

# Parsimonious Mixed-Effects HodgeRank for Crowdsourced Preference Aggregation

Qianqian Xu, Jiechao Xiong, Xiaochun Cao, and Yuan Yao

**Abstract**—In crowdsourced preference aggregation, it is often assumed that all the annotators are subject to a common preference or utility function which generates their comparison behaviors in experiments. However, in reality annotators are subject to variations due to multi-criteria, abnormal, or a mixture of such behaviors. In this paper, we propose a parsimonious mixed-effects model based on HodgeRank, which takes into account both the fixed effect that the majority of annotators follows a common linear utility model, and the random effect that a small subset of annotators might deviate from the common significantly and exhibits strongly personalized preferences. HodgeRank has been successfully applied to subjective quality evaluation of multimedia and resolves pairwise crowdsourced ranking data into a global consensus ranking and cyclic conflicts of interests. As an extension, our proposed methodology further explores the conflicts of interests through the random effect in annotator specific variations. The key algorithm in this paper establishes a dynamic path from the common utility to individual variations, with different levels of parsimony or sparsity on personalization, based on newly developed Linearized Bregman Algorithms with Inverse Scale Space method. In this dynamic scheme, three kinds of losses are systematically discussed, including L2 loss, Bradley-Terry, Thurstone-Mosteller. Finally the validity of the methodology are supported by experiments with both simulated examples and three real-world crowdsourcing datasets, which shows that our proposed method exhibits better performance (i.e. smaller test error) compared with HodgeRank due to its parsimonious property.

**Index Terms**—Preference Aggregation; HodgeRank; Mixed-Effects Models; Linearized Bregman Iterations; Personalized Ranking; Position Bias.

## 1 INTRODUCTION

With the Internet and its associated explosive growth of information, individuals today in the world are facing with the rapid expansion of multiple choices (e.g., which book to buy, which hotel to book, etc.). Inferring user's preference or utility over a set of alternatives has thus become an important issue. Among various methods to infer user viewpoint/preference, crowdsourcing technology is becoming a new paradigm, which collects voting data from a large crowd or population on Internet and pursue some statistical preference aggregations. For example, the following platforms are frequently used by researchers to crowdsource voting data of participants: [MTurk](#), [InnoCentive](#), [CrowdFlower](#), [CrowdRank](#), and [AllOurIdeas](#), etc. A typical and perhaps the simplest scenario is the pairwise comparison experiment. Specifically, there are a set of items to rank, and participants are asked to choose between various pairs among these items; the goal is to aggregate these

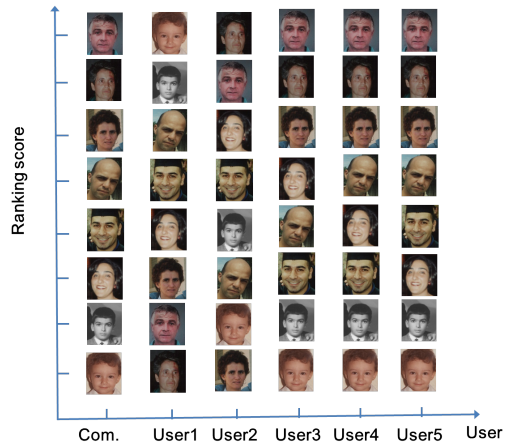


Fig. 1: An illustrative example on estimation of human ages, with the fixed effect of common ranking and random effects of user's personalized ranking. The faces in the first column are ordered according to a common ranking score aggregated from crowdsourced pairwise comparisons, while other columns are according to personalized rankings of different users. The ground truth ages in the first column, in a top-down order, are 46, 51, 36, 30, 25, 23, 10, and 2, respectively.

• Q. Xu and X. Cao are with State Key Laboratory of Information Security (SKLOIS), Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093 & BICMR, Peking University, Beijing 100871, China, (email: xuqianqian@iie.ac.cn; caoxiaochun@iie.ac.cn).

• J. Xiong and Y. Yao are with BICMR-LMAM-LMEQF-LMP, School of Mathematical Sciences, Peking University, Beijing 100871, China, (email: xiongjiechao@pku.edu.cn; yuanyao@math.pku.edu.cn).

pairwise comparisons into a global consensus ranking that summarizes the preference of all users. We have seen that researchers exploit such a paradigm to

evaluate the quality of multimedia content [4], predict image/video interestingness [12], estimate ages from face pictures [13], and rank taste of food [14] etc.

However, different individuals might very well have distinct preferences, such that participants of the crowdsourced experiments might vote under different criteria or conditions. It might be misleading to merely look at a global consensus while ignoring personal diversity. For example Fig. 1 shows an age estimation from photos that will be discussed in detail later in this paper. The majority is not always correct, as the common ranking by mistake thinks 46 older than 51! Moreover, even though User2 is largely deviated from the common ranking, it is noticeable that he/she makes correct judgements between the faces of year 46 and 51. So if one is looking for features that correctly predict the ages of these two particular faces, User2 is a better consultant than the majority. Particularly, User1 is clearly an adversarial voter, whose personalized ranking is largely against the common ranking reflecting the majority, so should be removed from a preference aggregation procedure.

Moreover, in crowdsourcing experiments, the participants are distributed over the Internet with a diverse environment. Even they might share the same preference or utility function in making choices, they might suffer various disturbances during the experiments. For example, i) one typically clicks one side more often than another. As some pairs are highly confusing or annotators get too tired, in these cases, some annotators tend to click one side hoping to simply raise their record to receive more payment; while for pairs with substantial differences, they click as usual. ii) some extremely careless annotators, or robots pretending to be human annotators, actually do not look at the instances and click one side all the time to quickly receive payment for work. Such a kind of behavior is called the annotator's position bias which has been studied in [7].

These examples above suggest us that we have to take into account of user or annotator specific variations in a crowdsourced preference aggregation task. As the classical social choice theory [1] points out, preference aggregation toward a global consensus is doomed to meet the conflicts of interests. *What is a suitable way to quantitatively analyze the conflicts of interests?*

In this paper, we pursue the Hodge-theoretic approach by [15] which decomposes the pairwise comparison data into three orthogonal components: the global consensus ranking, the local inconsistency as triangular cycles, and the global inconsistency as harmonic cycles. The latter two are both cycles, collectively decoding all the conflicts of interests in the data. Instead of the merely extracting from the data the global ranking component, often called *HodgeRank* which has been introduced into the quality assessment of multimedia by [26], we extend it here by

including some annotator-specific random effects to further decompose the cycles. To decipher the sources of the conflicts of interests, we mainly consider two types of annotator-specific variations: annotator's personalized preference deviations from the common ranking which characterize multi-criteria in data, and annotator's position bias which deteriorates the quality of data. This results in a linear mixed-effects extension of HodgeRank, called Mixed-Effects HodgeRank here.

To initiate a task of crowdsourced preference aggregation, we usually assume the majority of participants share a common preference interest and behave rationally, while deviations from that exist but are sparse. So a parsimonious model is assumed in this paper, with sparsity structure on personalized preference deviations and position biases. Due to the unknown amount of such sparse random effects in reality, it is natural to pursue a family of parsimonious models at a variety levels of sparsity. Algorithmically we adopt the Linearized Bregman Iteration, which is a simple iterative procedure generating a sequence of parsimonious models, evolving from the common global ranking in HodgeRank, to annotator's personalized ranking till a full model. Fig. 1 is in fact a result of our algorithm. As the algorithm iterates, typically the abnormal annotators with large preference deviations and/or position biases appear early, and the annotators who behave normally appear at a later stage. In practice when the number of participants is large and sample size is relatively small, early stopping regularization is needed to prevent the overfitting in full model.

Equipped with such a new scheme, given a set of entities, we choose a set of entity pairs and ask Internet crowds which entity is more preferable in each pair. Based on the feedback we not only can derive the common preference on population-level, but also can estimate rapidly an annotator's large preference/utility deviation in an individual-level, and an abnormal annotator's position bias. Individual preference deviations from the population common ranking are helpful to understand different criteria among annotators when they judges, and especially to monitor the adversarial users. On the other hand, annotator's position bias is a helpful tool to monitor the quality of his/her voting data, through the mixing behavior that the annotator simply clicks one side of the pair in comparisons without paying attention to their contents. Such a statistical mixed-effects framework simultaneously considers both the fixed effect of common ranking as the HodgeRank and the random effects of annotator-specific variations, which, up to the author's knowledge, has not been seen in literature.

As a summary, our main contributions in this new framework are highlighted as follows:

(A) A linear mixed-effects extension of HodgeRank

including both the fixed effect of common ranking, and the random effects of annotator's preference deviation with position bias;

- (B) A path of parsimonious estimates of the preference deviation and position bias at different sparsity levels, based on Linearized Bregman Iterations.
- (c) Three kinds of losses (i.e. L2 loss, Bradley-Terry, Thurstone-Mosteller) in the dynamic Linearized Bregman Iterations scheme.

This paper is an extension of our conference paper [27], where only L2 loss is considered. Though it is much easier to solve and the Hodge Theory is adapted to the new inner product with orthogonal decomposition, it may lose some efficiency than Maximum likelihood methods. Therefore, in this paper, three kinds of losses (i.e. L2 loss, Bradley-Terry, Thurstone-Mosteller) are taken into consideration with detailed comparative analysis.

The remainder of this paper is organized as follows. Sec.2 contains a review of related works. Then we systematically introduce the methodology for parsimonious mixed-effects HodgeRank estimation in Sec.3. Extensive experimental validation based on one simulated and three real-world crowdsourced datasets are demonstrated in Sec.4. Finally, Sec.5 presents the conclusive remarks.

## 2 RELATED WORK

### 2.1 Statistic Ranking Aggregation

Statistical preference aggregation, in particular ranking or rating from pairwise comparisons, is a classical problem which can be traced back to the 18<sup>th</sup> century. Various algorithms have been studied for this problem, including maximum likelihood under a Bradley-Terry model assumption, rank centrality (PageRank/MC3) [6], [17], HodgeRank [15], and a pairwise variant of Borda count [8] among others. In [22], it shows that under a natural statistical model, where pairwise comparisons are drawn randomly and independently from some underlying probability distribution, the rank centrality (PageRank) and HodgeRank algorithms both converge to an optimal ranking under a "time-reversibility" condition. However, PageRank is only able to aggregate the pairwise comparisons into a global ranking over the items. HodgeRank not only provides us a mean to determine a global ranking under various statistical models, but also measures the inconsistency of the global ranking obtained.

HodgeRank, as an application of combinatorial Hodge theory to the preference or rank aggregation problem from pairwise comparison data, was first introduced in [15], inspiring a series of studies in statistical ranking [20], [21] and game theory [3], in addition to traditional applications in fluid mechanics [5] and computer vision [30], etc. It is a general

framework to decompose paired comparison data on graphs, possibly imbalanced (where different video pairs may receive different number of comparisons) and incomplete (where every participant may only give partial comparisons), into three orthogonal components. In these components HodgeRank not only provides us a mean to determine a global ranking from paired comparison data under various statistical models (e.g., Uniform, Thurstone-Mosteller, Bradley-Terry, and Angular Transform), but also measures the inconsistency of the global ranking obtained. The inconsistency shows the validity of the ranking obtained and can be further studied in terms of its geometric scale, namely whether the inconsistency in the ranking data arises locally or globally. Local inconsistency can be fully characterized by triangular cycles, while global inconsistency involves cycles consisting nodes more than three, which may arise due to data incompleteness and once presented with a large component indicates some serious conflicts in ranking data. However through random graphs, we can efficiently control global inconsistency.

However, all of these methods have a major drawback: they aim to find one ranking thus cannot model the discrepancies across users. Therefore, in recent years, personalized ranking methods arise to fill in this gap. This task can be seen as rank aggregation analog to the standard collaborative filtering (CF) problem. There have been many CF algorithms, including Bayesian networks, clustering models, and latent semantic models, etc. Recent algorithms for collaborative filtering are mostly based on matrix factorization [23], [24]. The key idea behind them is to find a low rank user rating matrix that best approximates the observed ratings. Most recently, the application of the nuclear norm approach to CF was first proposed by [28], which shows good empirical evidence for using such a nuclear norm regularized based approach. The key difference between our study and the low rank matrix collaborative filtering algorithms is that we assume the majority of voters share a fixed effect of common ranking while some annotators might deviate from that significantly. Such parsimonious model from population to individual is a natural fit for crowdsourcing scenarios.

### 2.2 Linearized Bregman Iteration (LBI)

Linearized Bregman Iteration (LBI) was firstly introduced in [29] in the literature of variational imaging and compressive sensing. It is well-known that LASSO estimators are always biased [10]. On the other hand, [18] notices that Bregman iteration may reduce bias, also known as contrast loss, in the context of Total Variation image denoising. Now LBI can be viewed as a discretization of differential equations (inclusions), called *Inverse Scale Spaces*, which may produce unbiased estimators under nearly the same model selection consistency conditions as LASSO [19].

Beyond such a theoretical attraction, LBI is an extremely simple algorithm which combines an iterative gradient descent algorithm together with a soft thresholding. It is different to the well-known iterative soft-thresholding algorithm (ISTA) (e.g., [2], [9] and references therein) which converges to the biased LASSO solution. To tune the regularization parameter in noisy settings, one needs to run ISTA with many different thresholding parameters and chooses the best among them; in contrast, LBI only runs in a single path and regularization is achieved by early stopping like boosting algorithms [19], which may save the computational cost greatly and thus suitable for large scale implementation, e.g., distributive computation [31].

### 3 METHODOLOGY

In this section, we systematically introduce the methodology for parsimonious mixed-effects HodgeRank estimation. Specifically, we first start from extending the HodgeRank to a linear mixed-effect model. Then we present a simple iterative algorithm called Linearized Bregman Iterations to generate paths of parsimonious models at different sparsity levels. Early stopping regularization is discussed in the end.

#### 3.1 Mixed-Effects HodgeRank on Graphs

In crowdsourced pairwise comparison experiments, Let  $V = \{1, 2, \dots, n\}$  be the set of nodes and  $E = \{(u, i, j) : i, j \in V, u \in U\}$  be the set of edges, where  $U$  is the set of all annotators. Suppose the pairwise ranking data is given as  $y : E \rightarrow \mathbb{R}$ .  $y_{ij}^u > 0$  means  $u$  prefers  $i$  to  $j$  and  $y_{ij}^u \leq 0$  otherwise. The magnitude of  $y_{ij}^u$  can represent the degree of preference and it varies in applications. The simplest setting is the binary choice, where

$$y_{ij}^u = \begin{cases} 1 & \text{if } u \text{ prefers } i \text{ to } j, \\ -1 & \text{otherwise.} \end{cases} \quad (1)$$

The general purpose of preference aggregation is to look for a global score  $\theta : V \rightarrow \mathbb{R}$  such that

$$\min_{\theta \in \mathbb{R}^{|V|}} \sum_{i,j,u} \omega_{ij}^u l(\theta_i - \theta_j, y_{ij}^u), \quad (2)$$

where  $l(x, y) : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  is a loss function,  $\omega_{ij}^u$  denotes the confidence weights on  $\{i, j\}$  made by rater  $u$  (for simplicity, assumed to be  $\omega_{ij}^u = 1$  for the existing voting data), and  $\theta_i$  ( $\theta_j$ ) represents the global ranking score of item  $i$  ( $j$ , respectively). In HodgeRank, one benefits from the use of square loss  $l(x, y) = (x - y)^2$  which leads to fast algorithms to find optimal global ranking  $\theta$ , which becomes one component of a general orthogonal decomposition of paired comparison data [15], i.e.

$$y = \text{global ranking} \oplus \text{cycles},$$

where the component *cycles* can be further decomposed into

$$\text{cycles} = \text{local cycles} \oplus \text{global cycles}.$$

Local cycles are triangular cycles, e.g.  $i \succ j \succ k \succ i$ ; while global cycles, also called harmonic cycles, are loops involving nodes more than three (e.g.  $i \succ j \succ k \succ \dots \succ i$ ) and typically traversing all nodes in the graph. These cycles may arise due to conflicts of interests in ranking data. Therefore to analyze the statistical models of cycles is crucial to understand the conflicts of interests.

In crowdsourcing scenarios, the conflicts of interests are mainly due to two kinds of sources: the multi-criteria adopted by different annotators when they compare items in  $V$ ; the abnormal behavior of annotators in the experiments, e.g. simply clicking one side of the pair when they got bored, tired, or distracted. To quantitatively characterize these effects, we propose the following model of cycles

$$\text{cycles} = \text{personalized ranking} + \text{position bias} + \text{noise}.$$

To be specific, together with the global ranking component in HodgeRank, we consider the following linear mixed effects model for annotator's pairwise ranking:

$$y_{ij}^u = (\theta_i + \delta_i^u) - (\theta_j + \delta_j^u) + \gamma^u + \varepsilon_{ij}^u, \quad (3)$$

where

- $\theta_i$  is the common global ranking score, as a fixed effect;
- $\delta_i^u$  is the annotator's preference deviation from the common ranking  $\theta_i$  such that  $\theta_i^u := \theta_i + \delta_i^u$  becomes annotator  $u$ 's personalized ranking score, as a random effect;
- $\gamma^u$  is an annotator's position bias, which captures the careless behavior by clicking one side during the comparisons;
- $\varepsilon_{ij}^u$  is the random noise which is assumed to be independent and identically distributed with zero mean and being bounded.

Putting in matrix form, (3) becomes

$$y = d\theta + X\beta + \varepsilon, \quad (4)$$

where  $d \in \mathbb{R}^{|E| \times |V|}$  satisfies  $d\theta(u, i, j) = \theta_i - \theta_j$ ,  $\beta = [\delta, \gamma] \in \mathbb{R}^{(|V|+1)|U|}$  and  $X = [D, A]$ , where  $D \in \mathbb{R}^{|E| \times |V| \times |U|}$  satisfies  $D\delta(u, i, j) = \delta_i^u - \delta_j^u$  and  $A \in \mathbb{R}^{|E| \times |U|}$  satisfies  $\gamma(u, i, j) = \gamma^u$ .

Here  $\theta$  is population-level parameter which indicates some common score on  $V$ . In crowdsourcing studies, as the data vary greatly across individual annotators, we thus allow each annotator to have personalized parameters  $\theta^u$ . These personalized parameters can be obtained by adding some random effects  $\delta^u$  to the population parameter  $\theta$ , representing individual deviations from the population behavior. Besides,  $\gamma^u$  measures an annotator's position bias, i.e.



the tendency of  $u$  always clicking one side in paired comparison experiments. Under the random design of pairwise comparison experiments, a candidate should be placed on the left or the right randomly, so the position should not affect the choice of a careful annotator. However, some annotator might get confused, tired or distracted in experiments, such that he/she always clicks one side during some periods in experiments. Such a type of position bias captures a kind of noise in data not included in the zero mean  $\varepsilon$  and may severely deteriorates the quality of data. The remainder  $\varepsilon_{ij}^u$  measures the random noise in sampling which is of zero mean and bounded (hence subgaussian).

### 3.2 Parsimonious Paths with Linearized Bregman Iteration

In crowdsourced preference aggregation scenarios with good controls, it is natural to assume a parsimonious model. In such a model, the majority of annotators carefully follows the common behavior governed by the fixed effect parameter  $\theta$ , while only a small set of annotators might have nonzero personalized deviations and abnormal behavior in position bias. This amounts to assume that parameter  $\delta^u$  to be group sparse, i.e.  $\delta_i^u$  vanishes for all  $i$  simultaneously, and  $\gamma^u$  to be sparse as well, i.e. zero for most of careful annotators. Such a sparsity pattern motivates us to consider the following penalty function with a mixture of LASSO ( $L_1$ ) penalty on  $\gamma$  and group LASSO penalty on  $\delta$ :

$$P(\beta) = \|\gamma\|_1 + \sum_u \|\delta^u\|_2, \beta = (\delta, \gamma). \quad (5)$$

*Remark 1:* Usually a normalization factor  $\sqrt{n}$  is used before a group lasso penalty  $\|\delta^u\|_2$ , where  $n$  is the group size of  $\delta^u$ . But here all the  $\delta^u$  have the same group size, and  $\|D^u\|_F = \sqrt{2}\|A^u\|_F$ , so the column norm of  $D^u$  is on average  $\frac{\sqrt{2}}{\sqrt{n}}$  times of  $\|A^u\|_F$ , this basically cancels out the factor  $\sqrt{n}$ . So here we just use this simple formula.

Following the square loss in HodgeRank, the Euclidean distance (mean square error) in  $R^E$  can be used for the total loss function:

$$L(\theta, \beta) = \frac{1}{2m} \|y - d\theta - X\beta\|_2^2. \quad (6)$$

The following Linearized Bregman Iterations (LBI) give rise to a sequence of parsimonious (sparse) models:

$$\theta^{k+1} = \theta^k - \alpha \kappa \nabla_{\theta} L(\theta^k, \beta^k) \quad (7a)$$

$$z^{k+1} = z^k - \alpha \nabla_{\beta} L(\theta^k, \beta^k), \quad (7b)$$

$$\beta^{k+1} = \kappa \cdot \text{prox}_P(z^{k+1}), \quad (7c)$$

where  $\beta^0 = 0$ ,  $\theta^0 = \arg \min_{\theta} L(\theta, 0)$ , and variable  $z$  is an auxiliary parameter used for gradient descent,

$z = \rho + \beta/\kappa$ ,  $\rho \in \partial P(\beta)$ . Besides, the proximal map associated with the penalty function  $P$  is given by

$$\text{prox}_P(z) = \arg \min_{v \in R^{(|V|+1)|U|}} \left( \frac{1}{2} \|v - z\|^2 + P(z) \right).$$

The Linearized Bregman Iteration (7) generates a path of global ranking score estimators  $\theta^k$  and sparse estimators for preference deviation and position bias,  $\beta^k = (\delta^k, \gamma^k)$ . It starts from the common HodgeRank as  $\theta^0 = \arg \min_{\theta} L(\theta, 0)$ , and evolves into parsimonious mixed effect models with different levels of sparsity until the full model, often overfitted. To avoid the overfitting, early stopping regularization is required to find an optimal tradeoff between the model complexity and in-sample error. For more details, we refer the readers to see [19] and references therein. In this paper, we find that cross validation works to find the early stopping time that will be discussed in Sec. 3.4.

The Linearized Bregman algorithm was firstly introduced in [18] as a scalable algorithm for large scale image restoration with TV-regularization. It has several advantages than the widely used LASSO-type convex regularizations. First of all, it is simpler than LASSO in generating the sparse regularization paths: instead of a parallel run of several optimization problem over a grid of regularization parameters, a single run of LBI generates the whole regularization path. LBI is thus desired in dealing with big problems. Moreover, it can be less biased than LASSO as if nonconvex regularizations [11]. In fact, it is shown recently in [19] that as  $\kappa \rightarrow \infty$  and  $\alpha_t \rightarrow 0$ , the limit dynamics of Linearized Bregman Iterations in sparse linear regression with LASSO ( $L_1$ ) penalty may achieve the model selection consistency under nearly the same condition as LASSO yet return the unbiased Oracle estimator, while the LASSO estimator is well-known biased.

Here we give some remarks on the implementation details of the Linearized Bregman Iterations (7).

- The parameter  $\kappa$  determines the bias of the sparse estimators, a bigger  $\kappa$  leading to the less biased ones. The parameter  $\alpha$  is the step size which determines the precise of the path, with a large  $\alpha$  rapidly traversing a coarse grained path. However one has to keep  $\alpha\kappa$  small to avoid possible oscillations of the paths, e.g.  $\alpha\kappa \|X^T X\|_2/m < 2$ . The default choice in this paper is  $\alpha = \frac{m}{\kappa \|X^T X\|_2}$  as a tradeoff between performance and computation cost.

- The step (7a) can also be replaced by

$$\theta^{k+1} = \arg \min_{\theta} L(\theta, \beta^k)$$

if it is easy to solve.

- Now we turn to simplify the third step (7c) with an explicit formula for the proximal map with the particular penalty function defined in Eq.

(5). Recovering  $\beta^{k+1}$  from  $z^{k+1}$  is equivalent to the following group shrinkage on each group component of  $\beta$ , i.e.  $\gamma^u$  and  $\delta^u$ :

$$\begin{aligned} \beta^{k+1} &= \kappa \text{Shrinkage}(z^{k+1}) \\ &\triangleq \begin{cases} \delta^{u,k+1} = \kappa \max(0, 1 - 1/\|z_{\delta^u}\|_2) z_{\delta^u} \\ \gamma^{u,k+1} = \kappa \max(0, 1 - 1/|z_{\gamma^u}|) z_{\gamma^u} \end{cases} \end{aligned} \quad (8)$$

Now we are ready to give the following Linearized Bregman Algorithm for our Mixed-Effects HodgeRank as

---

**Algorithm 1** LBI for ME-HodgeRank

---

**Input:** Data  $(d, X, y)$ , damping factor  $\kappa$ , step size  $\alpha$ .

**Initialize:**  $\beta^0 = 0, \theta^0 = (d^T d)^{-1} d^T y, z^0 = 0, t^0 = 0$ .

**for**  $k = 0, \dots, K$  **do**

- 1)  $\theta^{k+1} = (d^T d)^{-1} d^T (y - X\beta^k)$ .
- 2)  $z^{k+1} = z^k + \frac{\alpha}{m} X^T (y - d\theta^k - X\beta^k)$ .
- 3)  $\beta^{k+1} = \kappa \text{Shrinkage}(z^{k+1})$
- 4)  $t^{k+1} = (k+1)\alpha$ .

**end for**

**Output:** Solution path  $\{t^k, \theta^k, \beta^k\}_{k=0,1,\dots,K}$ .

---

### 3.3 General Linear Model for Binary Data

For binary data, the linear model (3) may be not a proper interpreter, even though it still works in many datasets. A more interpretable model is General Linear Model (GLM) for binary data:

$$\begin{aligned} P(y_{ij}^u = 1) &= 1 - P(y_{ij}^u = -1) \\ &= \Phi((\theta_i + \delta_i^u) - (\theta_j + \delta_j^u) + \gamma^u) \end{aligned} \quad (9)$$

(9) where  $\Phi(t)$  is a symmetric cumulative distribution function (CDF). When  $\Phi(t) = \frac{1}{1+e^{-t}}$ , it becomes Bradley-Terry model and  $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-x^2/2} dx$  indicates Thurstone-Mosteller model.

Under GLM assumption, a natural estimation method is to minimize the negative log-likelihood:

$$L(\theta, \beta) = -\frac{1}{m} \sum_{i,j,u} \log \Phi(y_{ij}^u ((\theta_i + \delta_i^u) - (\theta_j + \delta_j^u) + \gamma^u)). \quad (10)$$

Here we use the fact  $\Phi(-t) = 1 - \Phi(t)$ . Plug it in (7), the corresponding LBI can be used to estimate a sequence of parsimonious sparse models for GLM.

Alg 2 is the detailed algorithm, where  $*$  and  $/$  means entry-wise multiplication/division and  $\phi(t) = \Phi'(t)$  is the probability distribution function corresponding to  $\Phi(t)$ . Here  $\phi(t) = \frac{e^t}{(1+e^t)^2}$  corresponds to Bradley-Terry model and  $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$  for Thurstone-Mosteller model.

### 3.4 Early Stopping Regularization

The Alg.1 actually returns a solution path with many estimators of different sparsity. So we need to find

---

**Algorithm 2** LBI for ME-GLM

---

**Input:** Data  $(d, X, y)$ , damping factor  $\kappa$ , step size  $\alpha$ .

**Initialize:**  $\beta^0 = 0, \theta^0 = 0, z^0 = 0, t^0 = 0$ .

**for**  $k = 0, \dots, K$  **do**

- 1)  $r^k = y * (d\theta^k + X\beta^k)$ .
- 2)  $\theta^{k+1} = \theta^k + \frac{\alpha\kappa}{m} d^T (\phi(r^k) * y / \Phi(r^k))$ .
- 3)  $z^{k+1} = z^k + \frac{\alpha}{m} X^T (\phi(r^k) * y / \Phi(r^k))$ .
- 4)  $\beta^{k+1} = \kappa \text{Shrinkage}(z^{k+1})$
- 5)  $t^{k+1} = (k+1)\alpha$ .

**end for**

**Output:** Solution path  $\{t^k, \theta^k, \beta^k\}_{k=0,1,\dots,K}$ .

---

an optimal stopping time among  $t^k = \alpha k$  to choose some best estimators and avoid overfitting. Here we sketch the procedure of cross-validation to choose the optimal stopping time:

- Given the training data, fix  $\kappa$  and  $\alpha$ , then split the data into  $K$  folds. Then choose a list of parameter  $t$ .
- **for**  $k = 1, \dots, K$  **do**
  - 1) Run Alg.1 on the training data except  $k$ -th fold to get the solution path.
  - 2) For pre-decided parameter list of  $t$ , use a linear interpolation to get  $(\theta(t), \beta(t))$ .
  - 3) On the  $k$ -th fold of training data, use the estimator  $(\theta(t), \beta(t))$  to predict, and then compute prediction error.
- end for**
- Return the optimal  $t_{cv}$  with minimal average prediction error.

**Remark:** Because the Alg.1 only returns the estimator at discrete  $\{t^k\}$  and may not contain the pre-decided parameter  $t$ , so we use a linear interpolation of the nearest two estimator  $(\theta^k, z^k)$  and  $(\theta^{k+1}, z^{k+1})$  to approximate  $(\theta(t), z(t))$ .  $\beta(t)$  is further obtained by using  $\text{Shrinkage}(z(t))$ .

## 4 EXPERIMENTS

In this section, four examples are exhibited with both simulated and real-world data to illustrate the validity of the analysis above and applications of the methodology proposed. The first example is with simulated data while the latter three exploit real-world data collected by crowdsourcing.

### 4.1 Simulated Study

**Settings** We validate the proposed algorithm on simulated data labeled by 100 annotators. Specifically, we first generate the true  $\theta_i \sim \mathcal{N}(0, 1)$ . Then each annotator has a probability  $p_1 = 0.4$  having a nonzero  $\gamma^u$  and a probability  $p_2 = 0.4$  having a nonzero  $\delta^u$ . Those nonzero  $\gamma^u$  is drawn randomly from  $\mathcal{N}(0, 2^2)$ . And each entry  $\delta_i^u$  of nonzero  $\delta^u$  is drawn randomly from  $\mathcal{N}(0, s^2)$  with  $s \sim \mathcal{U}(0, 3)$ . At last, we draw  $N^u$

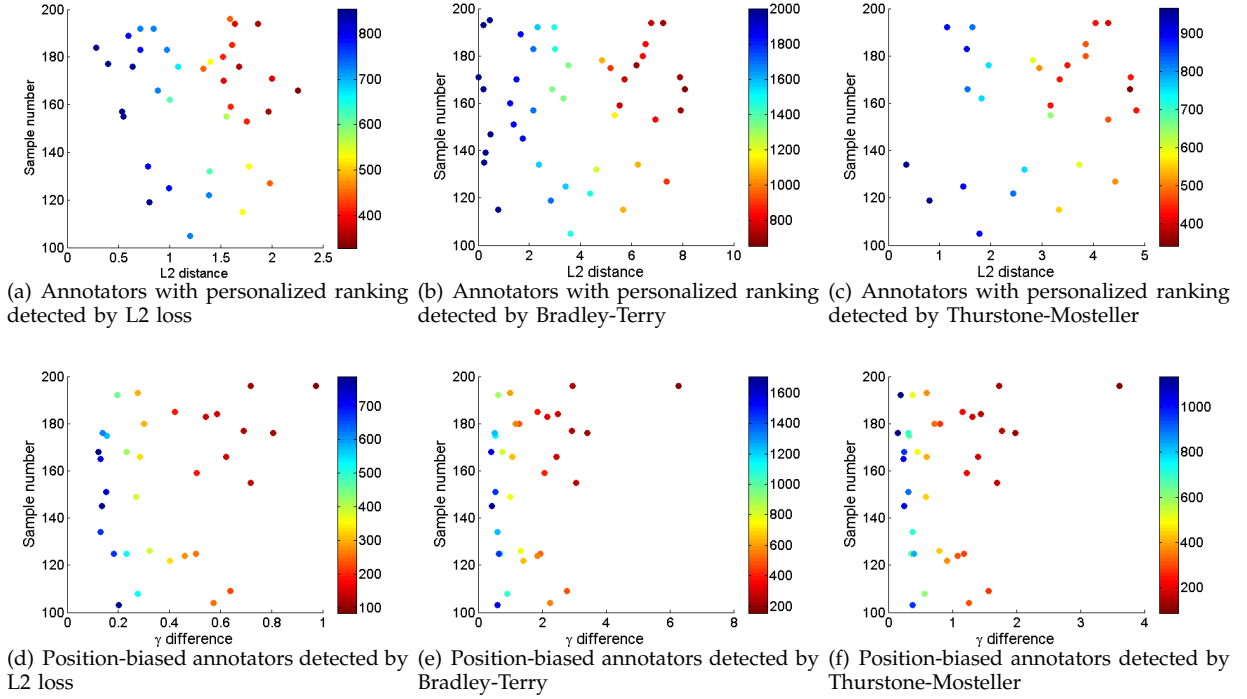


Fig. 2: Annotators with personalized ranking and position bias detected in simulated data. YY: what's the meaning of color bar? The time  $t$  that a path becomes nonzero?

samples for each user randomly with binary response  $y_{ij}^u$  following the model  $P(y_{ij}^u = 1) = \Phi((\theta_i + \delta_i^u) - (\theta_j + \delta_j^u) + \gamma^u)$ , where  $\Phi(t) = 1/(1 + e^{-t})$  (YY: so Bradley-Terry model is used here to generate the data?). The sample number  $N^u$  uniformly spans in  $[N_1, N_2] = [100, 200]$ . Here we choose  $n = |V| = 16$ , which is consistent with the first real-world dataset. Finally, we obtain a multi-edge graph with 15,067 pairwise comparisons annotated by 100 annotators.

**Results** Fig.2(a), 2(b), and 2(c) show the annotators exhibiting personalized preferences selected via three loss cases, where each color dot represents one annotator. The optimal  $t$  obtained via cross-validation is shown in Fig.3. The scores derived from each user are denoted as  $\hat{\theta}^u$  and the common ranking as  $\hat{\theta}$ . The X-axis represents the  $L_2$ -distance between each user's personalized ranking and the common ranking,  $\|\hat{\theta}^u - \hat{\theta}\|$ . The Y-axis counts the number of pairwise comparisons each user provides. Clearly one can see the larger the  $L_2$ -distance and sample size, the more earlier this user is treated as one with personalized ranking (from color red to blue). This indicates that users jumped out earlier are those with large deviation from the population's opinion and a big sample size indicating a high confidence. Similar results of position-biased user are illustrated in Fig.2(d), 2(e), and 2(f) in which the X-axis ( $\gamma$  difference) represents the absolute difference of  $\gamma$  between each user and the population.

Finally, to see whether our proposed method in

TABLE 1: HodgeRank vs. Mixed-effects model on test error (i.e. mismatch ratio) in simulated data.

	min	mean	max	std
HodgeRank	0.2573	0.2645	0.2760	0.0051
L2 Loss	0.1902	0.1983	0.2063	0.0044
Bradley-Terry	0.1930	0.1980	0.2053	0.0034
Thurstone-Mosteller	0.1912	0.1981	0.2056	0.0042

three loss cases could provide more precise preference function for users by introducing individual-specific parameters (i.e.,  $\delta$  and  $\gamma$ ), we split the data into training set (70% of the total pairwise comparisons) and testing set (the remaining 30%). To ensure the statistical stability, we repeat this procedure 20 times. Tab.1 shows the experimental results of the proposed mixed-effects model compared with HodgeRank, which indicates that our method exhibits smaller test error (i.e. mismatch ratio) due to its parsimonious property. Besides, it is worth mentioning that L2 loss could exhibit similar performance with Bradley-Terry and Thurstone-Mosteller models which are particular suitable for binary data in most cases. (YY: BT performs the best because we use BT model to generate the data?)

## 4.2 Human Age

**Settings** In this dataset, 30 images from human age dataset FG-NET<sup>1</sup> are annotated by a group of volun-

1. <http://www.fgnet.rsunit.com/>

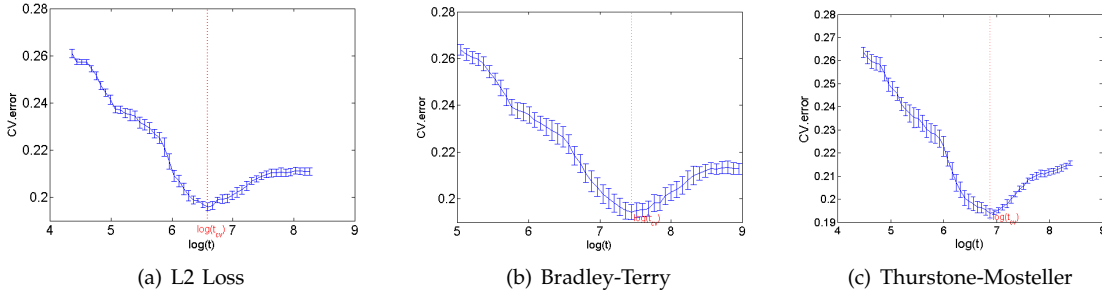
Fig. 3: Optimal  $t$  with minimal average prediction error in simulated data.

Fig. 4: Images in Human age dataset.

TABLE 2: HodgeRank vs. Mixed-effects model in Human age dataset.

	min	mean	max	std
HodgeRank	0.1997	0.2104	0.2202	0.0050
L2 Loss	0.1478	<b>0.1569</b>	0.1653	0.0055
Bradley-Terry	0.1191	<b>0.1255</b>	0.1341	0.0046
Thurstone-Mosteller	0.1214	<b>0.1283</b>	0.1356	0.0042

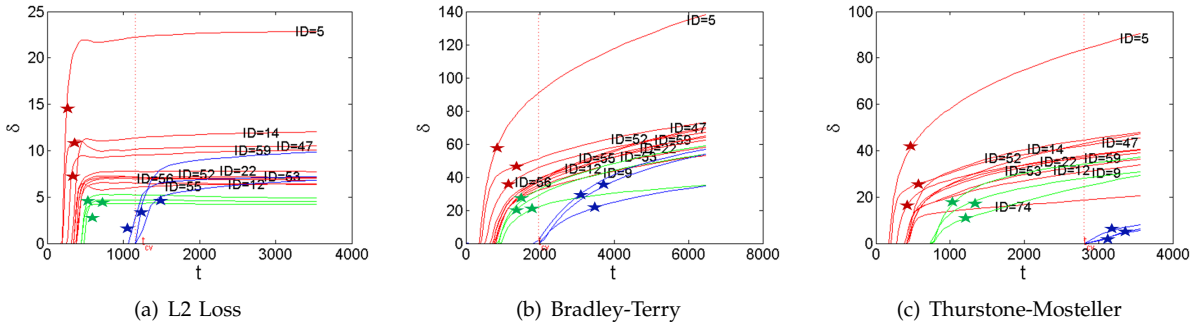


Fig. 5: Top 10 (marked with red color) annotators with personalized ranking in Human age dataset.

teer users on **ChinaCrowds** platform, as is illustrated in Fig.4. The annotator is presented with two images and given a binary choice of which one is older. Totally, we obtain 14,011 pairwise comparisons from 94 annotators.

**Results** Tab.2 shows the mean test error (70% data for training, 30% for testing) results of 20 times achieved by this scheme. It is shown that consistent with the simulated data, in this dataset, the mixed-effects model with three losses could also provide better approximate results of the annotators' preference than the HodgeRank estimator.

To further investigate the characteristics of anno-

tators with personalized ranking, Fig.5 illustrates annotator's LBI regularization paths of preference deviations with optimal  $t$  (i.e.,  $t_{cv}$ ) returned by cross-validation in three losses. The red curves in Fig.5 represent the top 10 annotators who jumped out early. It is easy to find that results returned by three losses are actually very similar except one or two differences. Besides, similar to the simulated data, annotators jumped out earlier are those with a large deviation ( $L_2$ -distance) from the common ranking and a big sample size showing high confidence of such deviations. Moreover, Fig.?? shows the order comparisons of common ranking (i.e., com.) and personalized ranking of 9 representative annotators at  $t_{cv}$ . The X-



axis represents user index: user = 2, 3, 4 jumped out early corresponding to paths labeled with red stars in Fig.5; user = 5, 6, 7 jumped out in the middle time corresponding to green stars; user = 8, 9, 10 jumped out late corresponding to blue stars. The order of faces in Y-axis is arranged from lower to higher (i.e., from color blue to red) according to the common ranking score calculated by our method. The color represents the ranking position returned by the corresponding user. It is easy to see users jumped out late exhibit almost consistent ranking order with the common ranking, while the earlier ones are almost the adversarial against the common. A subset of this has been shown in Fig.1 in introduction.

Moreover, Fig.?? illustrates the LBI regularization paths of annotator's position bias with red lines represent the top 10 annotators. It is easy to see that the corresponding results returned from these three loss functions are exactly the same. Tab.3 further shows the click counts of each side (i.e., Left and Right) for these top 10 position-biased annotators. It is easy to see that these annotators can be divided into two types: (1) click one side all the time (with ID in blue); (2) click one side with high probability (others). Although it might be relatively easy to identify the annotators of type (1) above by inspecting their inputs, it is impossible for eye inspection to pick up those annotators of type (2) with mixed rational and abnormal behaviors. Therefore it is essential to design such a statistical methodology to quantitatively detect these kind of position-biased annotators for crowdsourcing platforms in market.

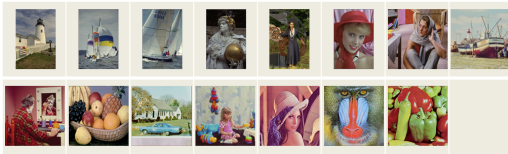


Fig. 8: Images in LIVE and IVC dataset.

### 4.3 Image Quality Assessment (IQA)

**Settings** Two publicly available datasets, LIVE [25] and IVC [16], are used in this work. The LIVE dataset contains 29 reference images and 779 distorted images. The distorted images are obtained using five different distortion processes—JPEG2000, JPEG, White Noise, Gaussian Blur, and Fast Fading Rayleigh. Considering the resolution limit of most test computers, we only choose 6 different reference images ( $480 \times 720$ ) and 15 distorted versions of each reference, for a total of 96 images. The second dataset, IVC, which is also a broadly adopted dataset in the community of multimedia quality evaluation, includes 10 reference images and 185 distorted images derived from four distortion types—JPEG2000, JPEG, LAR Coding, and

TABLE 4: HodgeRank vs. Mixed-effects model in IQA dataset (reference image 1).

	min	mean	max	std
HodgeRank	0.1678	0.1874	0.2019	0.0081
L2 Loss	0.1002	0.1094	0.1239	0.0065
Bradley-Terry	0.0638	0.0708	0.0817	0.0054
Thurstone-Mosteller	0.0625	0.0731	0.0825	0.0057

Blurring. Following the collection strategy in LIVE, we further select 9 different reference images ( $512 \times 512$ ) and 15 distorted images of each reference. Totally, we obtain a medium-sized image set that contains a total of 240 images from 15 references, as illustrated in Fig.8. Finally, 342 observers of different cultural background, each of whom performs a varied number of comparisons via Internet, provide 52,043 paired comparisons in total. The number of responses each reference image receives is different.

TABLE 5: Top 10 position-biased annotators in IQA dataset (reference image 1).

Order	ID	Left	Right	Order	ID	Left	Right
1	259	96	0	6	2	55	0
2	334	90	0	7	260	49	2
3	177	77	0	8	23	42	0
4	103	74	4	9	207	46	2
5	29	58	0	10	287	34	0

To validate whether the annotators' preference function we estimated is good enough, we randomly take reference image 1 as an illustrative example while other reference images exhibit similar results.

**Results** In terms of prediction performance, it can be seen that the mixed-effects model with three losses is more accurate than HodgeRank by significantly reducing the mean test error, as is shown in Tab.4.

Besides, Fig.?? shows the regularization paths of personalized ranking with top 10 annotators (red curves) selected early in the paths. Similar to the Age dataset, the common ranking vs. personalized ranking results of 9 representative users is shown in Fig.??, corresponding to the red/green/blue stars in Fig.??.

**Note:** Here, these exists large difference among the results returned by three losses, and we need more elegant words the explain this phenomenon.....It is easy to see that among the 10 annotators returned by L2 loss, 9 of them (except annotator with ID = 133) click one side all the time (i.e., position-biased annotators), while results returned by other two are not. From this viewpoint, maybe L2 loss is indeed less suitable for binary data on some specific cases compared with other two. Or, should we explain this phenomenon from a totally new perspective? add a remark here?

Moreover, the LBI regularization paths of position bias  $\gamma$  and click counts of top 10 annotators in this dataset are shown in Fig.?? and Tab.5. It is easy to see that annotators picked out under three losses are also exactly the same, which are mainly those clicking on one side almost all the time. Besides, it is interesting

to see that all these annotators highlighted with blue color in Tab.5 click the left side all the time. We then go back to the crowdsourcing platform and find out that the reason behind this is a *default choice* on the left button, as is shown in Fig.??, which induces some lazy annotators to cheat for the task.

#### 4.4 WorldCollege Ranking

**Settings** We now apply the proposed method to the WorldCollege dataset, which is composed of 261 colleges. Using the **Allourideas** crowdsourcing platform, a total of 340 distinct annotators from various countries (e.g., USA, Canada, Spain, France, Japan) are shown randomly with pairs of these colleges, and asked to decide which of the two universities is more attractive to attend. Finally, we obtain a total of 8,823 pairwise comparisons. The world map of all the annotators is illustrated in Fig.13.

**Results** We apply the proposed method to the resulting dataset and find out that, similar to the simulation and other two real-world datasets, the mixed-effects model could produce better performance than HodgeRank with smaller mean test error, shown in Tab.6. Noting that in this dataset, only 9 annotators are treated as annotators with distinct personalized rankings at optimal  $t$  (i.e.,  $t_{cv}$ ) selected via cross-validation in L2 loss case, as is shown in Fig.?. The reason why others with relative big  $\delta$  are not detected out lies in the small sample size they provide indicating high variances. (YY: so other losses did not detect any personalized deviation? The explanation here seems not clear to me.) To better illustrate comparison result of three losses, for other two, we also only pick up the top 9 annotators, as is shown in Fig.?? and ?. It is pleasing to see that the top 9 annotators returned by three losses are exactly the same in this dataset. The common ranking vs. personalized ranking of 9 representative users is shown in Fig.?? with a similar observation to the other two datasets. Besides, the regularization paths of position bias and click counts of top 10 annotators in this dataset are shown in Fig.?? and Tab.7. It is easy to see that similar to the human age dataset, these annotators are either clicking one side all the time, or clicking one side with high probability in mixed behaviors. Clearly, there is only one difference among these three cases, where L2 loss and Thurstone-Mosteller pick out annotator with ID=115, while Bradley-Terry treats annotator with ID=245 as position-biased one. A further inspection of the dataset confirms that such a detection result is reasonable, as the ratio of left/right clicks of these two annotators are 34:0 and 0:34 respectively, as is shown in Tab.7. (ID=115 and ID=245 are both reasonable? Why not finding them both?)

## 5 CONCLUSIONS

In this paper, we propose a parsimonious mixed-effects model based on HodgeRank to learn user's



Fig. 13: World map of all the annotators in WorldCollege dataset.

TABLE 6: HodgeRank vs. Mixed-effects model in WorldCollege ranking dataset.

	min	mean	max	std
HodgeRank	0.2979	0.3108	0.3230	0.0097
L2 Loss	0.2530	0.2670	0.2757	0.0071
Bradley-Terry	0.2456	0.2555	0.2678	0.0099
Thurstone-Mosteller	0.2553	0.2616	0.2696	0.0067

preference or utility function in crowdsourced ranking, which takes into account both the personalized preference deviations from the common and position biases of the annotators. To be specific, common preference scores indicate the consistent ranking on population-level which approximates the behavior of all users, while a small set of annotators might have nonzero personalized deviations and abnormal behavior in position bias. Equipped with the newly developed Linearized Bregman Iteration, which is a simple iterative procedure generating a sequence of parsimonious models, we establish a dynamic path from the common utility to individual variations, with different levels of parsimony or sparsity on personalization. In this dynamic scheme, three kinds of losses are systematically discussed, including L2 loss, Bradley-Terry, Thurstone-Mosteller. Experimental studies conducted on simulated examples and real-world datasets show that our proposed method could exhibit better performance (i.e. smaller test error) compared with HodgeRank. Our results suggest that the proposed methodology is an effective tool to investigate the diversity in annotator's behavior in modern crowdsourcing data.

## REFERENCES

- [1] K. Arrow. *Social Choice and Individual Values*, 2nd Ed. Yale University Press, New Haven, CT, 1963.
- [2] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.
- [3] O. Candogan, I. Menache, A. Ozdaglar, and P. A. Parrilo. Flows and decompositions of games: Harmonic and potential games. *Mathematics of Operations Research*, 36(3):474–503, 2011.
- [4] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei. A crowd-sourceable QoE evaluation framework for multimedia content. pages 491–500. ACM Multimedia, 2009.
- [5] A. Chorin and J. Marsden. *A Mathematical Introduction to Fluid Mechanics*. Texts in Applied Mathematics. Springer, 1993.
- [6] D. Cynthia, K. Ravi, N. Moni, and S. Dandapani. Rank aggregation methods for the web. In *International Conference on World Wide Web*, pages 613–622, 2001.
- [7] R. L. Day. Position bias in paired product tests. *Journal of Marketing Research*, 6(1):98–100, 1969.

- [8] J. de Borda. *Mémoire sur les Elections au Scrutin*. Histoire de l'Académie Royale des Sciences, 1781.
- [9] D. L. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627, 1995.
- [10] J. Fan and R. L. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, 96(456):1348–1360, 2001.
- [11] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, pages 1348–1360, 2001.
- [12] Y. Fu, T. M. Hospedales, T. Xiang, S. Gong, and Y. Yao. Interestingness prediction by robust learning to rank. In *European Conference on Computer Vision*, volume 8690, pages 488–503. 2014.
- [13] Y. Fu, T. M. Hospedales, T. Xiang, J. Xiong, S. Gong, Y. Wang, and Y. Yao. Robust subjective visual property prediction from crowdsourced pairwise labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, in press, 2016.
- [14] K. G. Jamieson and R. D. Nowak. Active ranking using pairwise comparisons. *Annual Conference on Neural Information Processing Systems*, pages 2240–2248, 2011.
- [15] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127(6):203–244, 2011.
- [16] P. Le Callet and F. Autrusseau. Subjective quality assessment irccyn/ivc database, 2005. <http://www.irccyn.ec-nantes.fr/ivcdb/>.
- [17] S. Negahban, S. Oh, and D. Shah. Iterative ranking from pairwise comparisons. In *Annual Conference on Neural Information Processing Systems*, pages 2483–2491, 2012.
- [18] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin. An iterative regularization method for total variation-based image restoration. *SIAM Journal on Multiscale Modeling and Simulation*, 4(2):460–489, 2005.
- [19] S. Osher, F. Ruan, J. Xiong, Y. Yao, and W. Yin. Sparse recovery via differential inclusions. *Applied and Computational Harmonic Analysis*, in press, 2016.
- [20] B. Ousting, C. Brune, and S. Osher. Enhanced statistical rankings via targeted data collection. In *International Conference on Machine Learning*, pages 489–497, 2013.
- [21] B. Ousting, J. Darbon, and S. Osher. Statistical ranking using the  $l_1$ -norm on graphs. *Inverse Problems & Imaging*, 7(3), 2013.
- [22] A. Rajkumar and S. Agarwal. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *International Conference on Machine Learning*, pages 118–126, 2014.
- [23] J. D. Rennie and N. Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *International conference on Machine learning*, pages 713–719, 2005.
- [24] R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *International conference on Machine learning*, pages 880–887, 2008.
- [25] H. Sheikh, Z. Wang, L. Cormack, and A. Bovik. LIVE image & video quality assessment database, 2008.
- [26] 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.
- [27] Q. Xu, J. Xiong, X. Cao, and Y. Yao. Parsimonious mixed-effects HodgeRank for crowdsourced preference aggregation. page preprint. ACM Multimedia, 2016.
- [28] J. Yi, R. Jin, S. Jain, and A. Jain. Inferring users' preferences from crowdsourced pairwise comparisons: A matrix completion approach. In *AAAI Conference on Human Computation and Crowdsourcing*, 2013.
- [29] W. Yin, S. Osher, J. Darbon, and D. Goldfarb. Bregman iterative algorithms for compressed sensing and related problems. *SIAM Journal on Imaging Sciences*, 1(1):143–168, 2008.
- [30] J. Yuan, G. Steidl, and C. Schnorr. Convex Hodge decomposition and regularization of image flows. *Journal of Mathematical Imaging and Vision*, 33(2):169–177, 2009.
- [31] K. Yuan, Q. Ling, W. Yin, and A. Ribeiro. A Linearized Bregman algorithm for decentralized basis pursuit. *European Signal Processing Conference*, pages 1–5, 2013.



Qianqian Xu received the B.S. degree in computer science from China University of Mining and Technology in 2007 and the Ph.D. degree in computer science from University of Chinese Academy of Sciences in 2013. She is currently an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. Her research interests include statistical machine learning, with applications in multimedia and computer vision. She has authored or coauthored 10+ academic papers in prestigious international journals and conferences, among which she has published 4 full papers with the first author's identity in ACM Multimedia (acceptance rate 17%). She served as member of professional committee of CAAI, and member of online program committee of VALSE, etc.



Jiechao Xiong is currently pursuing the Ph.D. degree in statistics at the School of Mathematical Science, Peking University, Beijing, China. His research interests include statistical learning, data science and topological and geometric methods for high-dimension data analysis.



Xiaochun Cao, Professor of the Institute of Information Engineering, Chinese Academy of Sciences. He received the B.E. and M.E. degrees both in computer science from Beihang University (BUAA), China, and the Ph.D. degree in computer science from the University of Central Florida, USA, with his dissertation nominated for the university level Outstanding Dissertation Award. After graduation, he spent about three years at ObjectVideo Inc. as a Research Scientist. From 2008 to 2012, he was a professor at Tianjin University. He has authored and coauthored over 100 journal and conference papers. In 2004 and 2010, he was the recipients of the Piero Zamperoni best student paper award at the International Conference on Pattern Recognition. He is a fellow of IET and a Senior Member of IEEE. He is an associate editor of IEEE Transactions on Image Processing.



Yuan Yao received the B.S.E and M.S.E in control engineering both from Harbin Institute of Technology, China, in 1996 and 1998, respectively, M.Phil in mathematics from City University of Hong Kong in 2002, and Ph.D. in mathematics from the University of California, Berkeley, in 2006. Since then he has been with Stanford University and in 2009, he joined the School of Mathematical Sciences, Peking University, Beijing, China, as a professor of statistics in the Hundred Talents Program. His current research interests include topological and geometric methods for high dimensional data analysis and statistical machine learning, with applications in computational biology, computer vision, and information retrieval. Dr. Yao is a member of American Mathematical Society (AMS), Association for Computing Machinery (ACM), Institute of Mathematical Statistics (IMS), and Society for Industrial and Applied Mathematics (SIAM). He served as area or session chair in NIPS and ICIAM, as well as a reviewer of Foundation of Computational Mathematics, IEEE Trans. Information Theory, J. Machine Learning Research, and Neural Computation, etc.



TABLE 3: Top 10 position-biased annotators in Human age dataset, together with the click counts of each side (i.e., Left and Right).

Order	ID	Left	Right	Order	ID	Left	Right
1	12	90	270	6	91	79	5
2	70	191	9	7	51	63	0
3	59	213	66	8	50	60	3
4	38	110	15	9	18	74	25
5	43	79	1	10	40	40	0

TABLE 7: Top 10 position-biased annotators in World-College ranking dataset.

Order	ID	Left	Right	Order	ID	Left	Right
1	268	148	0	6	270	20	70
2	209	127	0	7	267	45	0
3	156	189	67	8	276	16	54
4	320	253	324	9	166	35	0
5	87	11	62	10	115	34	0
				10	245	0	34