

Is this model for real?

Simulating data to reveal the proximity of a model to reality

Rinat B. Rosenberg-Kima¹, Zachary A. Pardos²

¹ Tel-Aviv University

rinat.rosenberg.kima@gmail.com

² University of California, Berkeley

pardos@berkeley.edu

Abstract. Simulated data plays a central role in Educational Data Mining and in particular in Bayesian Knowledge Tracing (BKT) research. The initial motivation for this paper was to try to answer the question: given two datasets could you tell which of them is real and which of them is simulated? The ability to answer this question may provide an additional indication of the goodness of the model, thus, if it is easy to discern simulated data from real data that could be an indication that the model does not provide an authentic representation of reality, whereas if it is hard to set the real and simulated data apart that might be an indication that the model is indeed authentic. In this paper we will describe analyses of 42 GLOP datasets that were performed in an attempt to address this question. Possible simulated data based metrics as well as additional findings that emerged during this exploration will be discussed.

Keywords: Bayesian Knowledge Tracing (BKT), simulated data, parameters space.

1 Introduction

Simulated data has been increasingly playing a central role in Educational Data Mining [1] and Bayesian Knowledge Tracing (BKT) research [1, 4]. For example, simulated data was used to explore the convergence properties of BKT models [5], an important area of investigation given the identifiability issues of the model [3]. In this paper, we would like to approach simulated data from a slightly different angle. In particular, we claim that the question "given two datasets could you tell which of them is real and which of them is simulated?" is interesting as it can be used to evaluate the goodness of a model and may potentially serve as an alternative metric to RMSE, AUC, and others. In a previous work [6] we started approaching this problem by contrasting two real datasets with their corresponding two simulated datasets with Knowledge Tracing as the model. We found a surprising close to identity between the real and simulated datasets. In this paper we would like to continue this investigation by expanding the previous analysis to the full set of 42 Groups of Learning Opportunities (GLOPs) real datasets generated from the ASSISTments platform [7].

Knowledge Tracing (KT) models are widely used by cognitive tutors to estimate the latent skills of students [8]. Knowledge tracing is a Bayesian model, which assumes that each skill has 4 parameters: two knowledge parameters include initial (prior knowledge) and learn rate, and two performance parameters include guess and slip. KT in its simplest form assumes a single point estimate for prior knowledge and learn rate for all students, and similarly identical guess and slip rates for all students. Simulated data has been used to estimate the parameter space and in particular to answer questions that relate to the goal of maximizing the log likelihood (LL) of the model given parameters and data, and improving prediction power [7, 8, 9].

In this paper we would like to use the KT model as a framework for comparing the characteristics of simulated data to real data, and in particular to see whether it is possible to distinguish between the real and simulated datasets.

2 Data Sets

To compare simulated data to real data we started with 42 Groups of Learning Opportunities (GLOPs) real datasets generated from the ASSISTments platform¹ from a previous BKT study [7]. The datasets consisted of problem sets with 4 to 13 questions in linear order where all students answer all questions. The number of students per GLOP varied from 105 to 777. Next, we generated two synthetic, simulated datasets for each of the real datasets using the best fitting parameters that were found for each respective real datasets as the generating parameters. The two simulated datasets for each real one had the exact same number of questions, and same number of students.

3 Methodology

The approach we took to finding the best fitting parameters was to calculate LL with a grid search of all the parameters (prior, learn, guess, and slip). We hypothesized that the LL gradient pattern of the simulated data and real data will be different across the space. For each of the datasets we conducted a grid search with intervals of .04 that generated 25 intervals for each parameter and 390,625 total combinations of prior, learn, guess, and slip. For each one of the combinations LL was calculated and placed in a four dimensional matrix. We used fastBKT [12] to calculate the best fitting parameters of the real datasets and to generate simulated data. Additional code in Matlab and R was generated to calculate LL and RMSE and to put all the pieces together².

¹ Data can be obtained here: <http://people.csail.mit.edu/zp/>

² Matlab and R code will be available here: www.rinatrosenbergkima.com/AIED2015/

4 What are the Characteristics of the Real Datasets Parameters Space?

Before we explored the relationships between the real and sim datasets, we were interested to explore the BKT parameter profiles of the real datasets. We calculated the LL with a grid search of 0.04 granularity across the four parameters resulting in a maximum LL for each dataset and corresponding best prior, learn, guess, and slip. Figure 1 present the best parameters for each datasets, taking different views of the parameters space. The first observation to be made is that the best guess and slip parameters fell into two distinct areas (see figure 1, guess x slip).

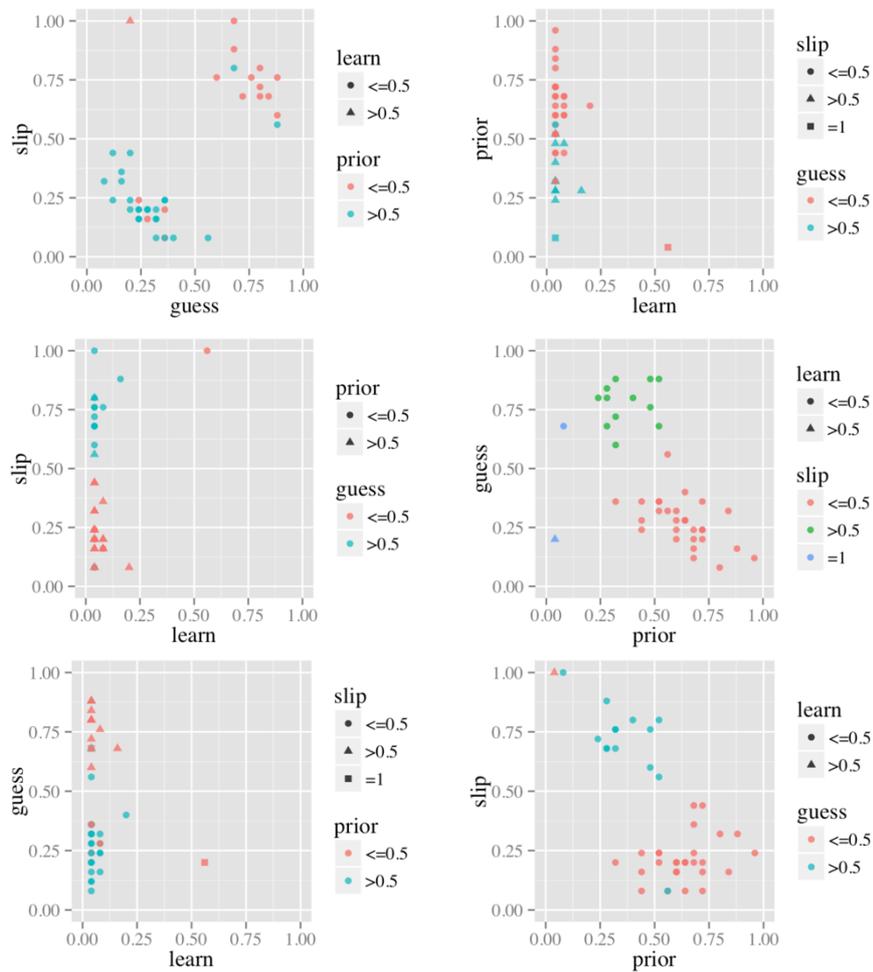


Figure 1. Best parameters across the 42 GLOP real datasets.

Much attention has been given to this LL space, which revealed the apparent co-linearity of BKT with two primary areas of convergence, the upper right area being a false, or “implausible” converging area as defined by [3]. What is interesting in this figure is that real data also converged to these two distinct areas. To further investigate this point, we looked for the relationships between the best parameters and the number of students in the dataset (see figure 2). We hypothesized that perhaps the upper right points were drawn from datasets with small number of students; nevertheless, as figure 2 reveals, that was not the case. Another interesting observation is that while in the upper right area (figure 1, guess x slip) most of the prior best values were smaller than 0.5, in the lower left area most of the prior best values were bigger than 0.5, thus revealing interrelationships between slip, guess, and prior that can be seen in the other views. Another observation is that while prior is widely distributed between 0 and 1, most of best learn values are smaller than 0.12.

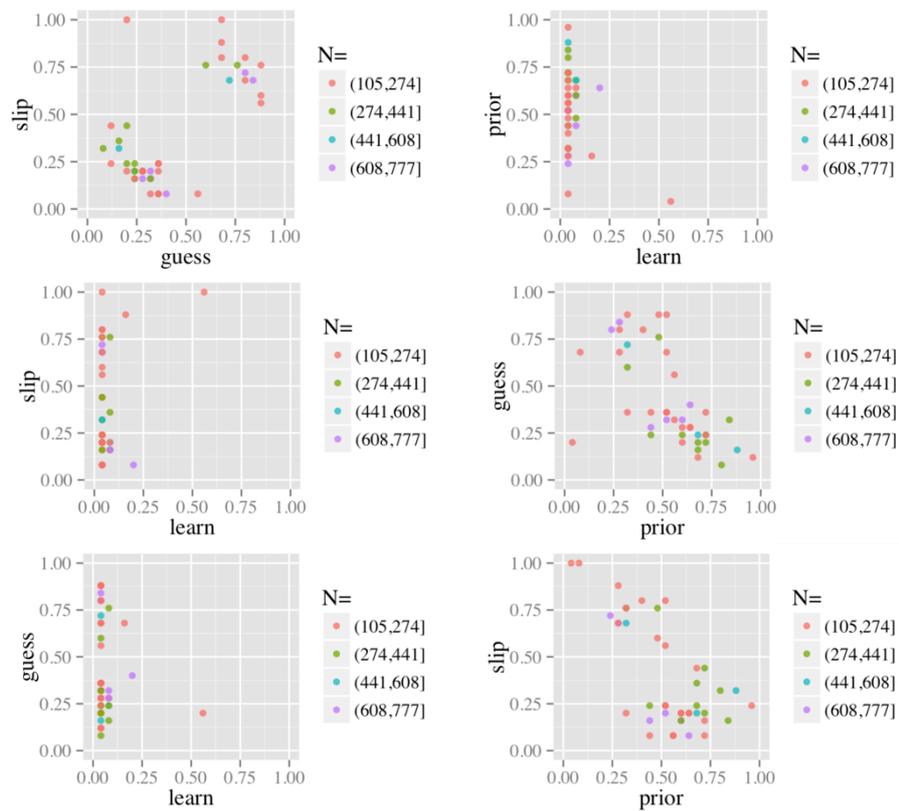


Figure 2. Best parameters across the 42 GLOP real datasets by number of students.

5 Does the LL of Sim vs. Real Datasets Look Different?

Our initial thinking was that as we are using a simple BKT model, it is not authentically reflecting reality in all its detail and therefore we will observe different patterns of LL across the parameters space between the real data and the simulated data. The LL space of simulated data in [5] was quite striking in its smooth surface but the appearance of real data was left as an open research question. First, we examined the best parameters spread across the 42 first set of simulated data we have generated. As can be seen in figure 3, the results are very similar (although not identical) to the results we received with the real data (see figure 1). This is not surprising, after all, the values of learn, prior, guess, and slip were inputs to the function generating the simulated data.

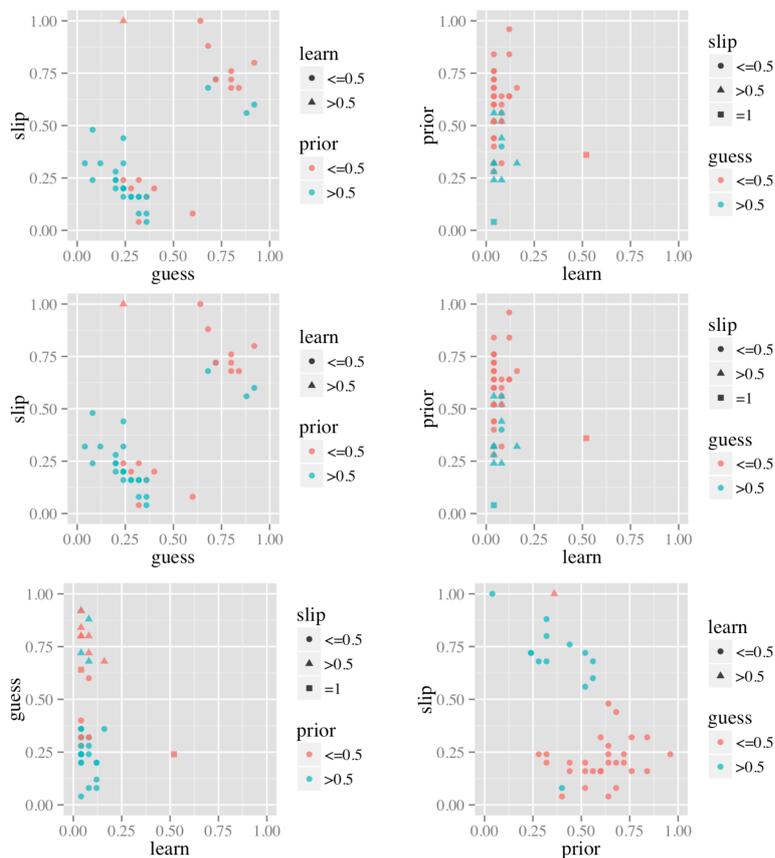


Figure 3. Best parameters across 42 GLOP simulated datasets.

In order to see if the differences between real and sim were more than just the difference between samples from the same distribution, we generated *two* simulated versions of each real dataset (sim1 and sim2) using the exact same number

of questions, number of students, generated with the best fitting parameters from the real dataset. We then visualized 2D LL heatmaps looking at two parameter plots at a time where the other two parameters were fixed to the best fitting values. For example, when visualizing LL heatmaps for the combination of guess and slip, we fixed learn and prior to be the best learn and the best prior from the real data grid search. To our surprise, when we plotted heatmaps of the LL matrices of the real data and the simulated data (the first column in figure 4 represents the real datasets, the second column represents the corresponding sim1, and the third column the corresponding sim2) we received what appears to be extremely similar heatmaps. Figure 4 and 5 displays a sample of 4 datasets, for each one displaying the real dataset heatmap and the corresponding two simulated datasets heatmaps.

The guess vs. slip heatmaps (see figure 4) prompted interesting observations. As mentioned above, the best guess and slip parameters across datasets fell into two areas (upper right and lower left). Interestingly, these two areas were also noticeable in the individual heatmaps. While in some of the datasets they were less clear (e.g., G5.198 in figure 4), most of the datasets appear to include two distinct global maxima areas. In some of the datasets the global maxima converged to the lower left expected area, as did the corresponding simulated datasets (e.g., G4.260 in figure 4), in other datasets the global maxima converged to the upper right “implausible” area, as did the corresponding simulated datasets (e.g., G6.208 in figure 4). Yet in some cases, one or more of the simulated dataset converged to a different area than that of the real dataset (e.g., G4.205 in figure 4). The fact that so many of the real datasets converged to the “implausible” area is surprising and may be due to small number of students or to other limitations of the model.

The learn vs. prior heatmaps were also extremely similar within datasets and exhibited a similar pattern also across datasets (see figure 5), although not all datasets had the exact pattern (e.g., G5.198 is quite different than the other 3 datasets in figure 5). While best learn values were low across the datasets, the values of best prior varied. As with guess vs. slip, in some cases the two simulated datasets were different (e.g., G4.205 had different best parameters also with respect to prior). Similar patterns of similarities within datasets and similarities with some clusters across datasets were also noticeable in the rest of the parameters space (learn vs. guess, learn vs. slip, prior vs. guess, prior vs. slip not displayed here due to space considerations).

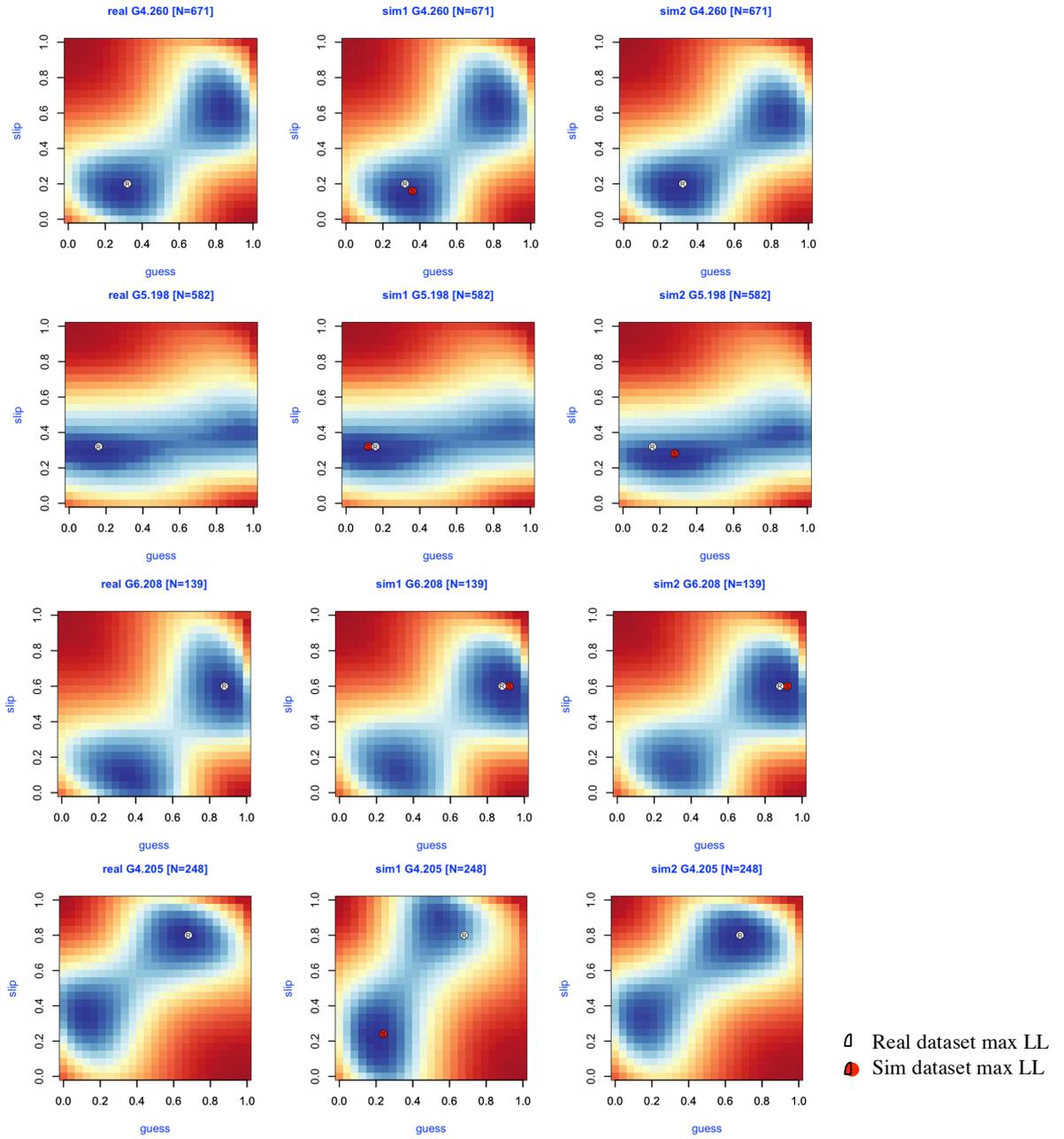


Figure 4. Heatmaps of (guess vs. slip) LL of 4 sample real GLOP datasets and the corresponding two simulated datasets that were generated with the best fitting parameters of the corresponding real dataset.

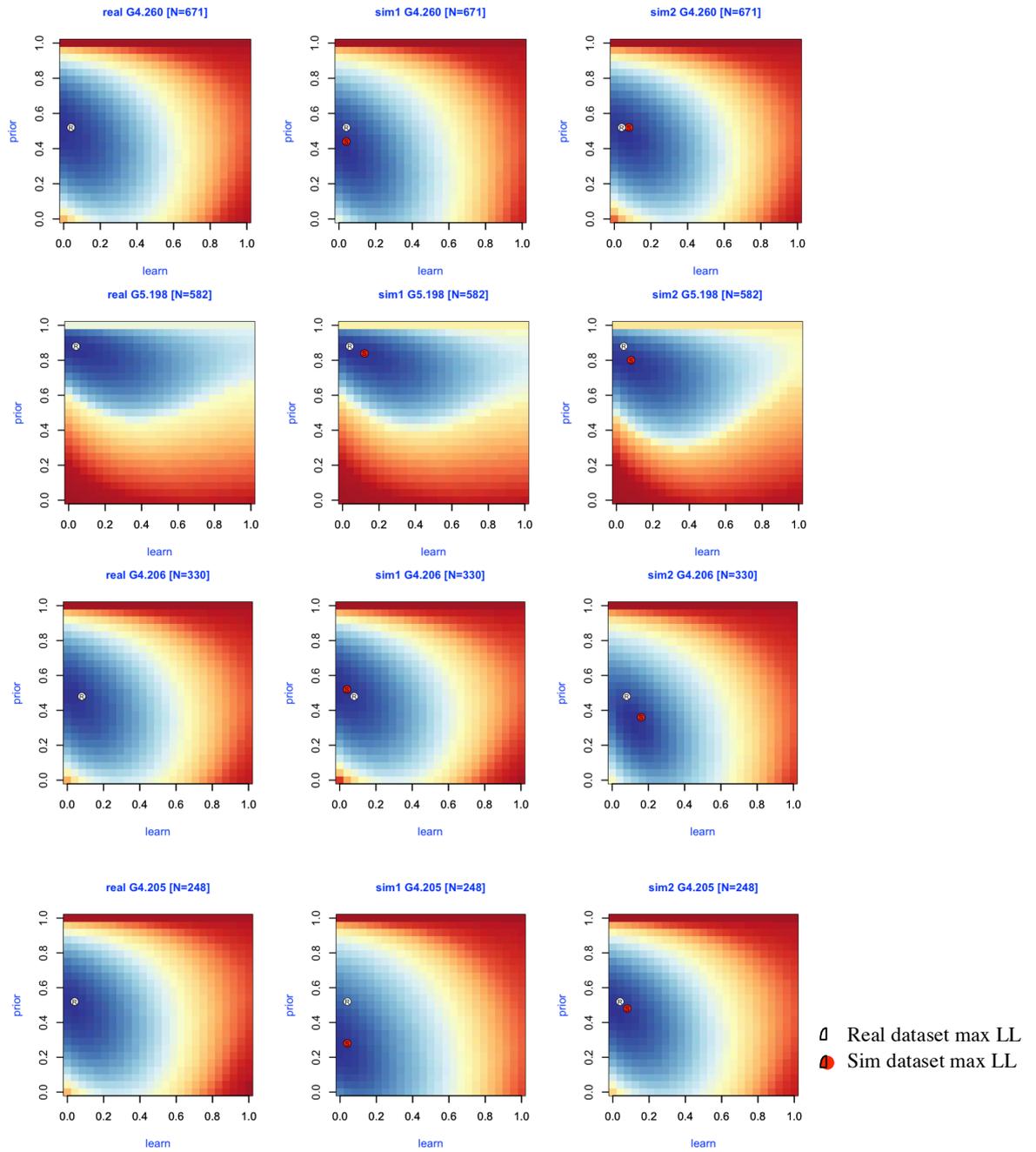


Figure 5. Heatmaps of (learn x prior) LL of 4 sample real GLOP datasets and the corresponding two simulated datasets that were generated with the best fitting parameters of the corresponding real dataset.

6 Exploring Possible Metrics Using the Real and Sim Datasets

In natural science domains, simulated data is often used as a mean to evaluate its underlying model. For example, simulated data is generated from a hypothesized model of the phenomena and if the simulated data appears to be similar to the real data observed in nature, it serves as evidence for the accuracy of the model. Then, if the underlying is validated, simulated data is used to make predictions (e.g., in the recent earthquake in Nepal a simulation was used to estimate the number of victims). Can this approach be used in education as well? What would be an indication of similarity between real and simulated data?

Figure 5 displays two preliminary approaches for comparing the level of similarity between the simulated and real data. First, the Euclidean distance between the real dataset parameters and the simulated data parameters was compared to the Euclidean distance between the two simulated datasets parameters. The idea is that if the difference between the two simulated datasets is smaller than the difference between the real and the simulated dataset this may be an indication that the model can be improved upon. Thus, points on the right side of the red diagonal indicate good fit of the model to the dataset. Interestingly, most of the points were on the diagonal and a few to the left of it. Likewise the max LL distance between the real and simulated datasets was compared to the max LL distance of the two simulated datasets. Interestingly, datasets with larger number of students did not result in higher similarity between the real and simulated dataset. Also, here we *did* find distribution of the points to the left and to the right of the diagonal.

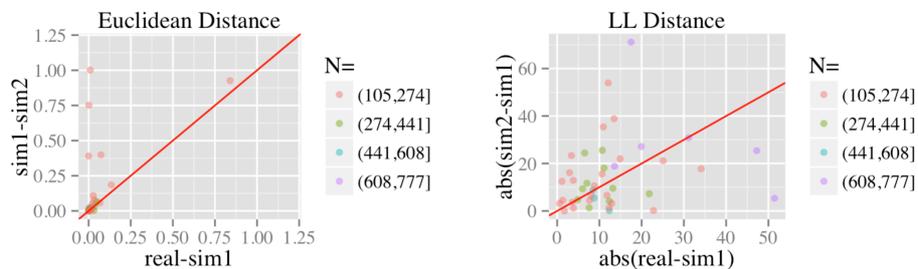


Figure 5. Using Euclidean distance and LL distance as means to evaluate the model.

7 Contribution

The initial motivation of this paper was to find whether it is possible to discern a real dataset from a simulated dataset. If for a given model it is possible to tell apart a simulated data from a real dataset then the authenticity of the model can be questioned. This line of thinking is in particular typical of simulation use in Science contexts, where different models are used to generate simulated data, and then if a simulated data has a good fit to the real phenomena at hand, then it may be possible to claim that the model provides an authentic explanation of the system [13]. We believe

that finding such a metric can serve as the foundation for evaluating the goodness of a model by comparing a simulated data from this model to real data and that such a metric could provide much needed substance in interpretation beyond that which is afforded by current RMSE and AUC measures. This can afford validation of the simulated data, which can then be used to make predictions on learning scenarios; decreasing the need to test them in reality, and at minimum, serving as an initial filter to different learning strategies.

References

- [1] R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *J. Educ. Data Min.*, vol. 1, no. 1, pp. 3–17, 2009.
- [2] M. C. Desmarais and I. Pelczer, "On the Faithfulness of Simulated Student Performance Data.," in *EDM*, 2010, pp. 21–30.
- [3] J. E. Beck and K. Chang, "Identifiability: A fundamental problem of student modeling," in *User Modeling 2007*, Springer, 2007, pp. 137–146.
- [4] Z. A. Pardos and M. V. Yudelson, "Towards Moment of Learning Accuracy," in *AIED 2013 Workshops Proceedings Volume 4*, 2013, p. 3.
- [5] Z. A. Pardos and N. T. Heffernan, "Navigating the parameter space of Bayesian Knowledge Tracing models: Visualizations of the convergence of the Expectation Maximization algorithm.," in *EDM*, 2010, pp. 161–170.
- [6] R. B. Rosenberg-Kima and Z. Pardos, "Is this Data for Real?," in *Twenty Years of Knowledge Tracing Workshop*, London, UK, pp. 141–145.
- [7] Z. A. Pardos and N. T. Heffernan, "Modeling individualization in a bayesian networks implementation of knowledge tracing," in *User Modeling, Adaptation, and Personalization*, Springer, 2010, pp. 255–266.
- [8] A. T. Corbett and J. R. Anderson, "Knowledge tracing: Modeling the acquisition of procedural knowledge," *User Model. User-Adapt. Interact.*, vol. 4, no. 4, pp. 253–278, 1994.
- [9] S. Ritter, T. K. Harris, T. Nixon, D. Dickison, R. C. Murray, and B. Towle, "Reducing the Knowledge Tracing Space.," *Int. Work. Group Educ. Data Min.*, 2009.
- [10] R. S. d Baker, A. T. Corbett, S. M. Gowda, A. Z. Wagner, B. A. MacLaren, L. R. Kauffman, A. P. Mitchell, and S. Giguere, "Contextual slip and prediction of student performance after use of an intelligent tutor," in *User Modeling, Adaptation, and Personalization*, Springer, 2010, pp. 52–63.
- [11] R. S. Baker, A. T. Corbett, and V. Aleven, "More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing," in *Intelligent Tutoring Systems*, 2008, pp. 406–415.
- [12] Z. A. Pardos and M. J. Johnson, "Scaling Cognitive Modeling to Massive Open Environments (in preparation)," *TOCHI Spec. Issue Learn. Scale*.
- [13] U. Wilensky, "GasLab—an Extensible Modeling Toolkit for Connecting Micro- and Macro-properties of Gases," in *Modeling and simulation in science and mathematics education*, Springer, 1999, pp. 151–178.