Statistical Analysis of Peer Grading: A Latent Variable Approach

Giuseppe Mignemi, Yunxiao Chen and Irini Moustaki

Bocconi Institute for Data Science and Analytics
giuseppe.mignemi@unibocconi.it
London School of Economics
Y.Chen186@lse.ac.uk I.Moustaki@lse.ac.uk

Abstract. Peer grading is an educational system in which students assess each other's work. It is commonly applied under Massive Open Online Course (MOOC) and offline classroom settings. Peer grading data have a complex network structure, where each student is a vertex of the network, and each peer grade serves as an edge connecting one student as a grader to another student as an examinee. We introduce a latent variable model framework for analyzing peer grading data and develop a fully Bayesian procedure for its statistical inference. The proposed approach produces more accurate aggregated grades by modelling the heterogeneous grading behaviour with latent variables and provides a way to assess each student's performance as a grader. It may be used to identify a pool of reliable graders or generate feedback to help students improve their grading. Thanks to the Bayesian approach, uncertainty quantification is straightforward when inferring the student-specific latent variables as well as the structural parameters of the model. The proposed method is applied to a real-world dataset.

Keywords: peer grading, rating models, cross-classified models, latent variable approach, Bayesian modelling

1 Introduction

Peer grading, also known as peer assessment, is a system of formative assessment in education whereby students assess and give feedback on one another's assignments. It substantially reduces teachers' burden for grading and improves students' understanding of the subject and critical thinking [9]. Consequently, it is widely used in many educational settings, including massive open online courses (MOOCs; [4]), large university courses [2], and small classroom settings [7].

2 Proposed Model

Consider N students who receive T assignments. Each student i's work on assignment t is randomly assigned to a small subset of other students to grade their work. We denote this subset as S_{it}, which is a subset of $\{1,...,i-1,i+1,...,N\}$. Each grader $g \in S_{it}$ gives a grade Y_{igt} to this work, following certain scoring rubrics. For simplicity, we consider the case when Y_{igt} is continuous while pointing out that extending the proposed model to ordinal data may be achieved using

an underlying variable formulation [1]. It is common, but not required, for the number of grades $|S_{it}|$ to be the same for all students and assignments. An aggregated score is then computed as a measure of student i's performance on the tth assignment student i's performance on the tth assignment, often by taking the mean or the median of the peer grades $Y_{igt}, g \in S_{it}$. We note that a simple aggregation rule, such as the mean and the median of the peer grades, fails to account for the grader effect and, thus, may not be accurate enough.

2.1 Proposed Model

Modelling Peer Grade Y_{igt}. We assume the following decomposition for the peer grade Y_{igt}:

$$Y_{igt} = \theta_{it} + \tau_{igt} - \delta_t, \quad i = 1,...,N, t = 1,\ldots,T, g \in S_{it}. \tag{1}$$

Here, δ_t captures the difficulty level of assignment t. A larger value of δ_t corresponds to a more difficult assignment. In addition, θ_{it} represents student i's true score for assignment t, and τ_{igt} is an error brought by the grader. We assume θ_{it}, τ_{igt} and δ_t to be independent.

Modelling True Score θ_{it}. For each student i, we assume that their true scores for different assignments θ_{it}, $t = 1,\ldots,T$, are independent and identically distributed, following a normal distribution

$$\theta_{it} \sim N(\alpha_i, \eta_i^2), \tag{2}$$

where the mean and variance are student-specific latent variables. The latent variable α_i captures the student's average performance over the assignments, and the latent variable η_i^2 measures their performance stability. This model assumes the true scores fluctuate randomly around the average score α_i without a trend. This assumption can be relaxed if we are interested in assessing students' growth over time.

Modelling Grader Effect τ_{igt}. Each student g grades multiple assignments from multiple students. We let $H_g = \{(i,t) : g \in S_{it}, t = 1,...,T\}$ be all the work student g grades. For each student g, we assume that τ_{igt}, for all $(i,t) \in H_g$, are independent and identically distributed (i.i.d.), following a normal distribution $N(\beta_g, \phi_g^2)$, where the mean and variance are student-specific latent variables. The latent variable β_g may be interpreted as the bias of student g as a grader. For two students g and g' satisfying $\beta_g > \beta_{g'}$, student g will give a higher grade on average than student g' when grading the same work. We say grader g is unbiased when $\beta_g = 0$. Moreover, the latent variable ϕ_g^2 is a measure of grader reliability. A smaller value of ϕ_g^2 implies that the grader tends to follow a consistent standard, while a larger value suggests that they may give erratic grades that lack a consistent standard. In other words, when grading multiple pieces of work with the same true score and assignment difficulty (so that ideally they should receive the same grade), a grader with a small ϕ_g^2 tends to give similar grades, and thus, the grades are more reliable. In contrast, a grader with a large ϕ_g^2 tends to give noisy grades that lack consistency. We remark that the grader effects τ_{igt}, $t = 1,...,T$, are assumed to be i.i.d. in the current setting, which means the grading quality remains the same

over time. It is possible to extend the model for τ_{igt} to capture its change. We leave it for future investigation.

Joint Modelling of Student-Specific Latent Variables. The above model specification introduces four student-specific latent variables $(\alpha_i, \beta_i, \eta_i^2, \phi_i^2)$ for each student i. We allow for dependence between these latent variables, which will enable us to borrow information between the performance data and the grading data of the same student when evaluating their performances as an examinee and as a grader. More precisely, we assume $(\alpha_i, \beta_i, \eta_i^2, \phi_i^2)$, $i = 1, ..., N$ are i.i.d., with $(\alpha_i, \beta_i, \log(\eta_i^2), \log(\phi_i^2))$ following a multivariate normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (\mu_1, ..., \mu_4)^\top$ and $\boldsymbol{\Sigma} = (\sigma_{ij})_{4\times 4}$. For identifiability purposes, we fix $\mu_1 = \mu_2 = 0$, so that the average score of each assignment (averaged across students and graders) is completely captured by the difficulty parameter δ_t. We note that no constraint is imposed on μ_3 and μ_4.

2.2 Bayesian Inference

We adopt a fully Bayesian procedure for drawing statistical inference under the proposed model.

Prior specification. We first specify the prior for the assignment difficulty parameters $\delta_1, \ldots, \delta_T$. When T is relatively small (e.g., $T \leq 5$), we simply assume each δ_t to have a weakly informative prior $N(0, 5^2)$. When T is larger, a hierarchical prior specification may be used by assuming $\delta_1, \ldots, \delta_T$ to be i.i.d. following a certain prior distribution (e.g., normal) with some hyper-parameters and further setting a hyper-prior distribution for the hyper-parameters. We proceed to specify a prior for the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in the joint distribution for the student-specific latent variables. Recall that μ_1 and μ_2 are constrained to zero, for which no prior needs to be set. For μ_3 and μ_4, we assume them to be independent, each following a weakly informative normal prior $N(0, 5^2)$. For the covariance matrix $\boldsymbol{\Sigma}$, we reparameterize

$$\boldsymbol{\Sigma} = \mathbf{S}\boldsymbol{\Omega}\mathbf{S},$$

where $\mathbf{S} = \text{diag}(\sqrt{\sigma_{11}}, \ldots, \sqrt{\sigma_{44}})$ is a 4×4 diagonal matrix with its diagonal entries being the standard deviations of $(\alpha_i, \beta_i, \log(\eta_i^2), \log(\phi_i^2))$, and $\boldsymbol{\Omega} = (\omega_{ij})_{4\times 4} = \mathbf{S}^{-1}\boldsymbol{\Sigma}\mathbf{S}^{-1}$ is the correlation matrix of $(\alpha_i, \beta_i, \log(\eta_i^2), \log(\phi_i^2))$. The prior distribution on $\boldsymbol{\Sigma}$ is imposed by the priors on $\mathbf{S}$ and $\boldsymbol{\Omega}$. For $\mathbf{S}$, we assume $\sqrt{\sigma_{11}}, \ldots, \sqrt{\sigma_{44}}$ to be i.i.d., each following a half-Cauchy distribution with location 0 and scale 5. For the correlation matrix $\boldsymbol{\Omega}$, we assume a Lewandowski-Kurowicka-Joe (LKJ) prior distribution with shape parameter 1 [6], which corresponds to the uniform distribution over the space of all correlation matrices.

Computational aspects. We adopt the No-U-Turn Hamiltonian Monte Carlo (HMC) sampler [5], a computationally efficient MCMC sampler, and implement it under the Stan programming language. Compared with classical MCMC samplers, such as the Gibbs and Metropolis-Hastings samplers, the No-U-Turn HMC sampler uses geometric properties of the target distribution to propose posterior samples and thus converges faster to high-dimensional target distributions [5].

Regarding the implementation, for all the models, 4 HMC chains are run in parallel for 2000 iterations, of which the first 1000 iterations were specified as warm-up. We use the rstan R package to analyze the resulting posterior samples, more specifically, it enables us to merge the MCMCs, compute the summary statistics of the posteriors and check the MCMC mixing and convergence. Moreover, the R package loo [8] and Bayesplot [3] are used separately for model comparisons and to plot the results, respectively.

3 Real Data Example

These peer grading data are from [10]. Participants are $N = 274$ American undergraduate students attending a Biology course. A double-blinded individual peer assessment was implemented for four different assignments, $T = 4$ throughout the course. Each student was graded, on average, by a random set of 5 other students for each assignment. The coursework was rated on a $1-7$ Likert scale with instructor-provided anchor descriptions for each rating level. For the current analysis, only the students who completed at least three assignments were considered. This allows us to fit the LGC model. It results in a sample size of $N = 212$ students.

Results. No mixing or convergence issues emerge from a graphical inspection of the MCMCs and also as suggested by the values of the $\hat{R} < 1.01$.
The assignment difficulty levels seem to be in increasing order (see Table 1). Conditional to the other parameters, the first and the fourth assignments are, respectively, the easiest and the most difficult ones. The 95% quantile-based credible intervals of the assignment difficulty parameters are moderately narrow, suggesting a low level of uncertainty for these parameters. The estimates of the location parameters of the latent variables, μ_3 and μ_4, suggest that students are on average more consistent than reliable. The posterior mean and the 95% credible intervals of the first quantity, respectively $\hat{\mu}_3 = -1.27$ and $(-1.46, -1.10)$, are considerably smaller than those related to the second one, $\hat{\mu}_4 = -0.46$ and $(-0.51, -0.41)$. This implies that, on average, the variance of the student's ability is smaller than the error variance of the grades they give. They are more consistent as an examinee than as a grader. From a substantive point of view, considering that they are not grader experts, it seems reasonable. Note that these quantities are expressed on a logarithmic scale, which implies that the average variance of the students' proficiency across different assignments is $exp(\hat{\mu}_3) = 0.28$, and, on average, their reliability parameter is $exp(\hat{\mu}_4) = 0.63$.
Students are moderately homogeneous in terms of their mean abilities, as suggested by the relatively small values of σ_1, whereas they are more variable in their systematic bias, which is indicated by the posterior values of the parameter σ_2. In other words, they are, on average, more similar as examinees than they are as graders. Moreover, students are widely different from each other concerning their consistency across assignments, as suggested by the posterior values of the structural parameter σ_3. There is slightly less variability among them concerning the reliability parameters as suggested by the values of σ_4.
Regarding the dependency among the latent variables, it emerges that higher values of students' proficiency are associated with higher values of consistency.

Indeed, there is evidence of a strong correlation, between the first and the second student-specific variable, respectively, α_i and $log(\eta_i^2)$, as suggested by the posterior mean and the 95% credible interval $\hat{\omega}_{13} = -0.86$ and $(-0.96, -0.74)$, respectively. In addition, higher values of mean bias are predictive of higher reliability levels. This is evidenced by the posterior mean and 95% credible interval of the parameter ω_{24}, respectively, $\hat{\omega}_{24} = -0.73$ and $(-0.83, -0.63)$. The estimates of the other correlation parameters do not provide clear evidence about any other dependency among the latent variables under the present model.
Going to the student-specific level, each student might be provided with a Score estimate and a 95% quantile-based credible interval for each assignment as a measure of uncertainty. The posterior mean of $\hat{\theta}_{it} - \hat{\delta}_t$ might be a point estimate for students' Scores. The posterior distributions of both the average bias and the reliability of each grader might be useful information to assess their grading behaviour.

	Parameter	Post. Mean	95% CI
Assignments	δ_1	-6.31	$(-6.39, -6.24)$
	δ_2	-5.38	$(-5.46, -5.32)$
	δ_3	-5.36	$(-5.43, -5.29)$
	δ_4	-4.96	$(-5.03, -4.89)$
Students	μ_3	-1.27	$(-1.46, -1.10)$
	μ_4	-0.46	$(-0.51, -0.41)$
	σ_1	0.23	$(0.19, 0.28)$
	σ_2	0.35	$(0.32, 0.39)$
	σ_3	0.66	$(0.53, 0.83)$
	σ_4	0.32	$(0.29, 0.37)$
	ω_{12}	-0.09	$(-0.27, 0.08)$
	ω_{13}	-0.86	$(-0.96, -0.74)$
	ω_{14}	0.17	$(-0.02, 0.36)$
	ω_{23}	-0.08	$(-0.29, 0.13)$
	ω_{24}	-0.73	$(-0.83, -0.63)$
	ω_{34}	0.12	$(-0.10, 0.34)$

Table 1. Estimated structural parameters. For each parameter, the posterior mean (Post. Mean) and the 95% quantile-based credible interval (CI) are reported. The parameter δ_t is the difficulty level of the assignment t; μ_3 and μ_4 are the location parameters of the third and the fourth latent variable; $\sigma_1, \ldots, \sigma_4$ are the standard deviations of the latent variables; ω_{mn} is the correlation parameter between the latent variables m and n.

Acknowledgments

We are sincerely grateful to Professor Oscar Luaces for sharing the data set on peer grading.

References

1. BARTHOLOMEW, D., KNOTT, M., AND MOUSTAKI, I. *Latent variable models and factor analysis: a unified approach*, 3 ed. New York: Wiley, 2011.
2. DOUBLE, K. S., MCGRANE, J. A., AND HOPFENBECK, T. N. The impact of peer assessment on academic performance: a meta-analysis of control group studies. *Educational Psychology Review 32* (2020), 481–509.
3. GABRY, J., SIMPSON, D., VEHTARI, A., BETANCOURT, M., AND GELMAN, A. Visualization in bayesian workflow. *J. R. Stat. Soc. A 182* (2019), 389–402.
4. GAMAGE, D., STAUBITZ, T., AND WHITING, M. Peer assessment in moocs: systematic literature review. *Distance Education 40*, 2 (2021), 1–22.
5. HOFFMAN, M. D., AND GELMAN, A. The no-u-turn sampler. *The Journal of Machine Learning Research 15* (2014), 1593–1623.
6. LEWANDOWSKI, D., KUROWICKA, D., AND JOE, H. Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis 100*, 9 (2009), 1989–2001.
7. SANCHEZ, C., ATKINSON, K., KOENKA, A., MOSHONTZ, H., AND COOPER, H. Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology 109*, 8 (2017), 1049–1066.
8. VEHTARI, A., GELMAN, A., AND GABRY, J. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing 27* (2017), 1413–1432.
9. YIN, S., CHEN, F., AND CHANG, H. Assessment as learning: how does peer assessment function in students' learning? *Frontiers in Psychology 13* (2022), 912568.
10. ZONG, Z., SCHUNN, C. D., AND WANG, Y. What aspects of online peer feedback robustly predict growth in students' task performance? *Computers in Human Behavior 124* (2021), 106924.