Improving Peer Assessment Accuracy by Incorporating Relative Peer Grades

Tianqi Wang
University at Buffalo
twang47@buffalo.edu

Qi Li
University of Illinois
qli22@buffalo.edu

Jing Gao
University at Buffalo
jing@buffalo.edu

Xia Jing
Tsinghua University
jingxyy@qq.com

Jie Tang
Tsinghua University
jery.tang@gmail.com

ABSTRACT

Massive Open Online Courses (MOOCs) have become more and more popular recently. These courses have attracted a large number of students world-wide. In a popular course, there may be thousands of students. Such a large number of students in one course makes it infeasible for the instructors to grade all the submissions. Peer assessment is thus an effective paradigm that can help grade the submissions at a large scale. However, due to the variance in the ability and standard of the student graders, peer grades may be noisy and biased. Aggregating peer grades to have an accurate and fair final grade for a submission is a challenging problem because the reliability and bias degrees of graders are usually unknown in practice. To address this issue, some probabilistic models considering the graders' reliability and bias are proposed. However, due to the sparsity of peer grade observations, it is difficult for these models to estimate the accurate reliability and bias of the graders as well as the true grades of the submissions. Compared with absolute peer grades, the relative peer grades, derived from the difference between the peer grades of two submissions graded by the same grader, are less sparse and more robust to the grader's bias. Thus relative peer grades are informative and helpful in cardinal peer grading estimation whose goal is to estimate the absolute numeric grades of submissions. In this paper, we propose two new probabilistic models to help improve the accuracy of cardinal peer grading estimation using the observed relative grades among submissions. In this way, the relation between the true grades among submissions is taken into consideration when deriving the final grades. Experimental results on real MOOC peer grading datasets show that the proposed models outperform baselines and the relation of true grades among submissions indeed contributes to the improvement in the grade estimation.

Keywords

Peer grading, relative peer grades, MOOCs

Tianqi Wang, Qi Li, Jing Gao, Xia Jing and Jie Tang "Improving Peer Assessment Accuracy by Incorporating Relative Peer Grades" In: *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, Collin F. Lynch, Agathe Merceron, Michel Desmarais, & Roger Nkambou (eds.) 2019, pp. 450 - 455

1. INTRODUCTION

Massive Open Online Courses (MOOCs) have provided millions of learners with open access to high quality courses via web. For a popular course, there may be thousands of students. Recently, several MOOC platforms offer verified certificates or even degree programs, and peer grading plays an important role in the student performance evaluation. The benefit of peer grading is two-folded. On one hand, it is helpful for the instructors to evaluate students performance, which is otherwise infeasible due to the large number of enrollment. On the other hand, it is also beneficial to the students: they can see peers' work from different aspects and increase their involvement in the course [5]. Especially, peer grading can be used when automatic grading cannot be applied, for example, on essays and projects. A typical process of peer assessment includes two steps: first, students are assigned to grade a subset of submissions and then the platform aggregates these peer grades to compute the final grades of these submissions.

Although peer grading is helpful, it is a challenging problem to aggregate these peer grades and determine the final grade of a submission. In this paper, we consider the case of cardinal peer grading (i.e., each submission receives a numerical grade as the final grade). Most platforms use the median of received peer grades as the final grade of a submission. However, the median grade may be inaccurate due to the different reliability and bias degrees of graders. Usually, the difference between the grade given by a grader and the true grade of the submission can be decomposed into bias and reliability degree. Suppose a grader grades multiple submissions, and then the bias represents the difference between the mean grades of this grader and the true grades on these submissions. The reliability degree of the grader is measured by the variance of the difference between the the grades that the grader gives and the true grades of these submissions. If a grader randomly assigns grades to the submissions, he/she is not a reliable grader. If the variance is small, then a grader grades the submission in a consistent way and is thus a reliable grader. It is important to consider the modeling of grader bias and reliability to derive more accurate estimates of the final grades. Therefore, there are some existing efforts towards this direction [7].

However, the mechanism of peer grading that each student only grades a small subset of submissions leads to a data sparsity issue. The sparsity of the observed grades makes these models difficult to correctly estimate reliability, bias

of the grader and the true grades of the submissions. In addition the observed grades are sensitive to the grader's bias. Compared with absolute observed grades, the relative peer grades between two submissions are less sparse and more robust to the grader's bias, since the relative peer grades are derived from the difference of the grades assigned by the same grader to two different submissions. Thus the relative peer grades are informative in estimating the true grades of submissions. However, all existing cardinal peer grading estimation models [7, 6, 2] only consider the absolute peer grades of each submission. None of these models considers the relative grades between two submissions.

Recognizing the importance of relative peer grades, we develop new probabilistic graphical models by leveraging relative peer grades between submissions to model the dependency between the true grades. The proposed probabilistic models estimate the true grades of submissions from the peer grades as well as relative peer grades by modeling the bias and reliability of graders. Gaussian distributions are applied to model the true grades, the bias of grader, the absolute peer grades, and relative peer grades in the proposed models. Two different distributions are proposed to estimate the reliability of the graders. In the first model, the reliability of the grader follows a Gamma distribution with the shape parameter determined by the grader's own true grade, while in the second model, it follows a Gaussian distribution with the mean equal to the grader's true grade. To evaluate the proposed models, experiments are conducted on peer grading datasets collected from a popular MOOC platform in China. Experimental results show that the proposed models improve the accuracy of the cardinal peer grading estimation by considering the dependency of true scores between two submissions. The main contributions of this paper are summarized as follows:

- We find that relative peer grades among submissions can help improve cardinal peer grading estimation accuracy.
- We propose new probabilistic graphical models by incorporating observed relative grades to model the dependency between the true grades of these two submissions.
- We evaluate the proposed models on real peer grading datasets and experimental results show that the proposed models can improve the accuracy of cardinal peer grading estimation.

2. RELATED WORK

Existing work on peer assessment aggregation can be divided into two categories based on the data types: the cardinal and ordinal peer grade estimation. The goal of ordinal peer grade estimation is to rank the students according to their submissions. Models based on pair comparison [10, 8], Bayesian generative approach [12] and matrix factorization are developed for the ordinal peer grades estimation [1].

For cardinal peer grading estimation, students are asked to grade their peers' submissions by assigning a specific numerical grade and the aim of cardinal grades estimation is to find the absolute true scores of the submissions. Below we summarize the existing work related to cardinal peer grading estimation respectively.

One major approach of cardinal peer grading estimation is to update grades and grader weights iteratively [4, 12, 3]. Another major category of methods are based on probabilistic graphical models [7, 6, 2]. The proposed models in this paper fall into this category. The main idea is to model the true grade of a submission, the reliability and bias of each grader as hidden random variables following certain distributions, and infer the model parameters by fitting the models on observed peer grades. In particular, the following methods [7, 6] (referred to as PG_1 to PG_5) are the most relevant to our proposed model. In [7], three probabilistic graphical models named PG_1, PG_2 and PG_3 are proposed. PG_1 is the basic model, which assumes that true grades, observed peer grades, and biases follow Gaussian distributions and the reliability of the grader follows a Gamma distribution. Upon PG_1, PG_2 links the bias of a grader among assignments, and PG_3 couples the grader's grade of his/her submission and the grader's reliability. In PG_3, the grader's reliability is modeled as a linear function of the grader's grade. To relax this assumption of linear relationship, two extensions of PG_3 referred as PG_4 and PG_5 are later proposed in [6]. Both PG_4 and PG_5 assume the reliability of a grader is related to the grader's own grade, and use either Gamma distribution or Gaussian distribution to model this reliability. Recently, social connections are also considered in the modeling of the dependencies of bias among students [2].

However, all existing cardinal peer grading estimation methods only consider absolute grades. In these methods, the true grades of different submissions are treated independently. None of these models takes the relative grades into consideration. In fact, leveraging the relative grades between submissions to model the dependency between true grades of these two submissions can help reduce the noise introduced by the bias of graders and alleviate the data sparsity issue, and thus can help to improve the accuracy of cardinal peer grading estimation. To the best of our knowledge, this is the first work that integrates relative grades into cardinal peer grading aggregation to achieve improved estimation.

3. PROBLEM DEFINITION

In this section, we first introduce some concepts and notations used in the rest of this paper. Then we formally define the problem.

The set of all the students is denoted as S and the set of all the graders is denoted as G. Under the peer grading setting, $G \subseteq S$, since the graders are students as well. The observed absolute grade (peer grade) of a submission submitted by student i graded by grader g is denoted as z_i^g, and the observed relative grades (relative peer grades) between submissions submitted by students i and j graded by grader g is denoted as d_{ij}^g. The relative peer grades are derived using absolute peer grades, which are the difference of the absolute peer grades. For example, if a grader g assigned a score of 4 to the submission submitted by student i and a score of 6 to the submission submitted by student j, then z_i^g is 4 and z_j^g is 6. We can derive that the relative grade $d_{ij}^g = z_i^g - z_j^g = 6 - 4 = 2$. The subset of students whose submissions are graded by an arbitrary grader $g \in G$ is described as S_g and the set of graders who assign grades to the submission submitted by student i is defined as G_i.

With these definitions introduced, we define the cardinal peer grading estimation problem as follows: Given a set of

students S, a set of graders G, a set of peer grades $\{z_i^g\}_{i \in S}^{g \in G}$ and relative peer grades $\{d_{ij}^g\}_{i,j \in S, i \neq j}^{g \in G}$, we want to estimate the true absolute grade for submission submitted by student i, $\forall i \in S$, and to learn the reliability and bias for each grader g, $\forall g \in G$.

4. METHODOLOGY

In this section, we describe our probabilistic graphical models named PG_6 and PG_7 for cardinal peer grading estimation. Both models specify a two-stage generation for the peer grades and relative peer grades. The first stage specifies the generation of graders' bias, reliability and true scores of submissions and the second stage generates the peer grades and relative peer grades given the grader's bias, reliability as well as the true scores of submissions.

True score generation: In the proposed models, the true score of the submission submitted by student i is modeled as a random variable following a Gaussian distribution.

Grader bias generation: The bias of grader g is denoted as b_g, which measures the constant grade inflation or deflation of a grader. We model the grader's bias as a random variable following a Gaussian distribution. Though different graders may have different bias, we can assume the average of all graders' bias is 0.

Grader reliability generation: The reliability of a grader reflects how consistent a grader assigns grades. A reliable grader keeps a stable bias when assigning grades to different submissions. Following the assumptions in [11], we assume that the reliability of a grader is related to his/her own grade, which reflects the grader's knowledge about the assignment. We assume that the grader with a higher grade of the assignment may be a more reliable grader for submissions of the same assignment. The reliability of a grader g is denoted as τ_g and modeled as a random variable following a Gamma distribution in the PG_6 model and a Gaussian distribution in the PG_7 model, respectively. In the PG_6 model the grader's true grade is used as the shape parameter of the Gamma distribution, while in the PG_7 model it is used as the mean value of the Gaussian distribution.

Peer grade generation: After generating the bias and reliability of graders as well as the true scores, the peer grades can be generated with these variables. The peer grade is modeled as a variable following a Gaussian distribution whose mean is the sum of the true grade of the submission and the bias of the grader, and its variance is inversely proportional to the reliability of the grader. In the PG_7 model, we introduce a hyper-parameter λ to tune the scale of the variance.

Relative peer grade generation: To incorporate more observations to estimate the reliability and bias of the grader and the true grade of the submission, the relative peer grade is generated. The generation of relative peer grade provides us with another view of true score of a submission in addition to the traditional way that models the true grade as the sum of observed peer grade and the bias of the grader. With the relative peer grade, the true grade s_i of submission i can be estimated by the sum of the true grade s_j of submission j and the relative peer grade between these two submissions. In such a way, the influence of grader bias is excluded.

Similarly to the generation process of peer grade, the relative peer grade is generated with the given true grades of two submissions and the reliability of the grader. We assume the relative peer grade follows a Gaussian distribution with mean value equal to the difference of the true grades between two submissions and variance inversely proportional to the grader's reliability. Also, in the PG_7 model, λ is used to specify the scale of the variance.

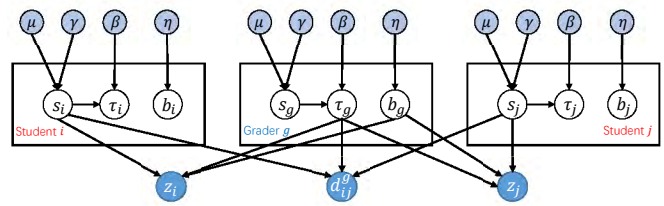

Figure 1: The plate notation of the PG_6 and PG_7 model.

Table 1: Notations

Notation	Description
S	set of all students
G	set of all graders
τ_g	reliability of grader g
b_g	bias of grader g
s_i	true grade of submission from student i
z_i^g	observed grade of submission from student i by grader g
d_{ij}^g	observed grade differences between submissions from student i and j by grader g

Figure 1 shows the graphical structure of the PG_6 and PG_7 models. The box in the middle indicate a grader g and the first and last box indicate student i and j whose submissions are graded by grader g. Table 1 summarizes the notations of variables.

In the PG_6 model and PG_7 model, the grader's reliability τ_g and bias b_g and the submission's true grade s_i are the latent variables that need to be estimated. However, these latent variables are related to each. To estimate the values of these latent variables, Gibbs sampling is applied in this work to draw samples of a latent variable from an approximated posterior distribution. After enough iterations, we discard the first few burn-in iterations and we use the mean value of sampled s_i as the final estimate of the true score of submission i. For s_i in PG_6 and τ_g in PG_7, we cannot find a closed form of the posterior distribution, so we use a discrete approximation to get the approximate posterior distribution of these two variables. Next we will describe the details of generation process and the inference of the PG_6 and PG_7 model separately.

4.1 The PG_6 Model

The generative process of the PG_6 model is as follows:

- For each submission submitted by student i
  - Draw true grade $s_i \sim \mathcal{N}(\mu, \frac{1}{\gamma})$
- For each grader g
  - Draw bias $b_g \sim \mathcal{N}(0, \frac{1}{\eta})$
  - Draw reliability $\tau_g \sim \Gamma(s_g, \beta)$
- For each peer grade z_i^g submitted by grader i graded by grader g
  - Draw peer grade $z_i^g \sim \mathcal{N}(s_i + b_g, \frac{1}{\tau_g})$

- For each relative peer grade d_{ij}^g between submissions submitted by student i and j graded by grader g
  - Draw relative peer grade $d_{ij}^g \sim \mathcal{N}(s_i - s_j, \frac{2}{\tau_g})$

In the PG_6 model, the posterior distribution of the true score of submission s_i does not have a closed form. To have an approximate distribution of this latent variable, in this paper, we discretized the true score of submission s_i from 0 to 15 (the full mark of the assignment) with an interval of 0.1. The variables are updated according to Eq. 1.

$$\begin{aligned} b &\sim \mathcal{N}(\frac{\sum_{i\in S_g}\tau_g(z_i^g - s_i)}{\eta + |S_g|\tau_g}, \frac{1}{\eta + |S_g|\tau_g}) \\ \tau &\sim \Gamma(s_g + \frac{|S_g|^2}{2}, \beta + \\ &\frac{\sum_{i\in S_g}(z_i^g - s_i - b_g)^2 + \sum_{i,j\in S_g}\frac{1}{2}(d_{ij}^g - (s_i - s_j))^2}{2}) \\ s &\propto \frac{\beta^{s_i}\tau_i^{s_i-1}}{\Gamma(s_i)} \times exp(\frac{-R}{2}(s_i - \frac{Y}{R})^2) \end{aligned} \tag{1}$$

where $R = \gamma + \sum_{g\in G_i}(\frac{1}{2}\tau_g(|S_g| + 1))$, and $Y = \mu\gamma + \tau_g(\sum_{g\in G_i}(z_i^g - b_g) + \sum_{g\in G_i}\sum_{j\in S_g}\frac{(d_{ij}^g + s_j)}{2})$.

4.2 The PG_7 Model

The difference between PG_7 model and PG_6 model lies in the grader reliability generation: PG_7 adopts Gamma distribution while PG_6 adopts Gaussian distribution. The generative process of the PG_7 model is as follows:

- For each submission submitted by student i
  - Draw true grade $s_i \sim \mathcal{N}(\mu, \frac{1}{\gamma})$
- For each grader g
  - Draw bias $b_g \sim \mathcal{N}(0, \frac{1}{\eta})$
  - Draw reliability $\tau_g \sim \mathcal{N}(s_g, \beta)$
- For each peer grade z_i^g submitted by grader i graded by grader g
  - Draw peer grade $z_i^g \sim \mathcal{N}(s_i + b_g, \frac{\lambda}{\tau_g})$
- For each relative peer grade d_{ij}^g between submissions submitted by student i and j graded by grader g
  - Draw relative peer grade $d_{ij}^g \sim \mathcal{N}(s_i - s_j, \frac{2\lambda}{\tau_g})$

In this model, the posterior distribution of the reliability of a grader τ_g does not have a closed form neither and we apply discrete approximation to approximate the posterior distribution of grader's reliability from 0 to 15 with an interval of 0.1. The variables are updated according to Eq. 2.

$$\begin{aligned} b &\sim \mathcal{N}(\frac{\sum_{i\in S_g}\frac{\tau_g}{\lambda}(z_i^g - s_i)}{\eta + |S_g|\frac{\tau_g}{\lambda}}, \frac{1}{\eta + |S_g|\frac{\tau_g}{\lambda}}) \\ \tau &\propto \tau_g^{\frac{|S_g|^2}{2}} \times exp(\frac{-\beta}{2}[\tau_g - (s_g - \\ &\frac{\sum_{i\in S_g}(z_i^g - s_i - b_g)^2}{2\lambda\beta} - \frac{\sum_{i,j\in S_g}(d_{ij}^g - s_i + s_j)^2}{4\lambda\beta})]^2) \\ s &\sim \mathcal{N}(\frac{Y}{R}, \frac{1}{R}) \end{aligned} \tag{2}$$

where $R = \gamma + \beta + \sum_{g\in G_i}\frac{\tau_g}{\lambda} + \sum_{g\in G_i}\frac{\tau_g*(|S_g|-1)}{2\lambda}$, and $Y = \gamma\mu + \beta\tau_i + \frac{\tau_g}{\lambda}(\sum_{g\in G_i}(z_i^g - b_g) + \frac{\sum_{g\in G_i}\sum_{j\in S_g}(d_{ij}^g + s_j)}{2})$.

Table 2: Dataset Statistics

	Question1	Question2	Question3
# of graders	100	237	105
# of submissions	126	288	141
# of peer grades	493	1121	516
# of instructor grades	114	257	123
full grades	15	15	15
observed mean	6.8	6.7	6.2
observed variance	0.11	0.12	0.14

5. EXPERIMENTAL RESULTS

We perform experiments on a real-world dataset with three questions to evaluate the performance of the proposed models, and we show the results in this section.

5.1 Dataset

The real dataset including peer grades for three questions was collected from a course named "Immortal Arts: Approaching the masters and classics" on the XuetangX platform [1]. For each question, students are asked to write an essay between 100 and 250 words. The peer graders for each submission are automatically assigned by the platform and the grading process is double-blind. After receiving the peer grades, the platform uses the median of peer grades as the final grades for submissions. The grades assigned by TAs are also available in this dataset, which we use as ground truth (true score) in evaluation. The overall statistics of this dataset is shown in Table 2.

5.2 Baselines

In order to evaluate the effectiveness of the proposed models, we compare them with 6 baselines, which are discussed as follows. including the median of peer grades, the mean of peer grades, the PG_1 model and the PG_3 mode in [7] and the PG_4 model and PG_5 model in [6].

- Median: This approach takes the media of peer grades as the final grade. This is the most frequently used method to aggregate peer grades in MOOC platforms such as Coursera[2] and XuetangX platform.
- Mean: This approach simply assigns the mean value of peer grades as the final grade to a submission. In some cases, using the mean value of peer grades as the final peer grades may achieve good performance according to [9].
- PG_1: This is the first probabilistic model for cardinal peer grading estimation that considers the reliability and bias of graders [7].
- PG_3: This is a probabilistic model that links the grader's reliability with the grader's own grade. This model assumes that the variance of distribution for the peer grades is inversely proportional to a linear function of the grader's grade [7].
- PG_4: This is a probabilistic model assuming that a grader's reliability follows a Gamma distribution with the shape parameter equal to the grader's own grade. The PG_6 model is an extension of this model [6].

[1] www.xuetangx.com
[2] www.coursera.org

Table 3: Experimental Results

	Question 1		Question 2		Question 3	
	Mean	Std	Mean	Std	Mean	Std
Mean	1.80		2.29		2.06	
Median	2.19		2.57		2.29	
PG_1	1.97	0.02	2.34	0.01	2.21	0.02
PG_3	1.69	0.07	2.85	0.01	1.92	0.01
PG_4	2.54	0.02	2.94	0.02	3.07	0.02
PG_6	1.31	0.01	**1.44**	**0.01**	1.38	0.02
PG_5	1.52	0.04	1.80	0.01	1.74	0.02
PG_7	**1.24**	**0.02**	1.45	0.01	**1.31**	**0.01**

- PG_5: This is a probabilistic model assuming that a grader's reliability follows a Gaussian distribution with the mean equal to the grader's own grade. The PG_7 model is an extension of this model.

5.3 Experimental Settings

As described before, many hyper-parameters are used in the proposed models and baselines, and it is important to set reasonable values for these hyper-parameters. In this section, we describe how to set the values of these hyper-parameters in our experiment.

Since the proposed models are the extensions of the PG_4 and PG_5 model in [6], to evaluate the effect of leveraging relative grades, we set the same values for the shared hyper-parameters in the proposed models and the PG_4 and PG_5 models. We use the mean and variance of the peer grades as the mean (μ) and variance ($\frac{1}{\gamma}$) of the prior distribution of the true grade (s_i). As claimed in [6], the β in the PG_4 model which decides the rate of the Gamma distribution for the grader's reliability and the λ in the PG_5 model which determines the variance of the Gaussian distribution for peer grades are the most important hyper-parameters. These parameters have a significant influence on the performance of these two models while other hyper-parameters influence the performance slightly if set in a reasonable range. Thus we mainly tune β in the PG_4 and PG_6 model and λ in the PG_5 and PG_7 model. We search for these two hyper-parameters in the range of $[50, 300]$ with the interval of 50 to get the best performance. We set η to 0.1 in our experiment, and in the PG_5 and PG_7 model, β is set to 0.1. For each latent variable, we sample it for 300 iterations and the first 60 iterations are the burn-in iterations that will be discarded. The average results over 10 runs with the hyper-parameter settings described above are reported.

5.4 Real Dataset Performance

We use Root-Mean-Square-Error (RMSE) to evaluate the performance of the proposed models and baselines on the datasets. The experimental results are shown in Table 3. From Table 3, we can find that on all these three questions, the PG_6 and PG_7 models outperform other baselines. The RMSE of the PG_6 and the PG_7 models which incorporate the relative observed grades to capture the dependency between true grades of submissions has dropped compared with that of the PG4 and PG5 models. The results demonstrate the effectiveness of incorporating relative peer grades in cardinal peer grade estimation.

To better illustrate the performance of the PG_6 and PG_7 models, we further compare the estimated grades with the ground truth on individual submissions in Figure 2. The submissions are sorted with an increasing order of the ground truth. Then we plot the estimated grades from *Mean* (the best naive method), PG_5 (the best baseline), and the proposed PG_7 model which has the best performance. We can find that the estimated grades by all three models show an increasing trend, but *Mean* shows a strong negative bias in the peer grades: the peer grades are consistently lower than the ground truth grade. Therefore, it is important to model the bias in graders to improve the aggregation results. PG_5 and PG_7 both show positive bias compared with the ground truth, but $PG5$'s bias is a bit higher. The comparison between PG_5 and PG_7 illustrates that the relative grades can also help estimate the bias more accurately. It may imply that although graders cannot give accurate absolute grades, they can assign accurate relative grades.

We further compare the experimental bias estimated by the proposed models with the real bias. The experimental bias is defined as the average difference between the peer grades assigned by a grader and the estimated true grades. The real bias is defined as the average difference between the peer grades assigned by a grader and the ground truth. For example, a grader g grades two submissions from student i and j, the experimental bias of this grader is $\frac{(z_i^g - s_i) + (z_j^g - s_j)}{2}$ and the real bias is $\frac{(z_i^g - s_i^*) + (z_j^g - s_j^*)}{2}$, where s_i and s_j are the estimated grades, s_i^* and s_j^* are the groundtruth grades for submission i and j. The results are illustrated in Figure 3, where x-axis denotes the real bias and y-axis denotes the experimental bias. We can see that most graders are harsh graders whose real biases are less than 0. The diagonal means that the estimated bias is the same as the real bias. The closer to the diagonal, the more accurate the bias estimation is. We can observe that our estimated bias is close to the real bias. With better bias estimation, the proposed models achieve more accurate cardinal estimation. This result again indicates the informativeness of relative grades in estimating final grades.

5.5 Sensitivity of Hyper-parameters

To show how the value of hyper-parameter β in the PG_6 model and the hyper-parameter λ in the PG_7 model will influence the performance, we conduct experiments using different values of these two hyper-parameters with all other hyper-parameter fixed. In the experiment to test the sensitivity of the models, the settings for other fixed hyper-parameters are the same as described above and the β in the PG_6 model and the λ in the PG_7 model are set from 50 to 300 with an interval of 50. The results in Figure 4 show that in a reasonable range these two models are robust to the value of the parameter and achieve acceptable performance.

6. CONCLUSIONS AND FUTURE WORK

With the popularity of the MOOCs, peer assessment has become an effective paradigm for large-scale grading. The aggregation of peer grades is a challenging problem due to the various levels of bias and reliability among graders that are unknown. Existing work contributes to the development of effective peer grading aggregation methods by modeling grader bias and reliability, but they ignore an important aspect in peer grading aggregation, which is the dependency relation among grades. In these models, the relative grades are not considered and the true grades of submission are

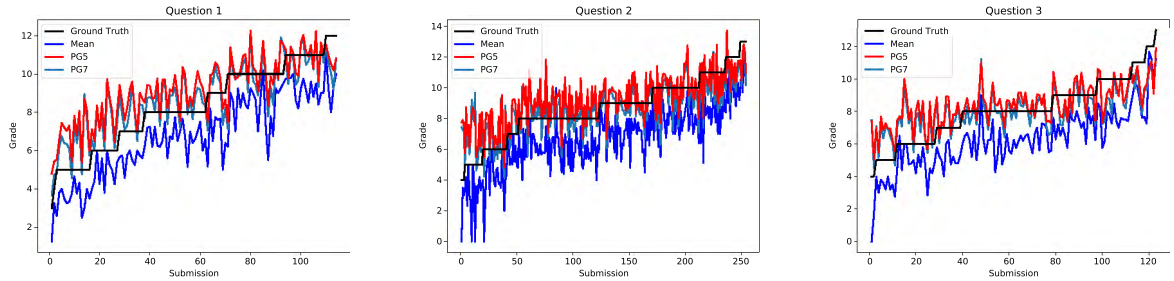

Figure 2: The estimated grades of three questions using mean, the PG_6 and PG_7 model and ground truth. The submissions are sorted by their ground truth.

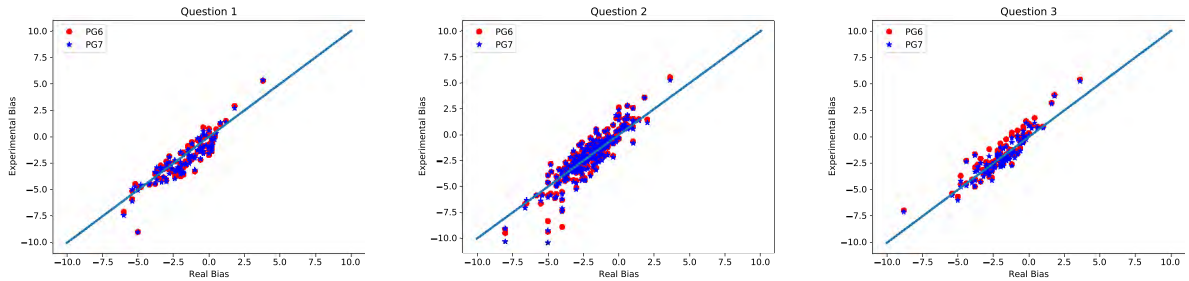

Figure 3: The comparison of experimental bias with real bias

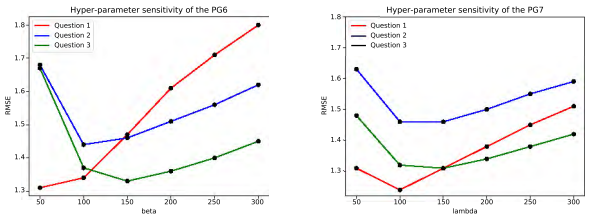

Figure 4: Hyper-parameter sensitivity of the PG_6 and PG_7 model

modeled independently. Modeling the dependencies among the true grades of different submissions can help improve the robustness of the aggregated grade estimation. In this paper, we propose two novel models that leverage relative grades to achieve improved estimation of final grades. In the proposed probabilistic models, we capture the distributions of true scores based on graders' bias and reliability degrees as well as their own submission scores which represents their knowledge about the question. In addition, the proposed models couple the true scores of different submissions via their differences. Effective inference algorithms are proposed to infer both model parameters and final scores. Experimental results demonstrate that the proposed models improve the accuracy of cardinal peer grading estimation. It can also be observed that the relative peer grades among submissions indeed contribute to the improvement in the accuracy of cardinal peer grading estimation.

In the future, we will investigate how to better model the ability of graders reflecting both reliability and bias of graders and how to cluster the graders and submissions into different groups to improve the peer assessment.

7. ACKNOWLEDGEMENTS

This work is sponsored by NSF IIS-1553411. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

8. REFERENCES

[1] A factorization approach to evaluate open-response assignments in moocs using preference learning on peer assessments. *Knowledge-Based Systems*, 85:322 – 328, 2015.

[2] H. P. Chan and I. King. Leveraging social connections to improve peer assessment in moocs. In *Proceedings of World Wide Web Companion*, pages 341–349, 2017.

[3] L. de Alfaro and M. Shavlovsky. Crowdgrader: A tool for crowdsourcing the evaluation of homework assignments. In *Proceedings of SIGCSE*, 2014.

[4] J. Hamer, K. T. K. Ma, and H. H. F. Kwong. A method of automatic grade calibration in peer assessment. In *Proceedings of Computing Education*, ACE '05, 2005.

[5] C. Kulkarni, K. P. Wei, H. Le, D. Chia, K. Papadopoulos, J. Cheng, D. Koller, and S. R. Klemmer. Peer and self assessment in massive online classes. *ACM Trans. Comput.-Hum. Interact.*, 20(6):33:1–33:31, Dec. 2013.

[6] F. Mi and D.-Y. Yeung. Probabilistic graphical models for boosting cardinal and ordinal peer grading in moocs. In *Proceedings of AAAI*, 2015.

[7] C. Piech, J. Huang, Z. Chen, C. B. Do, A. Y. Ng, and D. Koller. Tuned models of peer assessment in moocs. *CoRR*, abs/1307.2579, 2013.

[8] K. Raman and T. Joachims. Methods for ordinal peer grading. In *Proceedings of SIGKDD*, pages 1037–1046, 2014.

[9] M. S. Sajjadi, M. Alamgir, and U. von Luxburg. Peer grading in a course on algorithms and data structures: Machine learning algorithms do not improve over simple baselines. In *Proceedings of Learning @ Scale*, pages 369–378, 2016.

[10] N. B. Shah, J. K. Bradley, A. Parekh, and K. Ramchandran. A case for ordinal peer-evaluation in moocs. 2013.

[11] T. Walsh. The peerrank method for peer assessment. *arXiv preprint arXiv:1405.7192*, 2014.

[12] A. E. Waters, D. Tinapple, and R. G. Baraniuk. Bayesrank: A bayesian approach to ranked peer grading. In *Proceedings of Learning @ Scale*, 2015.