)$).

Note that in (9), the unitary matrix U_2^T is an orthogonal projection onto the subspace of cyclic ranking. Therefore, it enables the following interpretation of outlier detection LASSO via Hodge Decomposition. The outlier γ in (9) is a *sparse approximation of the projection of paired comparison data onto cyclic ranking subspace*. This leads us to an extension of outlier detection by TSR in complete case to incomplete settings.

3.5 Dichotomous Choice

In our crowdsourcing experiments on Internet, we often meet paired comparison data with dichotomous choices, i.e.,

$$Y_{ij}^{\alpha} = \begin{cases} 1 & \text{if } \textit{participant } \alpha \text{ prefers } i \text{ to } j, \\ -1 & \text{otherwise.} \end{cases} \tag{11}$$

In [33], four general linear models are compared in terms of their total inconsistency in explaining the data, and we find that the uniform model in (1) is nearly the best. Below we present an equivalent form of (7), which groups outlier variables and solves the problem in a more efficient way.

Proposition. If $Y_{ij}^{\alpha} \in \{1, -1\}$, and let $w_{ij}^{\pm} = |\{\alpha : Y_{ij}^{\alpha} = \pm 1\}|$, then the solution of (7) is equivalent to:

$$\min_{\sum_{i\in V} s_i = 0, \gamma_{ij}^{\pm}} \quad \sum_{i,j} [\tfrac{1}{2} w_{ij}^{+} (s_i - s_j + \gamma_{ij}^{+} - 1)^2 + \lambda w_{ij}^{+} |\gamma_{ij}^{+}| + \tfrac{1}{2} w_{ij}^{-} (s_i - s_j + \gamma_{ij}^{-} - 1)^2 + \lambda w_{ij}^{-} |\gamma_{ij}^{-}|]. \tag{12}$$

Here, we group all the outlier variables with the same preference on pair (i, j). This leads to a weighted LASSO with much smaller number of variables. Such a new formulation greatly improves the efficiency of the algorithm, which has been adopted in our experiments below. Similar tricks can be applied when paired comparisons have k discrete values.

Proof. We are going to prove the property for the situation when paired comparisons have k discrete values.

Suppose $Y_{ij}^{\alpha} \in K$, $|K| = k$. Let $A_{i,j,u} = \{(i, j, \alpha) : Y_{ij}^{\alpha} = u\}, u \in K$, be the group of comparisons with same preference u on pair (i, j), $w_{ij}^{u} = |A_{i,j,u}|$ is the number of comparisons in this group and $\gamma_{ij}^{u} = \frac{1}{w_{ij}^{u}} \sum_{A_{i,j,u}} \gamma_{ij}^{\alpha}$ is the average of γ_{ij}^{α} in this group. Then for $(i, j, \alpha) \in A_{i,j,u}$, (7) becomes

$$\begin{aligned}
&\sum_{A_{i,j,u}} \left[\frac{1}{2} (s_i - s_j + \gamma_{ij}^{\alpha} - u)^2 + \lambda |\gamma_{ij}^{\alpha}| \right] \\
= &\sum_{A_{i,j,u}} \Big\{ \frac{1}{2} [(s_i - s_j + \gamma_{ij}^{u} - u)^2 + (\gamma_{ij}^{\alpha} - \gamma_{ij}^{u})^2 \\
&+ 2(s_i - s_j + \gamma_{ij}^{u} - u)(\gamma_{ij}^{\alpha} - \gamma_{ij}^{u})] + \lambda |\gamma_{ij}^{\alpha}| \Big\} \\
= &\frac{w_{ij}^{u}}{2} (s_i - s_j + \gamma_{ij}^{u} - u)^2 + \sum_{A_{i,j,u}} \Big\{ \frac{1}{2} [(\gamma_{ij}^{\alpha} - \gamma_{ij}^{u})^2 \\
&+ 2(s_i - s_j + \gamma_{ij}^{u} - u)(\gamma_{ij}^{\alpha} - \gamma_{ij}^{u})] + \lambda |\gamma_{ij}^{\alpha}| \Big\} \\
= &\frac{w_{ij}^{u}}{2} (s_i - s_j + \gamma_{ij}^{u} - u)^2 + \sum_{A_{i,j,u}} \left[\frac{1}{2} (\gamma_{ij}^{\alpha} - \gamma_{ij}^{u})^2 + \lambda |\gamma_{ij}^{\alpha}| \right], \\
&\text{since } \gamma_{ij}^{u} = \frac{1}{w_{ij}^{u}} \sum_{A_{i,j,u}} \gamma_{ij}^{\alpha}, \\
\geq &\frac{w_{ij}^{u}}{2} (s_i - s_j + \gamma_{ij}^{u} - u)^2 + \sum_{A_{i,j,u}} \lambda |\gamma_{ij}^{\alpha}|,
\end{aligned}$$

where equality holds iff $\gamma_{ij}^{\alpha} = \gamma_{ij}^{u}$, hence (7) becomes

$$\frac{w_{ij}^{u}}{2} (s_i - s_j + \gamma_{ij}^{u} - u)^2 + w_{ij}^{u} \lambda |\gamma_{ij}^{u}|.$$

The result follows from the sum over all the pair (i, j) and $u \in K$. □

Moreover, we would like to point out that the outlier detection LASSO above can be applied to the scenario of subject-based outlier detection. A simple way is to compute the percentage of outliers for an assessor, based on which one can evaluate the reliability of each assessor subject. For example, one can drop those unreliable assessors whose input data involve a large number of outliers beyond certain threshold. In other words, subject-based outlier detection can be a straightforward extension from our proposed framework on paired comparison judgment outliers.

3.6 Parameter Tuning

We have suggested cross-validation on (9) to tune parameter λ based on random projections, rather than traditional cross-validation on the origin problem (7) by leaving-out samples. This is because of a special feature in outlier detection LASSO. Since each variable is associated with a sample, sample leaving-out will lose all the information about the associated variable. Therefore, traditional cross-validation with leaving-out samples is expected to be highly unstable. To achieve a cross-validation based on random projections, one can randomly draw l-rows from projection matrix U_2^T, which can be regarded as l-random projections onto the $\ker(X^T)$, as the training set. The remaining rows of U_2^T are used for validation set. Cross-validation is then applied to such random projection based measurements using training and validation sets.

In practice, although cross-validation works for sparse and large enough outliers, we find it might fail when outliers become dense and small in magnitudes. However, when cross-validation fails, we still find it informative to look at the regularization paths of (9) directly. From the order that variables γ_{ij}^{α} become nonzero as regularization parameter λ

Figure 1: Reference videos in LIVE database.

changes from ∞ to small, one can faithfully identify the tendency that a measurement Y_{ij}^{α} is contaminated by outliers, even when cross-validation fails. Therefore, we suggest to use regularization paths to inspect the outliers in applications.

Prior knowledge can also be used to tune the regularization parameter. For example, if one would like to drop a certain percentage of outliers, say 5%, then the top 5% variables appeared on regularization paths can be regarded as outliers and dropped. Moreover, the deviation magnitudes sometimes can be used to determine outliers. For example in dichotomous choice, we can just set $\lambda = 1$. If $s_i - s_j > 0$, and $Y_{ij}^{\alpha} = -1$ so the residual $|\gamma_{ij}^{\alpha}| = |s_i - s_j - Y_{ij}^{\alpha}| > 1$, then this comparison is easy to pick out. On the other hand, if $Y_{ij}^{\alpha} = 1$, $|\gamma_{ij}^{\alpha}| > 1$ iff $s_i - s_j > 2$, the sample can reasonably be selected as an outlier.

4. EXPERIMENTS

In this section, we systematically evaluate the performance of the proposed outlier detection algorithm for QoE assessment. First, we describe the datasets used for the experiments, which include both simulated and real-world data. Then, Algorithm 1 is applied to both datasets. To avoid the situations in which the optimal choice of λ^* by cross-validation is too conservative to detect outliers when they are dense and of small magnitudes (as discussed in Section 3.6), we look at the whole regularization path $\hat{\gamma}_{\lambda}$ by varying λ from ∞ to 0. AUC for ROC curves is used in the experiments to measure if the true outliers are detected by early appearance on regularization paths. The higher the AUC, the better the performance. Finally, some further discussions are provided.

4.1 Datasets

Three datasets are used in this work. The datasets include simulated data, PC-VQA (Paired Comparison based Video Quality Assessment) data [33], and PC-IQA (Paired Comparison based Image Quality Assessment) data [32].

In simulated data, we first create a random total order on n candidates V as the ground-truth and add paired comparison edges $(i, j) \in E$ to graph $G = (V, E)$ randomly, with the preference direction following the ground-truth order. To create sparse outliers, a random subset of E is reversed in preference direction. In this way, we simulate a paired comparison graph, possibly incomplete and imbalanced, with outliers.

The second dataset, PC-VQA, collected by [33], contains 38400 paired comparisons for LIVE dataset [3] (Figure 1) from 209 observers. One of the advantageous properties of this dataset is that the paired comparison data is complete and balanced.

The third dataset, PC-IQA, contains 15 reference images and 15 distorted versions of each reference, for a total of 240 images which come from two publicly available datasets, LIVE [3] and IVC [2] (Figure 2). Totally, 186 observers, each of whom performs a varied number of comparisons via

Figure 2: Reference images in LIVE and IVC databases. (The first six are from LIVE and the remaining nine are from IVC.)

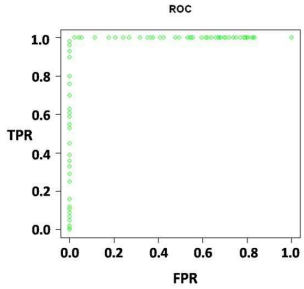


Figure 3: ROC curve of (2000,5%) for simulated data.

Internet, provide 23,097 paired comparisons (the resulting graph is incomplete and imbalanced) for subjective IQA.

These three datasets provide us both simulated and real-world paired data and hence can all be used for the experiments of QoE task. In the following, we first show the effectiveness of the proposed method on simulated data, then further confirm the effectiveness on real-world datasets.

4.2 Simulated Data

We choose $|V| = n = 16$ in the simulated graph $G = (V, E)$, which is consistent with the other real-world datasets. We make the following definitions of experimental parameters. The total number of paired comparisons occurred on this graph is SN (Sample Number), and the number of outliers is ON (Outlier Number). Finally, we define the outlier percentage OP = ON/SN.

In order to evaluate the performance of LASSO in outlier detection, for each pair of (SN,OP), we compute the regularization path $\hat{\gamma}^{\lambda}$ of LASSO by varying regularization parameter λ from ∞ to 0, which is solved by R-package `quadrupen` [12]. The order in which $\hat{\gamma}_{ij}^{\lambda}$ becomes nonzero gives a ranking of the edges according to their tendency to be outliers. Since we have the ground-truth outliers, the ROC curve can be plotted by thresholding the regularization parameter λ at different levels which creates different true positive rates (TPR) and false positive rates (FPR). For example, when SN = 2000 and OP = 5%, the ROC curve can be seen in Figure 3. With different choices of SN and OP, Area Under the Curve (AUC) are computed with standard deviations over 20 runs and shown in Table 1 to measure the performance of LASSO in outlier detection. It can be seen that when samples are large and outliers are sparse, AUC is close to 1. This implies that the regularization paths of LASSO give an accurate estimation of outliers (indicated by a small FPR with large TPR), where samples appearing early on LASSO paths are mostly contaminated by outliers. Figure 4 illustrates an example of such LASSO paths.

We note that when OP = 50%, i.e., half of the edges are reverted by outliers, Table 1 shows a rapid decrease of AUC to about 0.5, which is the performance of random guess.

Table 1: AUC over (SN,OP) for simulated data, 20 times repeat.

AUC (sd)	OP=5%	OP=10%	OP=15%	OP=20%	OP=25%	OP=30%	OP=35%	OP=40%	OP=45%	OP=50%
SN=1000	0.999(0)	0.999(0.001)	0.998(0.001)	0.996(0.003)	0.992(0.005)	0.983(0.010)	0.962(0.016)	0.903(0.038)	0.782(0.050)	0.503(0.065)
SN=2000	0.999(0)	0.999(0)	0.999(0)	0.998(0.001)	0.997(0.001)	0.992(0.004)	0.986(0.007)	0.956(0.019)	0.849(0.052)	0.493(0.086)
SN=3000	0.999(0)	0.999(0)	0.999(0)	0.999(0)	0.998(0)	0.996(0.002)	0.990(0.004)	0.971(0.013)	0.885(0.032)	0.479(0.058)
SN=4000	0.999(0)	0.999(0)	0.999(0)	0.999(0)	0.999(0)	0.997(0.001)	0.994(0.002)	0.980(0.008)	0.903(0.028)	0.519(0.055)
SN=5000	0.999(0)	0.999(0)	0.999(0)	0.999(0)	0.999(0)	0.998(0.001)	0.994(0.002)	0.984(0.009)	0.933(0.022)	0.501(0.066)

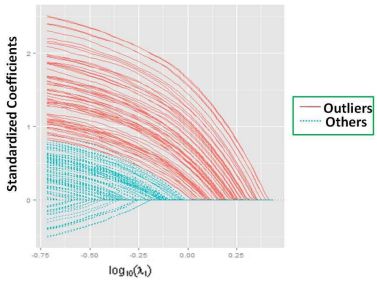
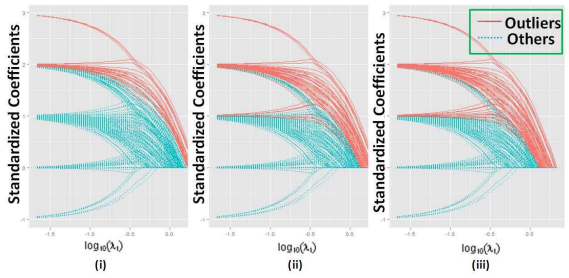


Figure 4: The regularization paths of (2000,5%) for simulated data. The true outliers (plotted in red color) mostly lie outside the majority of paths.

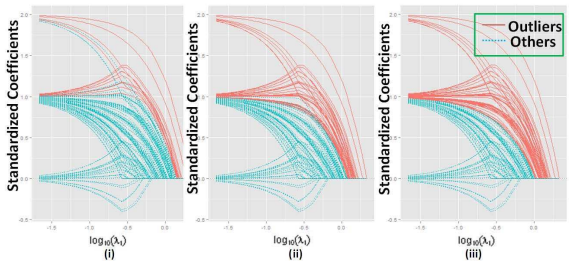
This is expected, since when more than half of the edges are perturbed, it is impossible to distinguish the signal from noise by any method. A phase transition can be observed in this table, that AUC rapidly approaches to 1 as long as OP drops below 50% and SN increases.

4.3 Real-world Data

As there is no ground-truth for outliers in real-world data, one can not exploit ROC and AUC as in simulated data to evaluate outlier detection LASSO here. In this subsection, we inspect the top p% pairs returned by LASSO regularization paths and compare them with the whole data to see if they are reasonably good outliers.



(a) Reference River Bed in PC-VQA dataset



(b) Reference 10 in PC-IQA dataset

Figure 5: Regularization paths. (i) Top 2% outliers; (ii) Top 5% outliers; (iii) Top 8% outliers.

For simplicity, in the PC-VQA dataset, we randomly take River Bed as an illustrative example (other reference videos exhibit similar results). Figure 5(a) shows the top 2%, 5%, and 8% outliers picked by regularization paths in River Bed. In order to allow a closer inspection of these pairs, outliers are marked in three colors (red for top 2%, green for top 2%-5%, and blue for top 5%-8%) in the paired comparison matrix in Table 2(a). The paired comparison matrix is constructed as follows (Table 2(b) is constructed in the same way). For each video pair $\{i, j\}$, let n_{ij} be the number of comparisons, for which a_{ij} raters agree that the quality of i is better than j (a_{ji} carries the opposite meaning). So $a_{ij} + a_{ji} = n_{ij}$ if no tie occurs. In the PC-VQA dataset, $n_{ij} \equiv 32$ for all videos. The order of the video ID in Table 2(a) is arranged from high to low according to the global ranking score calculated by the least squares method (3). It is interesting to see that the outliers picked out are mainly distributed in the lower left corner of this matrix, which implies that the outliers are those preference orders with a large deviation from the global ranking scores by L2. In addition, the earlier a pair is detected by LASSO as an outlier, the closer it will be to the lower left corner and the larger such a deviation is. Moreover, Figure 6 further confirms this phenomenon. Here, all the top 5% outliers are reversed preference arrows pointing from lower quality to higher quality videos. Clearly one can see that video V6, V12, V2, V11, and V8 are the top 5 videos which brought in the largest number of outliers in data collection.

To see the effect of outliers on global ranking scores, Table 3(a) shows the outcomes of three ranking algorithms, namely L2, HLASSO, and LASSO+L2. We choose the λ which finds 5% outliers. As we noted earlier, the global ranking scores directly returned by HLASSO are possibly biased in the estimation of large outliers, hence LASSO+L2 is used toward less-biased solution. While video V3 is nearly the same quality as V7 in L2, both HLASSO and LASSO+L2 put V3 worse than V7. Removal of the top 5% outliers in LASSO+L2 further changes the orders of some competitive videos, such as V15 and V10, V2 and V11, which shows that the effect of outliers is mainly within the highly competitive groups.

The experiments above demonstrate the effectiveness of the proposed method on complete and balanced data. To illustrate the detection ability on incomplete and imbalanced data, the PC-IQA dataset is taken into consideration. Figure 5(b), Table 2(b), and Figure 7 show the experimental results on a randomly selected reference (image 10 in Figure 2). Similar observations as above can be made and we note that outliers distributed on this dataset are much sparser than PC-VQA, shown by many zeros in the lower left corner of Table 2(b). The outcomes of three ranking algorithms with the top 5% outliers are shown in Table 3(b) for PC-IQA. Based on 5% outliers detection, both HLASSO and LASSO+L2 differ with L2 in that image ID =

Table 2: Paired comparison matrixes, with red pairs for top 2% outliers, green for top 2%-5% outliers, and blue for top 5%-8% outliers.

(a) Reference River Bed in PC-VQA dataset

Video ID	1	13	9	14	5	15	10	3	7	16	4	8	2	11	12	6
1	0	29	28	31	32	30	32	31	30	27	27	31	25	31	32	32
13	3	0	23	22	24	24	14	22	25	29	28	27	25	27	28	31
9	4	9	0	25	5	20	26	25	15	26	23	27	26	27	30	31
14	1	10	7	0	14	29	19	22	14	23	23	23	27	23	26	25
5	0	8	27	18	0	14	11	12	22	13	22	18	25	30	31	30
15	2	8	12	3	18	0	20	24	13	23	18	17	28	22	30	29
10	0	18	6	13	21	12	0	18	11	13	18	27	26	26	28	29
3	1	10	7	10	20	8	14	0	14	17	19	19	18	28	28	32
7	2	7	17	18	10	19	21	18	0	18	8	23	9	15	30	30
16	5	3	6	9	19	9	19	15	14	0	18	24	23	20	30	30
4	5	4	9	9	10	14	14	13	24	14	0	17	15	23	30	30
8	1	5	5	9	14	15	5	13	9	8	15	0	25	13	19	24
2	7	7	6	5	7	4	6	14	23	9	17	7	0	18	17	29
11	1	5	5	9	2	10	6	4	17	12	9	19	14	0	24	26
12	0	4	2	6	1	2	4	4	2	2	2	13	15	8	0	19
6	0	1	1	7	2	3	3	0	2	2	2	8	3	6	13	0

(b) Reference 10 in PC-IQA dataset

Image ID	1	6	9	12	10	2	16	7	15	11	8	13	14	3	4	5
1	0	11	10	12	11	20	11	15	13	14	16	15	14	17	14	13
6	3	0	5	10	8	11	11	12	10	10	12	13	12	11	14	16
9	3	7	0	5	7	6	3	10	4	5	8	6	5	11	14	17
12	3	2	3	0	6	9	13	7	8	5	8	13	13	13	16	17
10	5	4	0	2	0	7	2	5	6	7	9	5	6	12	17	19
2	0	3	5	4	4	0	8	9	9	13	6	11	12	12	13	13
16	3	1	1	2	2	4	0	6	16	8	7	16	16	15	20	14
7	0	2	1	1	4	2	4	0	7	5	12	8	9	13	17	16
15	0	4	1	4	1	4	3	2	0	8	7	12	16	16	16	14
11	0	0	0	0	0	0	3	4	1	0	9	2	6	11	14	17
8	0	0	0	0	0	4	0	0	0	0	0	5	8	10	15	17
13	0	0	0	0	0	0	0	0	4	2	1	0	12	13	15	16
14	0	0	1	0	0	0	1	0	0	2	1	2	0	6	12	10
3	0	0	0	0	0	0	0	0	0	0	0	0	9	0	10	10
4	0	0	0	0	0	0	0	0	0	2	0	2	4	1	0	9
5	0	0	0	0	0	0	0	0	0	1	0	0	5	1	5	0

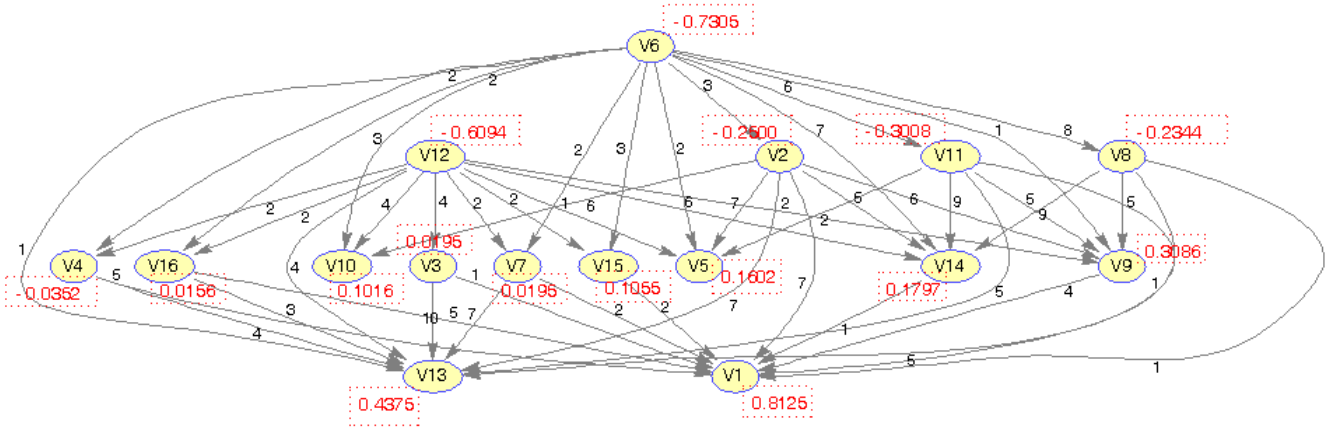


Figure 6: Top 5% outliers for reference River Bed in PC-VQA dataset. The integer on each curve represents a_{ij} defined in subsection 4.3 and the red decimal in dotted rectangle represents the global ranking score returned by L2 algorithm.

Table 3: Comparison of different rankings (5% outliers are considered). Three ranking methods are compared with the integer represents the ranking position and the number in parentheses represents the global ranking score returned by the corresponding algorithm.

(a) Reference River Bed in PC-VQA dataset

Video ID	L2	LASSO+L2	HLASSO
1	1 (0.8125)	1 (0.9245)	1 (0.8414)
13	2 (0.4375)	2 (0.6003)	2 (0.4615)
9	3 (0.3086)	3 (0.3620)	3 (0.3182)
14	4 (0.1797)	4 (0.3034)	4 (0.1978)
5	5 (0.1602)	5 (0.1996)	5 (0.1659)
15	6 (0.1055)	7 (0.1172)	6 (0.1098)
10	7 (0.1016)	6 (0.1458)	7 (0.1084)
3	8 (0.0195)	10 (-0.0021)	9 (0.0193)
7	8 (0.0195)	8 (0.0043)	8 (0.0194)
16	10 (0.0156)	9 (0.0015)	10 (0.0133)
4	11 (-0.0352)	11 (-0.0538)	11 (-0.0390)
8	12 (-0.2344)	12 (-0.2758)	12 (-0.2403)
2	13 (-0.2500)	14 (-0.3758)	13 (-0.2695)
11	14 (-0.3008)	13 (-0.3587)	14 (-0.3130)
12	15 (-0.6094)	15 (-0.7181)	15 (-0.6347)
6	16 (-0.7305)	16 (-0.8743)	16 (-0.7586)

(b) Reference 10 in PC-IQA dataset

Image ID	L2	LASSO+L2	HLASSO
1	1 (0.8001)	1 (0.8876)	1 (0.8144)
6	2 (0.6003)	2 (0.7034)	2 (0.6143)
9	3 (0.5362)	3 (0.6048)	3 (0.5484)
12	4 (0.4722)	4 (0.4886)	4 (0.4752)
10	5 (0.3472)	6 (0.2698)	5 (0.3368)
2	6 (0.3044)	5 (0.2859)	6 (0.3105)
16	7 (0.2756)	7 (0.2677)	7 (0.2757)
7	8 (0.1403)	8 (0.1398)	8 (0.1374)
15	8 (0.0965)	9 (0.0540)	9 (0.0865)
11	10 (-0.1609)	10 (-0.1815)	10 (-0.1563)
8	11 (-0.2541)	11 (-0.2813)	11 (-0.2620)
13	12 (-0.2964)	12 (-0.2927)	12 (-0.2958)
14	13 (-0.6215)	14 (-0.6478)	14 (-0.6361)
3	14 (-0.6315)	13 (-0.6246)	13 (-0.6315)
4	15 (-0.7822)	15 (-0.8098)	15 (-0.7889)
5	16 (-0.8262)	16 (-0.8639)	16 (-0.8287)

3 (`fruit_flou_f3.bmp` in IVC [2] database) is better than image ID = 14 (`fruit_lar_r1.bmp`). Such a preference is in agreement with the pairwise majority voting of 9:6 votes (Table 2(b)). Therefore, the example shows that under sparse outliers L2 ranking may be less accurate.

Moreover, LASSO+L2 further suggests that image ID = 2 should be better than image ID = 10, in contrast to the other two algorithms. A further inspection of the dataset confirms that such a suggestion by LASSO+L2 is reasonable. Figure 8 shows the two images (ID = 2 and ID = 10) from the IVC [2] database. There is a blurring effect in image ID = 2 and a blocking effect in the background of ID = 10. In the dataset, 4 raters agree that the quality of ID = 2 is better than ID = 10, while 7 raters have the opposite opinion. Clearly LASSO+L2 chooses the preference of the minority, based on aggregate behavior over population after removal of some outliers. Why does this happen? In fact, when a participant compares the quality between ID = 2 and ID = 10, his preference depends on his attention — on the foreground or on the whole image. A rater with foreground attention might be disturbed by the blurring effect, leading to $10 \succ 2$. On the other hand, a rater with holistic attention may notice the blocking effect in the background, leading to $2 \succ 10$. Which criterion might be dominant? To explore this question, we further collected more clean data (i.e., 20 more persons provide careful judgments in controlled lab conditions), among which a dominant percentage (80%) agrees with $2 \succ 10$, consistent with the LASSO+L2 prediction after removal of outliers. This suggests that most observers assess the quality of an image from a global point of view. Another less stable way is to select a subset of clean

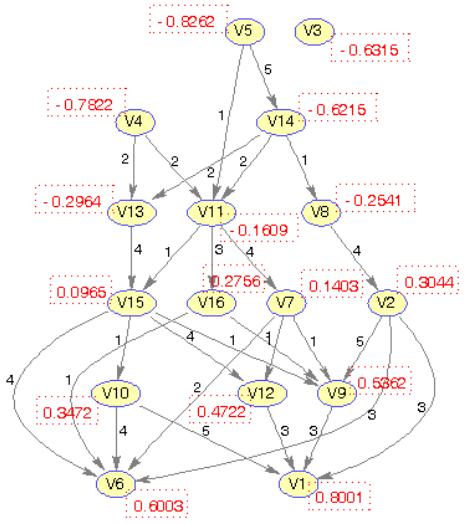


Figure 7: Top 5% outliers for reference 10 in PC-IQA dataset. The integer on each curve and the red decimal in dotted rectangle carry the same meanings with those in Figure 6.

data without outliers for validation, which points generally towards LASSO+L2<HLASSO<L2 in least squares error. Such a result suggests that for those highly competitive or confused alternative pairs, a large number of samples are expected to find a good ranking in majority voting; on the other hand, by exploiting intermediate comparisons of good confidence with other alternatives, it is possible to achieve a reliable global ranking with a much smaller number of samples, such as what LASSO+L2 does here.

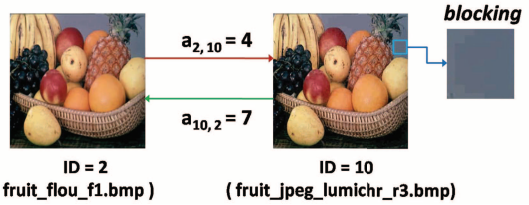


Figure 8: Dissimilar judgments due to multi-criteria in paired comparisons among users. The image is undistinguishable due to its small size, so image names in IVC [2] database are printed here.

4.4 Discussion

As we have seen, after dropping outliers, LASSO+L2 gives a different order on some images compared with L2, indicating that outliers may cause different rank aggregations. The differences mostly lie among the highly competitive alternatives. The order returned by HLASSO somewhere lies between the two orders returned by L2 and LASSO+L2, showing that the bias in HLASSO estimation makes it less sensitive to changes. From these results, we may conclude that outlier detection by our proposed framework is effective, and that using the framework, one can prune unreliable paired comparisons and achieve a good global ranking score with less data collected.

In our experiments, we observe that cross-validation may work when outliers with large magnitudes are sparse enough. When cross-validation works it helps to identify an optimal regularization parameter and sparsity pattern. However, in our QoE evaluation scenario, $|Y_t(i,j)|$ is bounded by 1. For

this reason, it is hard for sparse outliers to satisfy the large magnitude assumption. In this case, cross-validation may exhibit conservative and unstable behavior, either picking out all the outliers or none of them. Therefore, as we suggested earlier, users applying our framework should compute regularization paths whenever possible to inspect outliers.

Additionally, we find that when raters are asked to consider different pairs of images, they implicitly take different salient features into account. Dissimilar judgments due to noise will vanish when the sample size goes to infinity while disagreements due to multi-criteria will persist with the increase of sample size. This phenomenon will require additional investigation in the future.

5. CONCLUSIONS

In this paper, we have proposed a framework for robustly assessing the QoE of multimedia content, in which outliers are automatically detected and robust global ranking scores are obtained after outlier removal. In such a framework, outliers are sparse approximations of cyclic ranking projections of paired comparison data and regularization paths of outlier detection LASSO provides us an informative procedure. In particular, outlier detection LASSO, provides benefits in crowdsourcing experiments with the Erdös-Rényi random graph designs. Experiments are presented with both simulations and two real-world datasets including both representative images and videos in QoE with crowdsourcing data. It is shown that a two-stage method (outlier detection LASSO followed by a least squares on cleaned data) can remove sparse outliers and achieve a global rating more consistent with the population than the least squares method applied to the whole data. The framework enables us to derive reliable global ratings after purifying data, and as such provides a helpful tool for the multimedia community, which exploits crowdsourceable paired comparison data for robust ranking. Finally, we would like to point out that with the rapid development of technologies for rich user interfaces, the proposed framework can be extended via stochastic optimization to assess users' experience in an online setting, which is a largely unexplored field and one of our future directions.

6. ACKNOWLEDGMENTS

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400 and 2012CB825501, in part by National Natural Science Foundation of China: 61025011 and 61071157.

7. REFERENCES

[1] *ITU-R Recommendation P.800. Methods for subjective determination of transmission quality*, 1996.

[2] Subjective quality assessment irccyn/ivc database. `http://www2.irccyn.ec-nantes.fr/ivcdb/`, 2005.

[3] LIVE image & video quality assessment database. `http://live.ece.utexas.edu/research/quality/`, 2008.

[4] K. Arrow. A difficulty in the concept of social welfare. *Journal of Political Economy*, 58(4):328–346, 1950.

[5] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[6] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei. A crowdsourceable QoE evaluation framework for multimedia content. pages 491–500. ACM Multimedia, 2009.

[7] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[8] P. Erdös and A. Rényi On random graphs I. *Publicationes Mathematicae-Debrecen*, 6:290–297, 1959.

[9] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, pages 1348–1360, 2001.

[10] I. Gannaz. Robust estimation and wavelet thresholding in partial linear models. *Statistics and Computing*, 17:293–310, 2007. `arXiv:math/0612066v1`.

[11] B. Gardlo, M. Ries, and T. Hoßfeld. Impact of screening technique on crowdsourcing QoE assessments. International Conference Radioelektronika 2012, Special Session on Quality in multimedia systems, 2012.

[12] Y. Grandvalet, C. Ambroise, and J. Chiquet. Sparsity by worst-case quadratic penalties. 2012. `arXiv:1210.2077`.

[13] R. Herbrich, T. Graepel, and K. Obermayer. *Large margin rank boundaries for ordinal regression*. MIT Press, 2000.

[14] V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

[15] J. Howe. The rise of crowdsourcing. *Wired Magazine*, 14(6):176–183, 2006.

[16] P. J. Huber. *Robust Statistics*. New York: Wiley, 1981.

[17] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 2010.

[18] C. Keimel, J. Habigt, and K. Diepold. Challenges in crowd-based video quality assessment. In *Forth International Workshop on Quality of Multimedia Experience (QoMEX 2012)*, 2012.

[19] M. Kendall and B. Smith. On the method of paired comparisons. *Biometrika*, 31(3-4):324–345, 1940.

[20] A. Kittur, E. Chi, and B. Suh. Crowdsourcing user studies with Mechanical Turk. pages 453–456. SIGCHI conference on Human factors in computing systems, 2008.

[21] W. Lin and C.-C. J. Kuo. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation*, 22(4):297–312, 2011.

[22] B. Osting, J. Darbon, and S. Osher. Statistical ranking using the L_1-norm on graphs. *AIMS'*, 2012.

[23] M. Penrose. *Random Geometric Graphs (Oxford Studies in Probability)*. Oxford University Press, 2003.

[24] T. Saaty. A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology*, 15(3):234–281, 1977.

[25] Y. She and A. B. Owen. Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106(494):626–639, 2011.

[26] R. Tibshirani. Regression shrinkage and selection via the LASSO. *J. of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[27] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using L_1-constrained quadratic programming (LASSO). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

[28] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, (393):440–442, 1998.

[29] N. Wormald. Models of random regular graphs. pages 239–298. In Surveys in Combinatorics, 1999.

[30] J. Xiong, X. Cheng, and Y. Yao. Robust ranking on graphs. *Journal of Machine Learning Research*, submitted, 2013.

[31] Q. Xu, Q. Huang, T. Jiang, B. Yan, W. Lin, and Y. Yao. HodgeRank on random graphs for subjective video quality assessment. *IEEE Transactions on Multimedia*, 14(3):844–857, 2012.

[32] Q. Xu, Q. Huang, and Y. Yao. Online crowdsourcing subjective image quality assessment. pages 359–368. ACM Multimedia, 2012.

[33] Q. Xu, T. Jiang, Y. Yao, Q. Huang, B. Yan, and W. Lin. Random partial paired comparison for subjective video quality assessment via HodgeRank. pages 393–402. ACM Multimedia, 2011.