IRT Mixture Modeling with NWEA MAP

Klint Kanopka

4/22/2020

1 Model Specification

The model fit was:

$$P(X_{ij} = 1|\theta_i, b_{je}, b_{jl}, \pi_{ij}) = \pi_{ij}\sigma\big(\theta_i - b_{je}\big) + (1 - \pi_{ij})\sigma\big(\theta_i - b_{jl}\big)$$

Where: - θ_i is student i's latent ability - b_{je} is item j's difficulty when encountered early in the test - b_{jl} is item j's difficulty when encountered late in the test - σ is the standard logistic sigmoid function, specified by:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

- π_{ij} is a mixing parameter for the ith student's exposure to item j. This parameter is specified as a function of sequence number:

$$\pi_{ij} = \sigma\left(\frac{k_i - s_j}{c}\right)$$

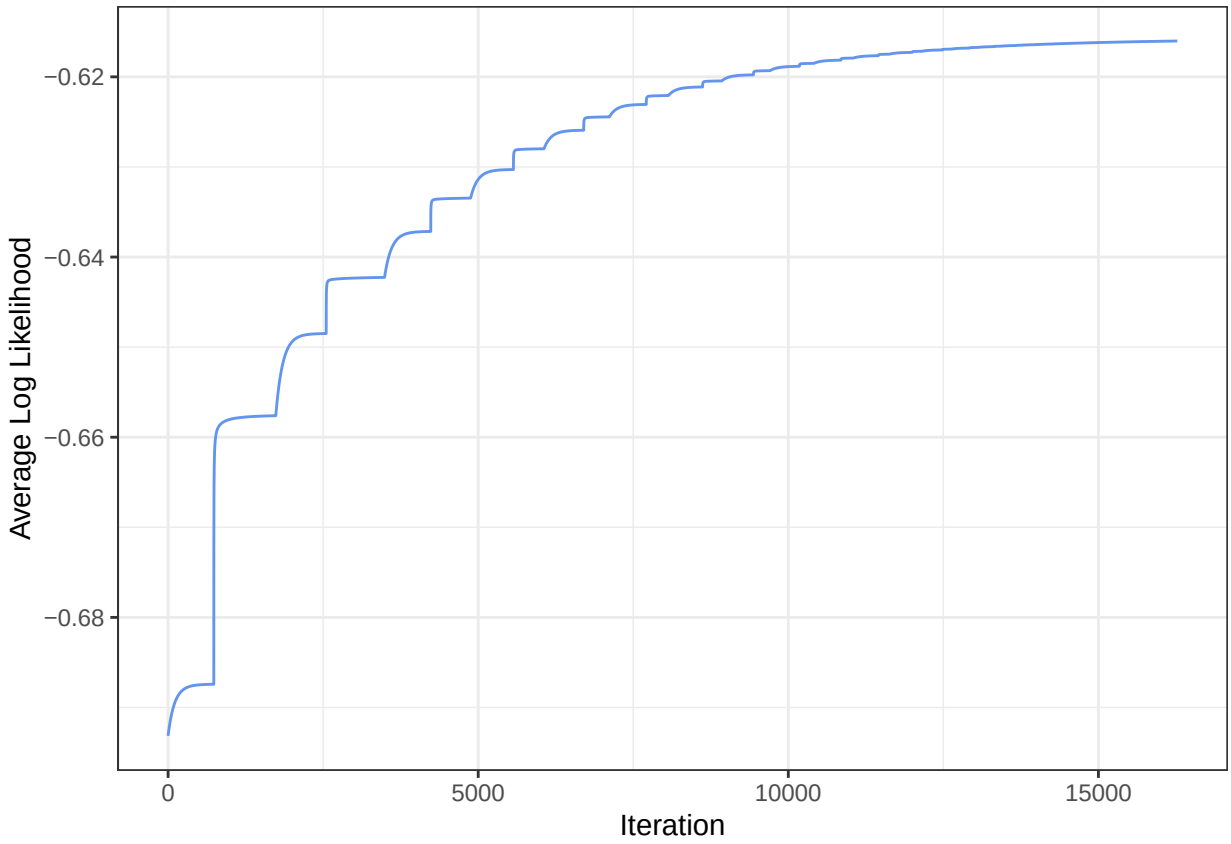
where s_{ij} is the *sequence number*, or the position in the test sequence where student i saw item j, k_i is a learned parameter that corresponds to the sequence number where student i's item response function is a 50/50 mixture of early and late test behavior, and c is a scale parameter that is fixed across all students. For this analysis, the scale parameter was fixed to $c = 10$.

1.1 Data

Data is taken from a spring administration of the NWEA MAP assessment. It consists of 7256 individual items taken by 464,117 individual students across multiple grades. Each student sees, on average, 39 different items, resulting in 18,187,294 observed item responses.

1.2 Fitting

The model was fit in Python (3.6.9) primarily using NumPy (1.17.4), optimized using joint maximum likelihood estimation and (full batch) gradient ascent. A plot of average log likelihood after each gradient ascent update is shown below. Note that a convergence threshold of $\varepsilon = 10^{-5}$ was used. From this, we observe that a smaller value of ε could be used in the future.

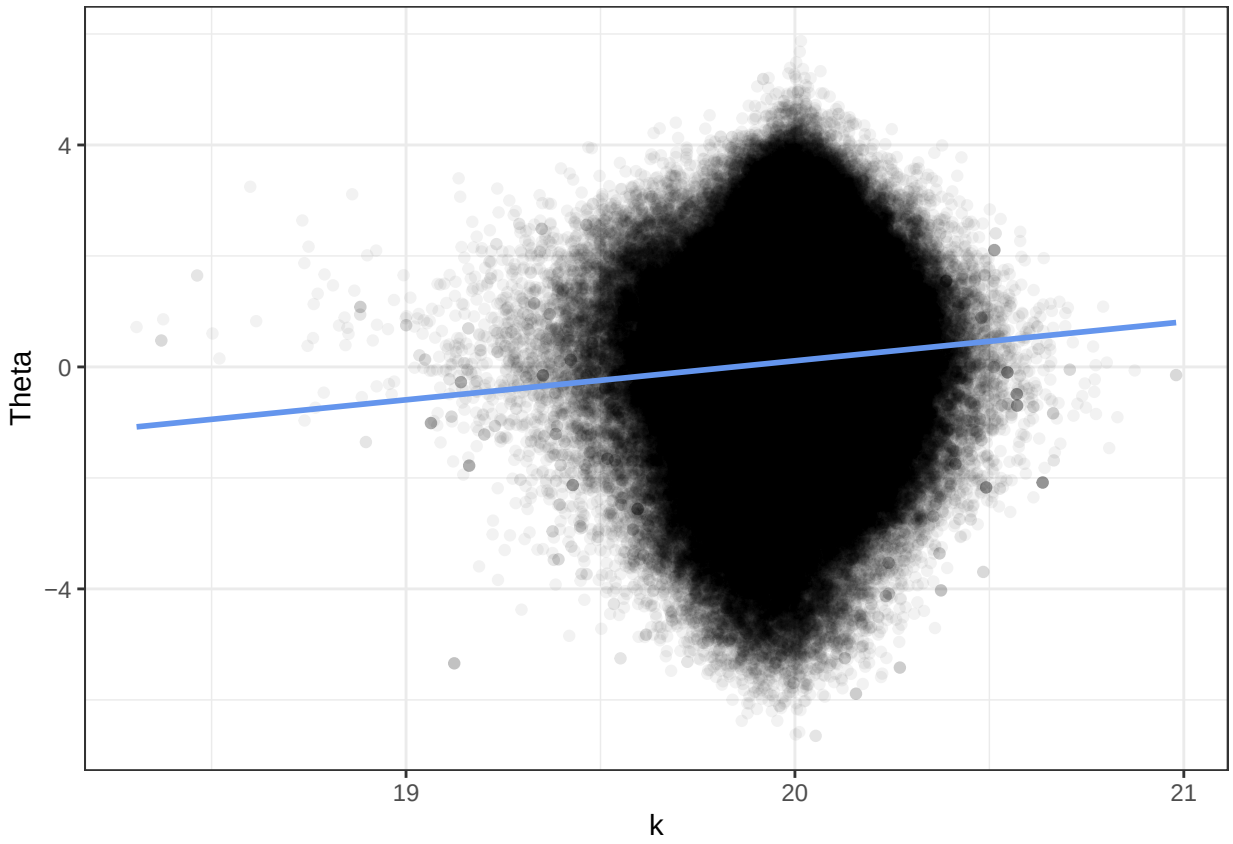


2 Initial Descriptives

2.1 Persons

First we look to see there is a relationship between the estimated abilities (θ_i) and how late early sequence behavior persists in students k_i. The figure below shows θ vs. k, with a linear fit superimposed. We first note that the range of estimate k values is small. This could be an artifact of the choice of c or ε. Either way, we notice that there is modest positive association - students of higher estimated ability also tend to have higher estimated k parameters. Note that a more flexible (GAM) smoother also showed a similar relationship.

```
## `geom_smooth()` using formula 'y ~ x'
```



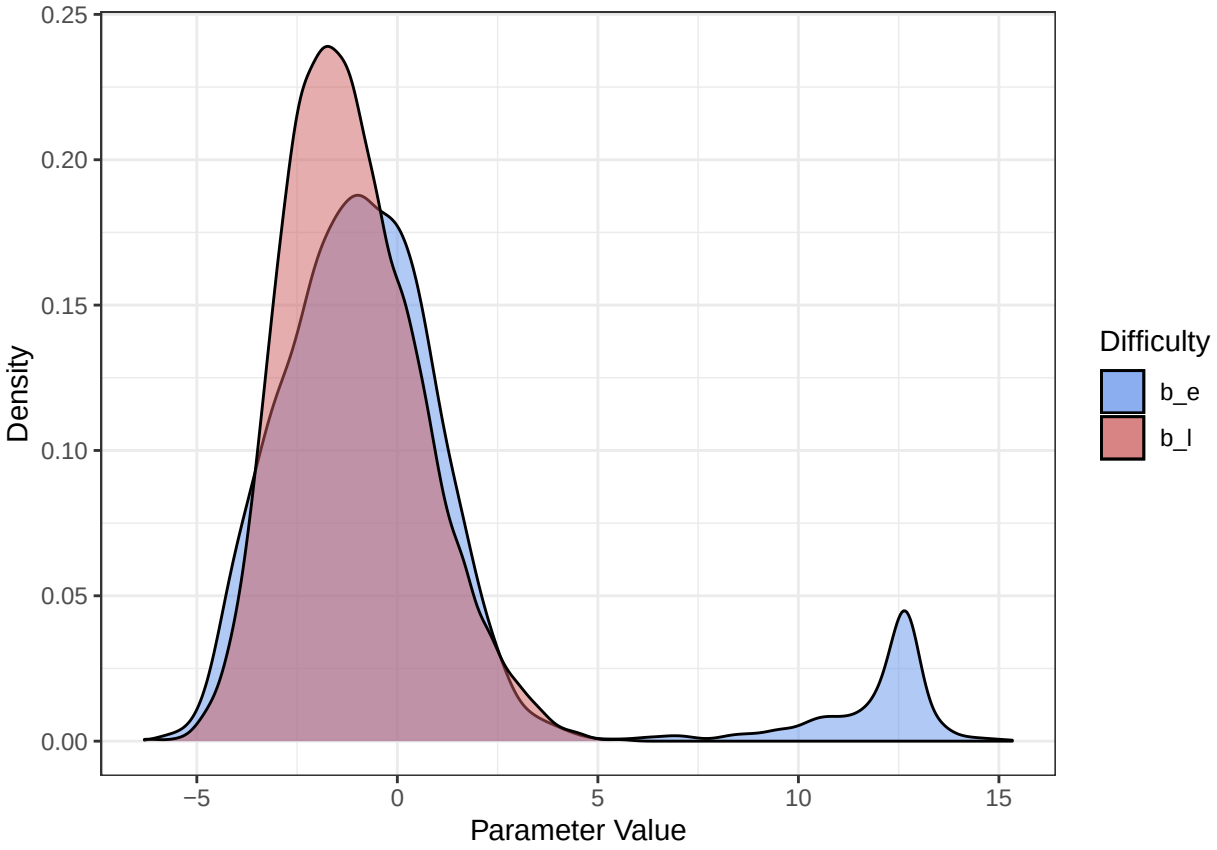
Next we regress θ on k find the relationship significant:

Table 1:

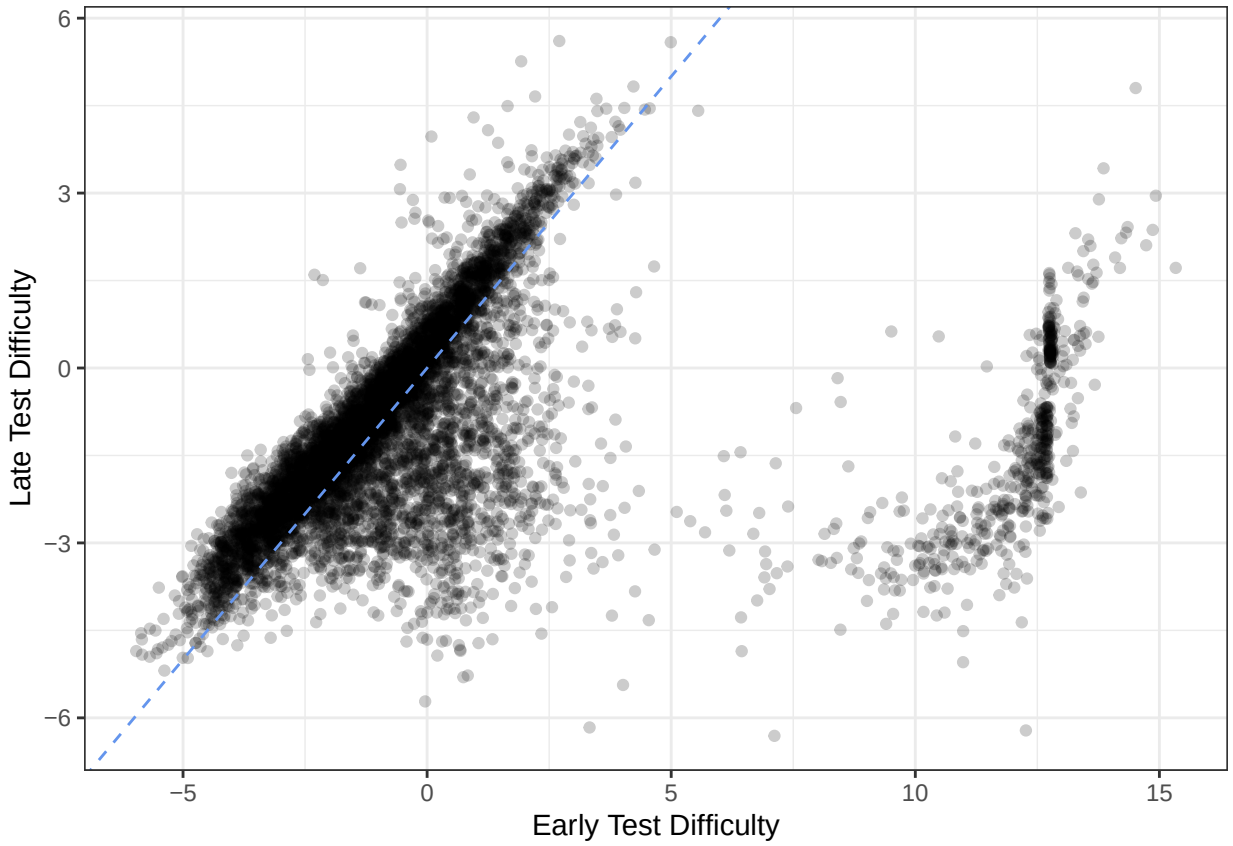
	Dependent variable:
	theta
k	0.703^{***}
	(0.015)
Constant	-13.952^{***}
	(0.295)
Observations	464,117
R^2	0.005
Adjusted R^2	0.005
Residual Std. Error	1.432 (df = 464115)
F Statistic	$2{,}267.438^{***}$ (df = 1; 464115)
Note:	*p<0.1; **p<0.05; ***p<0.01

2.2 Items

When examining items, we first look at the distribution of estimated early and late test difficulties:

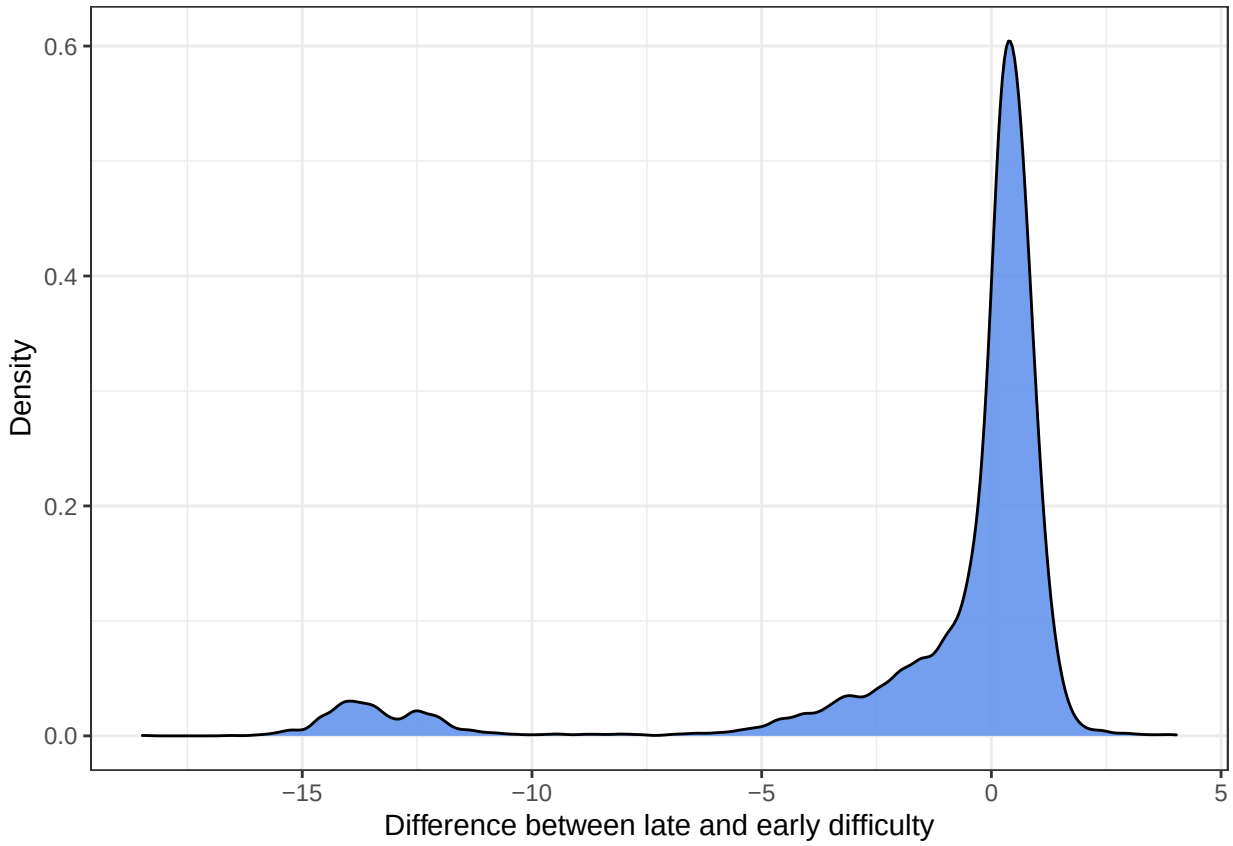


Strangely, we observe that late test difficulties appear to be lower than early test difficulties. Next we look at the relationship between estimated early test difficulties and estimated late test difficulties by item. A dashed line is overlaid to guide the eye and show where $b_{je} = b_{jl}$.



Here we see items seem to come from two distinct distributions. The first, just above the dashed line, are items that are slightly harder when they come later in the test. The second, below the dashed line, is items that are much easier when they occur later in the test.

Next we look at the magnitude of the difference between late and early test difficulty by item. First, we observe that the proportion of items estimated to be harder later in the test is 0.607. Next we look at the distribution of differences in difficulty and see, as expected from the previous plot, that items that get harder only get modestly harder, while items that get easier can get much easier:



3 Next Steps

- Reduce the value of ε. This is currently running.
- Experiment with different values of c. Is there enough data to learn a per-person c value (I don't think so)?
- Relate MMIRT estimated θ, k, b_e, b_l parameters to NWEA estimated θ, b parameters.
- Dig deeper into the estimated item parameters. Is there something about typical sequence position (are these pilot items?) or number of exposures that can account for the differences in estimated difficulties?
- The idea of modeling sequence number is meant to uncover information about different response processes students may be using. We also have response time data - how can that be included in this framework?