Moderation—A Bayesian Model

Chen Chen
Measurement and research services

01/02/2019

Contents

1 Introduction

This report proposes a Bayesian method which moderates the Raw School Assessment using Examination Marks. The model simply finds a middle point, so called change point, and connects this point to the top and bottom scores.

The model was built a while ago, and it was found to be unstable to fit by maximum likelihood. However, the recent purchase of the Beast makes a Bayesian simulation of the model finishes in seconds.

2 Log odds of EM over MSA follows Laplace distribution

Figure 1 shows that the log odds $\log \frac{\text{EM}}{\text{MSA}}$ follows a Laplace distribution with mean zero and scale $s = 0.0125$[1]. Hence I model the Moderated School Assessment as

$$\text{EM} = \text{MSA}e^{\epsilon}, \tag{1}$$

where Laplacian noise ϵ has a density function

$$f(\epsilon) = \frac{1}{2s} \exp\left(-\frac{|\epsilon|}{s}\right). \tag{2}$$

Noise ϵ models the variation of students' performance, depending on the status on the Examination day. Taking logarithm of model (1) gives

$$\epsilon = \log \text{EM} - \log \text{MSA}. \tag{3}$$

[1]The $s = 0.0125$ is learned by Bayesian Monte Carlo model. Maximum likelihood estimator of s is 0.0122. A mixture model gives $s = 0.0121$. These estimates are all consistent and $0.0125$is used from this point.

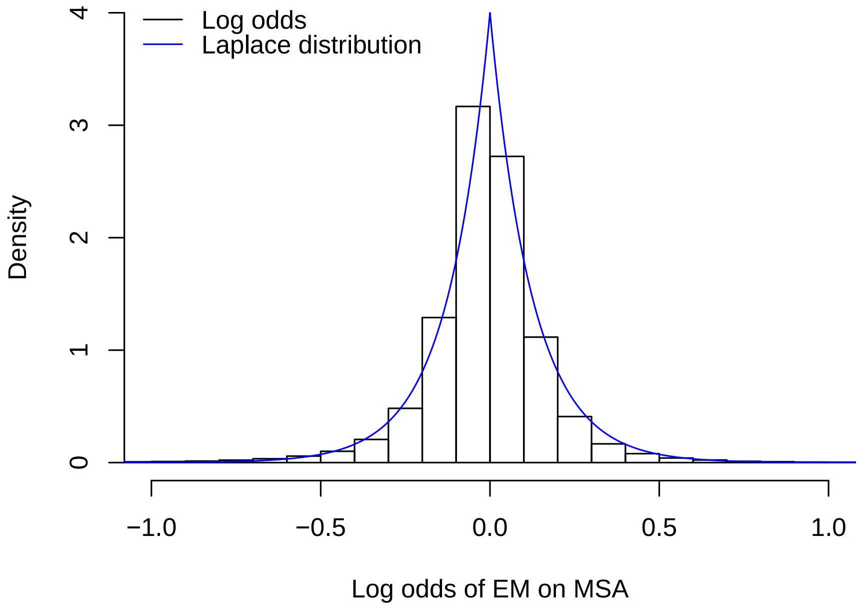

Figure 1: The log odds of Examination Mark and Moderated School Assessment follows a Laplace distribution. The blue line is the density plot for Laplace distribution with mean zero and scale 0.125. The boxes are the histogram of the Log odds for all courses and all school in the years 2016 and 2017. This plot excludes all students who rank one in both Raw School Assessment and Examination Markbecause they will have zero log odds. Fitting with the Maximum likelihood method excludes these students. When fitting Bayesian and Mixture models, the proportion of students who have zero log odds were modelled as a parameter.

3 Model MSA

I model the moderated school assessment by a linear function

$$\mathrm{MSA}(\mathrm{SA}) = a \times \mathrm{SA} + b,\ \ a > 0. \tag{4}$$

where intercept b models a shift from teacher's marking scale to the examiner's making scale and slope a models a scaling factor (i.e. stretch or compress). A very small scaling factor a reflect the teacher likes to give very extreme Raw School Assessment, and large scaling factor a reflect the teacher gives students similar marks.

However, the same teacher may usually have different scales for capable and poor students. Let $i = 1, 2, \ldots, N$ to denote the student whose SA is the ith highest. Define $i = \tau$ to be the cut-off between capable and poor. I propose a change point model

$$\mathrm{MSA}(\mathrm{SA}) = \begin{cases} a_1 \times \mathrm{SA} + b_1, & \mathrm{SA} < \mathrm{SA}_\tau \\ a_2 \times \mathrm{SA} + b_2, & \mathrm{SA} \geq \mathrm{SA}_\tau \end{cases}. \tag{5}$$

Due to political reasons, the student with highest Raw School Assessment will have a Moderated School Assessment equals to the highest Examination Mark in the group. Thus I set

$$b_2 = \max \mathrm{EM}_i - a_2 \mathrm{SA}_1.$$

Also, to ensure continuity at $i = \tau$, I set

$$b_1 = b_2 + (a_2 - a_1)\mathrm{SA}_\tau.$$

4 Frequentist approach is unstable

The frequentist approach of fitting the change point model (5) is by maximum likelihood. Multiplying density function (2), the likelihood is

$$L(a_1, a_2, \tau) = \frac{1}{2^N s^N} \exp\left(-\frac{\sum_{i=1}^N |\epsilon_i|}{s}\right).$$

Substitute equation (3) into this likelihood

$$L(a_1, a_2, \tau) = \frac{1}{2^N s^N} \exp\left(-\frac{\sum_{i=1}^N |\log \mathrm{EM}_i - \log \mathrm{MSA}_i|}{s}\right).$$

Substitute the change point model (5)

$$\begin{aligned} L(a_1, a_2, \tau) &= \frac{1}{2^N s^N} \exp\left(-\frac{1}{s} \sum_{i=1}^{\tau-1} |\log \mathrm{EM}_i - \log\left(a_1 \times \mathrm{SA}_i + b_1\right)|\right) \\ &\quad \times \exp\left(-\frac{1}{s} \sum_{i=\tau}^{N} |\log \mathrm{EM}_i - \log\left(a_2 \times \mathrm{SA}_i + b_2\right)|\right). \end{aligned} \tag{6}$$

Finding the maximum of this likelihood function is equivalent to minimise the loss function

$$\sum_{i=1}^{\tau-1} |\log \mathrm{EM}_i - \log\left(a_1 \times \mathrm{SA}_i + b_1\right)| + \sum_{i=\tau}^{N} |\log \mathrm{EM}_i - \log\left(a_2 \times \mathrm{SA}_i + b_2\right)|.$$

This loss function is known as the mean absolute error loss. Minimising this loss function by an optimisation package gives the maximum likelihood estimator of parameters a_1, a_2 and τ.

Mean absolute error is known to be very robust to outliers, that is, students performed atypically in the exam. It is robust because the corresponding distribution, Laplace distribution, has fat tails. However, this loss function tends unstable result because of the non-smoothness of absolute value. Unstableness means that a small variation in student marks may cause a large jump in the parameters a_1, a_2 and τ, especially when the number of student in a school group is small. This is not desired because the Moderated School Assessment of one student should not depend heavily on another student.

5 Bayesian Prior and model

To produce a more stable result, I use the Bayesian approach to simulate a change point model (5).

A Bayesian approach requires us to supply Prior distributions for parameters a_1,a_2 and τ. Prior distributions can be informative and uninformative. For example, for every school group, the change point τ can be at different positions, depending on the teacher. Hence I specify change point τ to have equal probability for all student

$$p(\tau = i) = \frac{1}{N} \text{ for all } i = 1, 2, \ldots, N. \tag{7}$$

However, NESA has plenty of information about the slopes a_1 and a_2. Firstly, on average, teacher's scale should be similar to Examiner's scale, so $E(a_i) \approx 1$ for $i = 1, 2$. Also, Dr. Bob proposed that a good model should avoid slope smaller than 0.1. So $p(a_i < 0.1)$ should be close to zero.

More systematically, I calculate the prior distributions for the slopes a_1 and a_2 from historical data by bootstrapping. Fitting 5000 different school groups in the year of 2016 and 2017 with replacement, I estimate the prior for a_1 and a_2 as

$$\log(a_1) \sim N(-0.01, 1.5) \text{ and } \log(a_2) \sim N(0.02, 1.5). \tag{8}$$

For these two distributions, $E(a_1) \approx 0.99$ and $E(a_2) \approx 1.02$. Probabilities $p(a_1 < 0.1) \approx 0.0632$ and $p(a_2 < 0.1) \approx 0.0607$. Prior distributions (7), (8) and likelihood (6) fully specifies the Bayesian change point model. I implement a Gibbs sampler to simulate this model by `Openbugs`.

6 Bayes factor decides the number of change points

I use Bayes factor to systematically assess whether such a point exists. The idea is very straightforward: I calculate the Bayesian likelihood of the simple linear model (4) and the change point model (5). I use the model with a larger likelihood. I implement the same approach to assess whether a second change point is required. Among over 100 simulations, I found no statistical evidence for the second change point for any of the school groups.

7 Visualisation of Bayesian results

Examples included in the slides attached visualises the new moderation results. To generate further examples, please run the code `Bayesian_change_point_tota`
The code randomly selects a course and produce moderation result. `Openbugs` is required. Also, one needs to install R packages `BRugs,mrfDepth,outliers`.