Tracking student progress in a game-like learning environment with a Monte Carlo Bayesian knowledge tracing model

G.-H. Gweon
Department of Physics
University of California
Santa Cruz, CA, USA
1-831-459-1806
gweon@ucsc.edu

Hee-Sun Lee
Department of Physics
University of California
Santa Cruz, CA, USA
1-831-459-2326
hlee58@ucsc.edu

Chad Dorsey
The Concord Consortium
25 Love Lane
Concord, MA 01742
1-978-405-3100
cdorsey@concord.org

Robert Tinker
The Concord Consortium
25 Love Lane
Concord, MA 01742
1-978-405-3225
bob@concord.org

William Finzer
The Concord Consortium
25 Love Lane
Concord, MA 01742
1-510-984-4380
wfinzer@concord.org

ABSTRACT

The Bayesian knowledge tracing (BKT) model is a popular model used for tracking student progress in learning systems such as an intelligent tutoring system. However, the model is not free of problems. Well-recognized problems include the identifiability problem or the empirical degeneracy problem. Unfortunately these problems are still poorly understood and it is not clear at all how to deal with them. The origin of these problems boils down to the difficulty to answer the question "how to best determine the four BKT parameter values, given the student activity data?". Here, we analyze the mathematical structure of the BKT model, identify a source of the difficulty, and construct a simple Monte Carlo BKT model to analyze the problem in real data. Using the student activity data obtained from the ramp game module at the Concord Consortium, we find that such a Monte Carlo BKT analysis is capable of detecting the identifiability problem and the empirical degeneracy problem, and, more generally, gives an excellent summary of of the student data. In particular, a useful by-product of this work is the identification of a student activity monitoring parameter M.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Bayesian knowledge tracing, fit, regression, educational data mining

1. INTRODUCTION

The Bayesian knowledge tracing (BKT) model [5] is widely used in the context of educational data mining [3, 4, 2, 6, 9]. It offers a simple model in which the student knowledge can be estimated as the student activity is scored in a structured online environment such as an intelligent tutoring system.

However, a major problem is that estimating student knowledge is often ambiguous, and a general method to overcome such ambiguity is not known. Recognizing this issue began with seminal papers by by Beck and Chang [3, 4], who pointed out that the BKT model, while important in the field of student modeling, suffers from a fundamental "identifiability problem." Namely, completely different sets of parameters can produce the same student performance curve.

While solutions to this well-recognized problem were sought, another type of problem, the "empirical degeneracy" problem has been pointed out as well—sometimes the model would seem to predict low knowledge on high performance or high knowledge on low performance. Contextualizing certain parameters [1] or limiting the ranges of slip and guess parameters of the model has been put to practice, to avoid these problems, and a theoretical analysis [9] has been presented. However, it remains unclear what it actually means to avoid this problem by restricting parameter values by hand.

Here, we analyze the BKT model from a simple mathematical point of view. This analysis sheds light on the identifiability problem, as it clearly shows that there is an underlying parameter degeneracy in the BKT theory. While this degeneracy has been understood by similar mathematics recently [9], we differ completely in the assessment of this issue in a real data setting. Our main message here is the finding that we must take this parameter degeneracy into account

when applying the BKT model to describe real situations. Considering this degeneracy, we propose that a new parameter, M (to be defined in Eq. 6 below), should play a central role in summarizing BKT analysis results. This parameter plays the role of summarizing the overall trend of the student activity. So, M may provide an important diagnostic parameter during real time activities. In addition, M is also related to the identifiability problem and the empirical degeneracy problem.

In this paper, we combine theoretical analysis, new algorithm, and data analysis. The new algorithm is a simple Monte Carlo BKT model, which gives a wealth of information when applied to the analysis of real data. The data are taken from the student activity data obtained from the "ramp-game module" of the Concord Consortium [7].

2. THE BKT MODEL IN THEORY

We shall briefly review the BKT model, with a focus on its mathematical structure.

2.1 The BKT model

The BKT model has been proposed by Corbett and Anderson[5]. This model involves four parameters, each of them having the numerical range from 0 to 1.

$p(L_1)$ This is the initial knowledge that a student has prior to taking on any learning activities.

$p(T)$ This is probability that the student will transition from an unknowing state to a knowing state, as the result of using the knowledge during a unit of activity.

$p(G)$ This is the "guess parameter" that corresponds to the probability that the student will choose the correct answer in an activity, while the student has not acquired the required knowledge.

$p(S)$ This is the "slip parameter" that corresponds to the probability that the student will choose an incorrect answer in an activity, while the student has acquired the required knowledge.

Here, we define n as the activity index. Then, $p(L_n)$ is the knowledge level estimated right before activity n (or, equivalently, after activity $n-1$). The activity index n goes from n goes from 1 to N, the total number of activities.

Another important curve as a function of n is given by $p(C_n)$, which is the probability that the student will get the correct answer at activity n. So, $p(C_n)$ is the student performance curve, while $p(L_n)$ is the latent knowledge.

The unobservable latent knowledge, $p(L_n)$, and the observable student performance, $p(C_n)$, are predicted in parallel by this model, and it is by fitting the observed performance data with the theoretical $p(C_n)$ curve that one can infer the student knowledge. As the student goes through activities, the typical outcome is that $p(L_n)$ increases—however, the model also allows the opposite case in which $p(L_n)$ decreases as n increases.

The knowledge parameter is updated according to the following equation

$$p(L_{n+1}) = p(L_n|E) + \big(1 - p(L_n|E)\big)p(T) \tag{1}$$

where $p(L_n|E)$ is the posterior probability that the student has the knowledge given the evidence E. Here, the evidence is the student score. The posterior probability is given by the Bayes theorem:

$$p(L_n|C_n) = \frac{(1-p(S))\cdot p(L_n)}{p(C_n)}, \tag{2}$$

$$p(L_n|I_n) = \frac{p(S)\cdot p(L_n)}{1-p(C_n)}. \tag{3}$$

Here, $p(C_n)$ is the probability that the student will get it right for activity n, and is given by

$$p(C_n) = p(L_n)\cdot(1-p(S)) + (1-p(L_n))\cdot p(G), \tag{4}$$

and $p(I_n) = 1 - p(C_n)$ is the probability that the student will get it wrong.

2.2 The BKT model without measurement

Here, we envision a theoretical process. Suppose that a BKT process as described in the previous section goes on, but the student withholds her/his answers. From the point of view of an educational researcher who likes to assess the learning, then, there are no data to analyze, since there are no "measurements" in the form of student provided answers.

Now, even in such a theoretical case, it is clear that the student likely is acquiring knowledge, as long as the student is engaged.

What would be the description of the student knowledge in such a case? While only a theoretical description is possible in such a case, the description has the benefit of being "noiseless," not suffering from statistical noises[1] that are part of the measurement process.

We call this situation as using "the BKT model without measurement." This situation corresponds to that considered by Beck and Chang [3, 4]. The level of the mathematical analysis required for this problem is basic, and it seems that the first published solution is due to van de Sande [9], with which our solution agrees.

In this measurement-less situation, the theoretical value for $p(C_n)$ is the best estimate of the real student performance. Also, it is most reasonable to put $p(L_n)$ equal to $p(L_n|E)$. Then, Eq. 1 turns into a simple recursion relation

$$p(L_{n+1}) = p(T) + (1-p(T))p(L_n). \tag{5}$$

The subsequent geometric series for $p(L_n)$ can be readily summed up, and this also makes it possible to express $p(C_n)$ (Eq. 4) in a closed form. The results are the following.

$$M \equiv (1-p(S)-p(G))\cdot(1-p(L_1)), \tag{6}$$

$$p(C_n) = 1 - p(S) - M\cdot(1-p(T))^{n-1}, \tag{7}$$

$$p(L_n) = 1 - (1-p(L_1))\,(1-p(T))^{n-1}. \tag{8}$$

[1]It is important to note that these statistical noises are due to the fundamental statistical nature of events, whether or not there is a finite value for $p(G)$ or $p(S)$.

Model	$p(L_1)$	$p(T)$	$p(G)$	$p(S)$	M
1	0.56	0.1	0.00	0.05	0.418
2	0.36	0.1	0.30	0.05	0.416
3	0.01	0.1	0.53	0.05	0.4158

Table 1: The values of the BKT model parameters used by Beck and Chang[3, 4] to generate curves shown in Fig. 1 to demonstrate the "identifiability problem" of the BKT model without measurement. The last column for a new parameter M (Eq. 6) is added in this work. All numbers are presented, assuming infinite precision.

If we define

$$n_T \equiv -\frac{1}{\log(1-p(T))} \tag{9}$$

then we can re-write our results for $p(C_n)$ and $p(L_n)$ as

$$p(C_n) = 1 - p(S) - Me^{-(n-1)/n_T}, \tag{10}$$

$$p(L_n) = 1 - (1-p(L_1))\, e^{-(n-1)/n_T}. \tag{11}$$

So, n_T tells us how fast or slow $p(C_n)$ and $p(L_n)$ approach their respective asymptotes, $1-p(S)$ and 1. It is a **scale parameter** that represents the order of magnitude for the number of activities required in order for the learning to be perfected.

These results clearly explain the origin of the identifiability problem of the BKT model without measurement. While the theory has four parameters, $p(C_n)$ depends on *only three independent parameters*, M, $p(T)$, and $p(S)$.

Indeed, when the values of M corresponding to the three models for Fig. 1 are listed in Table 1, we find that they are basically identical, explaining why Beck and Chang's three models give the same $p(C_n)$ curve. Going further, one can show that their curves shown in Fig. 1 *precisely* follow our functions for $p(C_n)$ and $p(L_n)$, Eqs. 7 and 8.

So, in fact, the identifiability problem, as pointed out by Beck[3, 4] is due to *an exact mathematical degeneracy of the BKT model without measurement*: as far as $p(C_n)$ is concerned, the parameter space is three dimensional, not four dimensional. Indeed, as also pointed out by van de Sande [9], it is not just these three models but an infinite number of models whose $p(G)$ value and $p(L_1$ value give $M = (0.95-p(G))(1-p(L_1)) = 0.418$ will give the identical result for $p(C_n)$ as model 1.

2.3 The BKT model with measurement

More commonly, though, measurements are involved. Students submit activity results, and their scores on activities are examined and their progress is checked. The student score may be boolean (true or false) or continuous (0 to 1). The latter kind can be considered a generalization of of the first kind, and so we will assume that student score s satisfies

$$0 \le s \le 1, \qquad \text{student score.} \tag{12}$$

In fact, the actual data that we analyze in this paper are of this kind.

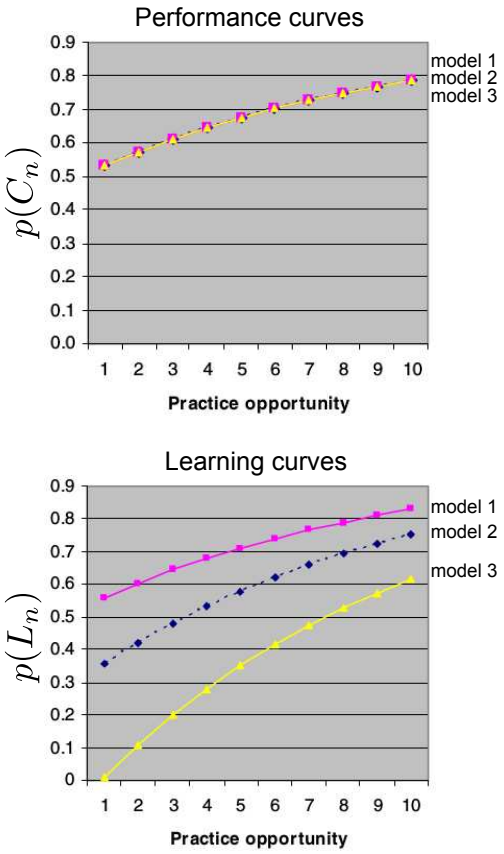

Figure 1: An example that demonstrates the identifiability problem[3, 4] of the BKT model without measurement: three identical student "performance curves" (upper panel) are generated from three entirely different "learning curves" (lower panel). The panels were taken as digital images from Beck and Chang[4]. We have re-labeled the y axes, in accordance with our symbol definition, and renamed the three models that they used simply as, simply, models 1, 2, and 3. The practice opportunity corresponds to our activity index n.

Now, based on Eqs. 2 and 3, the posterior probability given the evidence of score s is given by

$$\begin{aligned} p(L_n|s) &= s \cdot p(L_n|C_n) + (1-s) \cdot p(L_n|I_n) \\ &= \left[\frac{s \cdot (1-p(S))}{p(C_n)} + \frac{(1-s)p(S)}{1-p(C_n)}\right] p(L_n). \end{aligned} \quad (13)$$

This equation, along with Eqs. 1 and 4, then completely specify the BKT inference iteration, using which the student's actual performance can be fitted with theory and the student knowledge can be traced.

Our model here has the advantage of being general as it encompasses other models of interest in the following sense. If we replace s with the theoretical value $p(C_n)$, then we get $p(L_n|s = p(C_n)) = p(L_n)$, leading to identical results of the BKT model without measurement. If we restrict the s variable to be boolean 0 or 1, then $p(L_n|s)$ becomes either $p(L_n|C_n)$ or $p(L_n|I_n)$, and so our model reduces to the "boolean" BKT model, more common in the literature [2, 6].

Of course, Eq. 13 for the general case of s makes the recursion relation used in the previous section, Eq. 5, no longer valid. Accordingly, all results of the previous section no longer apply. This immediately leads to the following question.

Is the identifiability problem absent in the BKT model *with* measurement, then?

Since all BKT models used in practice are ones with measurements, this question is a very important one. As we will show in this paper, with real data, the answer is no, in general. The identifiability problem does persist even in the BKT model with measurement, unless some other feature of the data places a strong constraint on $p(L_1)$ or $p(G)$.

From a theoretical point of view, also, it seems a bit too optimistic to conclude [9] that the identifiability problem does not exist in the BKT model with measurement just because the model now involves all four parameters in predicting the student performance. The reason why all four parameters are involved is the probabilistic nature of the measurement process, or the probabilistic nature of a student score. Then, it is to be expected that when a few student scores are examined and fit with theory, at least *some* of the underlying identifiability problem shows up, since the fit strives to describe the average behavior of the data.

3. THE BKT MODEL IN PRACTICE

Motivated by the ideas put forth in the previous section, here we put the BKT model to practice, using a simple Monte Carlo BKT model. As we shall see, this model provides answers to the important "identifiability problem question" posed near the end of the previous section, and gives other useful insights as well.

3.1 Practical meanings of BKT fit parameters

The fitting of data with a complex model (non-linear models with many parameters) is a non-trivial task [10] with many potential pitfalls. As a rule of thumb, if each fit parameter used cannot be identified with the practical description of a salient feature of the data, fit results may well turn out to be unreliable as there is a high chance that some fit parameters are redundant or mis-used.

Therefore, it seems important to discuss the practical meanings of the BKT parameters at the outset. To do so, let us take note of the fact that the following equation is valid for the BKT model with or without measurement (see Eq. 4 or 7):

$$p(C_1) = 1 - p(S) - M. \quad (14)$$

This describes the initial value of the performance curve. In practice, there may be noises in the data, and $p(C_1)$ must be assessed with such noise filtered out.

Now, let us assume that the following **positive eventual outcome scenario** is realized.

$$p(L_{n\to\infty}) \to 1, \quad (15)$$

$$p(C_{n\to\infty}) \to 1 - p(S). \quad (16)$$

The second equation follows from the first for the BKT model with or without measurement (see Eq. 4 or 7). It should be noted that here $n \to \infty$ really means $n \gg n_T$, and so the actual n value for which these positive asymptotic behaviors show up can be quite small: sometimes $n = 4$ or $n = 3$ may be large enough, as we will show in our actual data analysis.

Assuming that the positive eventual outcome scenario is realized, the following is a summary of the meaning of BKT parameters from the practical point of view.

1. $p(S)$ is the difference between 1 and the average student score in the end.
2. T gives the scale parameter n_T (Eq. 9), which describes how many activities it takes for $p(C_n)$ (and $p(L_n)$) to expnentially approach the final behavior.
3. M is the difference between $1 - p(S)$ and the average student score in the beginning.

More importantly, perhaps, the M parameter can be used in a more general context to give the following information.

1. If the value of M is large and positive, say $1/3 \leq M \leq 1$, then it means that *the learning is progressing* as evidenced by measured student activities.
2. If M is close to zero, say $-1/3 < M < 1/3$, then *the learning is stalling* as evidenced by measured student activities. This includes the case when the student learns quickly and there is nothing more to learn.
3. If M is large and negative, say $-1 \leq M \leq -1/3$, then *the learning is regressing* as evidenced by measured student activities.

As we will see, this conjecture[2] regarding the meaning of M gains support from our data. Indeed, it is a main claim of

[2]This is a conjecture only in the BKT model with measurement. In the BKT model with no measurement, it is a readily provable statement.

this paper that the value of M can be used in this fashion to "monitor student activities with one number."

In other words, M can be regarded as a sort of **student activity monitoring parameter** or a **student activity quality measure**. It must be kept in mind that even in cases 2 (stalling) and 3 (regressing) with imperfect performance scores, it is possible that student latent knowledge is in fact increasing. If so, then the knowledge is apparently not firm enough to be clearly evidenced through measurements.

Before we discuss our actual data analysis, here are some points to keep in mind if we consider that the identifiability problem, as discussed near the end of the previous section, may persist in the actual data analysis.

1. Generally, the BKT analysis is *not* able to determine values of $p(L_1)$ and $p(G)$ precisely. In such a case, M is a better quantity to discuss.
2. There may be exceptions where certain features of the data place more stringent constraints on parameters. An example is when the initial value of $p(C_n)$ is very small. In this case, both $p(L_1)$ and $p(G)$ *are necessarily* very small and can be determined with high precision (by Eqs. 14 and 6).
3. The indeterminate values of $p(L_1)$ and $p(G)$, and better determined value of M, would mean a negative correlation between $p(L_1)$ and $p(G)$ according to Eq. 6.

3.2 Ramp game data

Here, we briefly discuss the set-up for the "ramp game module," whose activity logs were the source of the student score data analyzed here. The ramp game module is discussed in more detail elsewhere [7].

In the ramp task, students were asked to determine a height so that the car could land on a particular location. The ramp task consisted of five challenges requiring students to apply more and more sophisticated knowledge about the ramp system as follows.

Challenge 1: relationship between height and a fixed landing location.

Challenge 2: relationship between height and moving landing locations.

Challenge 3: relationship between height and moving landing locations when a friction value is changed from the previous challenge.

Challenge 4: relationship between height and moving landing locations when mass of the car is changed.

Challenge 5: relationship between friction and moving landing locations when starting height and mass are fixed.

It is important to note that each challenge level was aimed at teaching and testing a single concept as listed above, making the ramp game module very suitable for the BKT analysis. Each challenge level was comprised of multiple steps: 3 steps for Challenges 1 and 4; 4 steps for Challenges 2 and 3; 6 steps for Challenge 5. Students' performances were scored automatically on a 0 to 100 scale based on how close the car landed from the specified landing location. If students scored 75 points or higher, then they were allowed to move to the next step within the level. If students finished all required steps within the level, they moved to the first step of the next Challenge.

For our analysis, students' scores were normalized to a 0 to 1 scale, in accordance with Eq. 12.

Finally, note that in ramp game activities, a small group of students worked as a unit. So, when we say "a student" in this paper in reference to ramp game data, what we mean is in fact a group of students.

3.3 Monte Carlo BKT

Given the theoretical ideas discussed so far, how might one extract fit parameter values from the data? Here, we propose a simple Monte Carlo method. The idea is well-motivated. Allowing for a possibility that there are multiple or infinite minima of fit residuals (χ^2; or, multiple maxima of the likelihood function, more generally), we must approach the fitting of even a single data set with *distributions* of fit parameter values in mind. For instance, if indeed there is a massive degeneracy for values of $p(L_1)$ and $p(G)$, then there will be infinite combinations of these values at which χ^2 is a local minimum.

Our Monte Carlo BKT algorithm is the following.

1. A random set of fit parameters, $p(G)$, $p(L_1)$, $p(S)$, and $p(T)$ are chosen and the standard Levenberg-Marquardt non-linear least squares fit algorithm [8] is applied using those values as the initial values for finding a local minimum of χ^2.
2. Successfully converged fit results are collected. At least a certain minimum number of good fits, 200 for this work, are required, to ensure statistics.
3. If the average value of each and every parameter value converge within tolerance, 5e-4 for this work, then the program is stopped and the success is declared.

After the program stops successfully, the average fit parameter values can be taken as representing the given data set. Those average parameter values are the ones used in Ref. [7][3].

In comparison with past work, it seems that our work can be regarded as a generalized version of the brute force grid search mechanism used in Ref. [1], where the minimum χ^2 was identified by a grid search. However, it is important to note that in our work, we do not place any restriction on parameter values: any of the four parameters listed above

[3]For that work, however, tolerance 1.e-3 was used; the different tolerance does not affect the discussion of parameter values in that work in any substantial way.

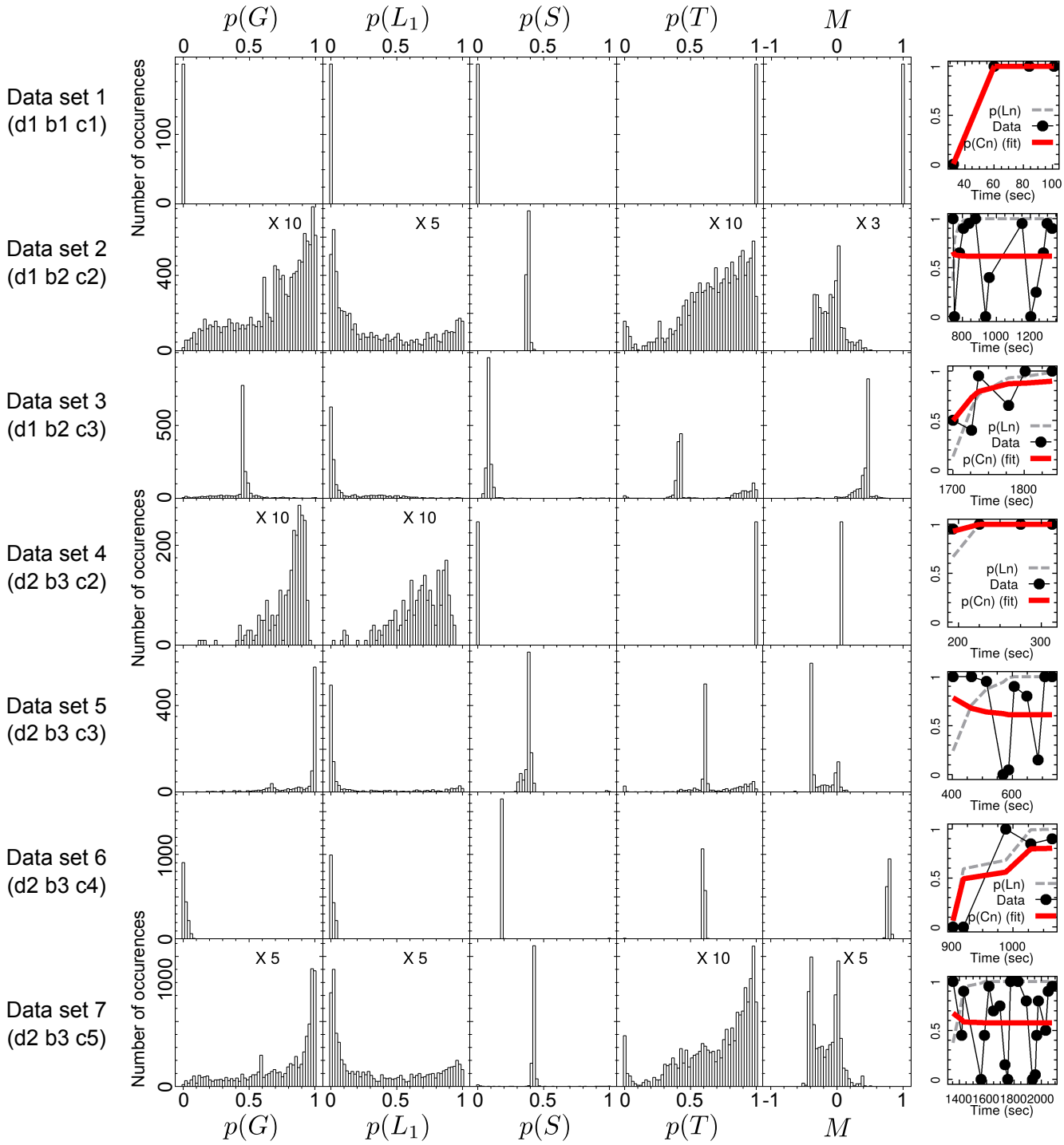

Figure 2: Monte-Carlo samples of BKT parameters, $p(G)$, $p(L_1)$, $p(S)$, $p(T)$, and M. M is a derived parameter, defined by Eq. 6. Each data set is labeled as "d<l> b<m> c<n>" which stands for "day <l> batch <m> challenge/level <n>" in the notation of Ref. [7]. For each data set, we show the histograms for Monte Carlo samples of fit parameters. Some histograms were scaled up by the shown factors, for clarity. On the right most column, data (connected black symbols) are shown with the fit (thick solid red lines) and the knowledge estimate (gray dashed lines), calculated using the averaged fit parameter values. The data are plotted as a function of time after session login by student. The sample sizes (N_{total}) that were required for convergence and other basic statistics are given in Table 2 in numerical forms. Some interesting statistics are visualized and analyzed further in Fig. 3.

is given an initial value randomly chosen between 0 and 1. This makes our analysis completely unbiased.

The results of the Monte Carlo BKT fit are shown in Fig. 2 and Table 2.

Fig. 2 shows various behaviors of the data captured by the analysis. For instance, we would characterize the data set 1 as "an unfamiliar but easy activity set." Apparently, the student does not "know it" before the first try, but immediately gets it after the first try, which means that $T \to 1$ and $n_T \to 0^+$. So, it takes no time for the student to learn, and the analysis results are very clean—there are no ambiguities for any parameter values. The next data set in terms of the sure analysis is data set 6. While data set 6 is clearly noisier than data set 1, what is common between the two data sets is that the initial value of $p(C_n)$ is practically zero, which places a strong constraint that $p(G)$ and $p(L_1)$ are small. All parameter distributions are very sharp. For data sets 1 and 6, the value of M is at or near its maximum value, 1.

The next data set of interest is data set 3. This data set shows a gradual improvement of performance, like data sets 1 and 6, but the main difference is that $p(C_n)$ starts at around 0.5. For this data set, all parameters have non-trivial distributions, while all distributions remain rather sharp. The value of M is moderate (0.4), but it is still in the "progress" domain (see Section 3.1), consistent with the positive trend shown in the overall performance pattern.

The next data set to discuss is data set 4, providing a very clear demonstration of the "identifiability problem" in real data. Here, parameters $p(S)$, $p(T)$ and M are determined with no ambiguity. However, parameters $p(L_1)$ and $p(S)$ are very broadly distributed. In comparison to data set 1, the only difference for this data set is that the first point of $p(C_n)$ is at a large value. This causes a great uncertainty in estimating values of $p(G)$ and $p(L_1)$, since there is no way of telling whether the first point is due to guessing or high initial knowledge. The value of M in this case is close to 0. So, the student is stalling, but in a high knowledge state.

All other data sets have negative values of M on average, and accordingly we might suspect that these are the data sets for which the activities did not go very smoothly (student is regressing). For these data sets, the data might be indicative of confusion. It is however remarkable that even for these "confused" data sets, the values of $p(S)$ are sharply defined.

Table 2 summarizes simple statistics for the fit parameters, including the total number of good fits N_{total} required for the convergence of the Monte Carlo iteration. Not surprisingly, data with practically no noise (data sets 1 and 4) converge very quickly, while others require many more iterations[4].

In view of our discussion in Section 3.1, the correlation between $p(G)$ and $p(L_1)$ is interesting to look at and it is listed in the last column. This correlation is almost always negative, supporting our idea that for a single value of M there is a significant degeneracy for the values of $p(G)$ and $p(L_1)$. The classic example among our data sets is data set 4, which shows the largest negative correlation, -0.79, when M is sharp with no ambiguity. For other data sets also, negative correlations are found except for one data set. An exception is found for data set 6, which is characterized by a very large positive correlation, as well as very small values of $p(G)$ and $p(L_1)$. As explained at the end of Section 3.1, the degeneracy associated with M will disappear in the limit of vanishing $p(C_1)$, which necessarily leads to the vanishing of both $p(G)$ and $p(L_1)$, and so the exception found for data set 6 does not raise any issue.

[4]The number of iterations required here can be reduced greatly, if the goal is to obtain the average value only.

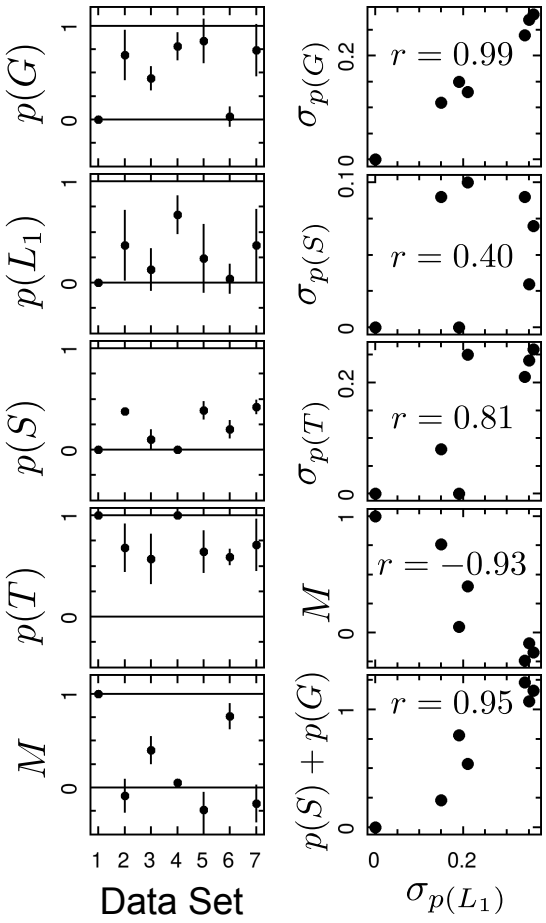

Figure 3: The basic statistics of fit parameters in Table 2 are summarized in the left column. Short vertical bars at data points represent the standard deviations. The standard deviation for random variable x is denoted as σ_x ($x = p(G)$, $p(L_1)$, $p(S)$, $p(T)$, and M) in the right column. Various standard deviation values or other parameters of interest (M and $p(S) + p(G)$) are plotted as a function of $\sigma_{p(L_1)}$ and the correlation coefficient (r) for the data plotted in each plot is also reported.

4. DISCUSSION

It seems that a few good insights are gained already on topics of high interest from our analysis.

4.1 There is an identifiability problem

Data	N_{total}	$p(G)$	$p(L_1)$	$p(S)$	$p(T)$	M	Corr(G, L_1)
1	200	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)	1.00 (0.00)	-0.61
2	1211	0.69 (0.27)	0.37 (0.35)	0.38 (0.03)	0.68 (0.24)	-0.09 (0.18)	-0.49
3	1564	0.44 (0.13)	0.13 (0.21)	0.10 (0.10)	0.57 (0.25)	0.40 (0.15)	-0.20
4	247	0.78 (0.15)	0.67 (0.19)	0.00 (0.00)	1.00 (0.00)	0.05 (0.00)	-0.79
5	1217	0.84 (0.24)	0.24 (0.34)	0.39 (0.09)	0.64 (0.21)	-0.24 (0.19)	-0.77
6	1725	0.03 (0.11)	0.04 (0.15)	0.20 (0.09)	0.59 (0.08)	0.76 (0.14)	0.89
7	1367	0.74 (0.28)	0.37 (0.36)	0.42 (0.07)	0.71 (0.26)	-0.17 (0.20)	-0.66

Table 2: Basic statistics for Monte Carlo BKT fit parameters summarized in Fig. 2. Average values are presented with standard devisions (in parentheses).

Data set 4 provides a clear proof for this statement. However, it must be noted that for some data sets there is no identifiability problem. Data sets 1 and 6 are examples.

Generally speaking, as the data become noisier, fit parameter distributions become broad, giving rise to substantial identifiability problem. However, $p(S)$ remains very robust, in comparison to other parameters. This is easy to notice from the left column of Fig. 3.

4.2 The $p(S)$ parameter estimate is robust

This is related to the fact that the positive eventual outcome scenario (Eqs. 15, 16) tends to be realized when the BKT analysis is applied. To see this point, note that for all plots shown in the right most column of Fig. 3, $p(L_n)$ has reached 1 at the end.

We can trace the reason back to Eq. 8, which shows that it is the nature of the BKT model that, if enough attempts are made, the latent knowledge eventually approaches 1. So, $p(S)$ becomes the only parameter that fits the behavior at large n, which is the reason why it can be so robustly estimated.

Since the latent knowledge approaches 1 at the end in all cases shown of Figure 2, should we say that all students acquired perfect knowledge? Clearly, this is not the case [7]. It seems that a practical choice is to take $1 - p(S)$ as the *demonstrable* level of expertise on the subject. In other words, $1 - p(S)$ might be taken as the *practical* or *demonstrable* knowledge, as opposed to the latent knowledge.

4.3 The empirical degeneracy can be detected

The concept of the empirical degeneracy [2] was discussed briefly in the introduction. In the literature, certain procedures such as limiting the values of $p(S)$ and $p(G)$ to small values have been employed to avoid this problem.

Our analysis shows that such a procedure is unnecessary, as the empirical degeneracy is detected from our analysis itself. In other words, the empirical degeneracy need not be avoided in the analysis by hand, since it is automatically detected.

The theoretical analysis of the fixed point behavior of the BKT inference iteration [9] shows that the empirical degeneracy condition corresponds to $p(S) + p(G) > 1$. According to this condition, the fits to our data sets 2, 5, and 7 show the empirical degeneracy.

What are the common features of our fits for these data sets? The answer is that the BKT fit shows that the knowledge is increasing while the performance is decreasing (and leveling off)! Indeed, one might call this a questionable result. It seems that these two opposing trends can be used as the *definition* of the "empirical degeneracy," similar in spirit to other definitions in the literature [2, 9].

If we take our new definition that the empirical degeneracy refers to the increasing knowledge with decreasing performance, then it must mean that $p(S)$ is large (to explain the low performance at end), $p(G)$ is large (to explain the initial high performance), and M is negative (since the overall student performance is decreasing). Then, one might guess that M or $p(S)+p(G)$ are highly correlated. Indeed, we find that the correlation between their values computable from Table 2 is -0.99, not surprising at all given the definition of M (Eq. 6).

A bit more interesting finding is that the empirical degeneracy and the identifiability problem are very highly correlated. In the right column of Fig. 3, we plot the width estimates of $p(L_1)$ distributions ($\sigma_{p(L_1)}$) with various quantities of interest. (1) The fact that $\sigma_{p(G)}$ has an extreme high correlation (0.99) with $\sigma_{p(L_1)}$ corroborates our finding that this is due to the identifiability problem. (2) The fact that $p(S)$ has the weakest correlation with $\sigma_{p(L_1)}$ corroborates our finding that $p(S)$ is a robust quantity, determined more or less independently of all other quantities. (3) Lastly, it shows that the M parameter has a correlation close to -1 with $\sigma_{p(L_1)}$ and the empirical degeneracy parameter $p(S)+p(G)$ has a perfect correlation with $\sigma_{p(L_1)}$.

4.4 The nature of data

As the analysis in this work is performed for continuous data sets (Eq. 12), while many researches deal with boolean data sets [1, 6], a question arises. How might our findings be applied to boolean data cases?

It seems a safe assumption that boolean data are more susceptible to noises than continuous data sets, and so it is presumably the case that such data sets are more difficult to analyze and more prone to identifiability and empirical degeneracy problems. However, other than the inherently greater noise, it is not clear whether there are any fundamental barriers to applying the current analysis framework to boolean data.

5. CONCLUSIONS

In this paper, we presented data from student on-line learning activities and their analysis using a Monte Carlo BKT model. We also analyzed the BKT model mathematically.

Our work shows that it is possible to approach the BKT model without any initial constraint or bias. We do not have to try and avoid the identifiability problem or the empirical degeneracy problem, since they are detected by the analysis.

Using our analysis, we can distinguish robust parameters from indeterminate parameters of the model. We can also detect problems in student activities, through the parameter M. The numerical procedure takes about 20 seconds of wall clock time per data set on a typical portable computer and so it is amenable to real time implementation in educational settings.

It is also an important finding of ours that the slip parameter $p(S)$ is the most robustly estimated parameter in the BKT model, and it is an important parameter to use for assessing the outcome of the lesson.

6. ACKNOWLEDGMENTS

We thank ...

7. REFERENCES

[1] R. S. 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*, pages 52–63. Springer, 2010.

[2] R. S. J. d. Baker, A. T. Corbett, and V. Aleven. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In B. P. Woolf, E. AÃŕmeur, R. Nkambou, and S. Lajoie, editors, *Intelligent Tutoring Systems*, number 5091 in Lecture Notes in Computer Science, pages 406–415. Springer Berlin Heidelberg, Jan. 2008.

[3] J. E. Beck. Difficulties in inferring student knowledge from observations (and why you should care). In *Educational Data Mining: Supplementary Proceedings of the 13th International Conference of Artificial Intelligence in Education*, pages 21–30, 2007.

[4] J. E. Beck and K.-m. Chang. Identifiability: A fundamental problem of student modeling. In C. Conati, K. McCoy, and G. Paliouras, editors, *User Modeling 2007*, number 4511 in Lecture Notes in Computer Science, pages 137–146. Springer Berlin Heidelberg, Jan. 2007.

[5] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, Dec. 1994.

[6] J. D. Gobert, M. Sao Pedro, J. Raziuddin, and R. S. Baker. From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, 22(4):521–563, Sept. 2013.

[7] H.-S. Lee, G.-H. Gweon, C. Dorsey, R. Tinker, and W. Finzer. Ramp game paper. This journal, 2015.

[8] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing, Second Edition*. Cambridge University Press, Cambridge ; New York, 2 edition edition, Oct. 1992.

[9] B. v. d. Sande. Properties of the bayesian knowledge tracing model. *JEDM - Journal of Educational Data Mining*, 5(2):1–10, July 2013.

[10] M. K. Transtrum, B. B. Machta, and J. P. Sethna. Why are nonlinear fits to data so challenging? *Physical Review Letters*, 104(6):060201, Feb. 2010.