Accounting for Slipping and Other False Negatives in Logistic Models of Student Learning

Christopher J. MacLellan
Human-Computer Interaction
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213
cmaclell@cs.cmu.edu

Ran Liu
Human-Computer Interaction
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213
ranliu@andrew.cmu.edu

Kenneth R. Koedinger
Human-Computer Interaction
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213
koedinger@cmu.edu

ABSTRACT

Additive Factors Model (AFM) and Performance Factors Analysis (PFA) are two popular models of student learning that employ logistic regression to estimate parameters and predict performance. This is in contrast to Bayesian Knowledge Tracing (BKT) which uses a Hidden Markov Model formalism. While all three models tend to make similar predictions, they differ in their parameterization of student learning. One key difference is that BKT has parameters for the slipping rates of learned skills, whereas the logistic models do not. Thus, the logistic models assume that as students get more practice their probability of correctly answering monotonically converges to 100%, whereas BKT allows monotonic convergence to lower probabilities. In this paper, we present a novel modification of logistic regression that allows it to account for situations resulting in false negative student actions (e.g., slipping on known skills). We apply this new regression approach to create two new methods AFM+Slip and PFA+Slip and compare the performance of these new models to traditional AFM, PFA, and BKT. We find that across five datasets the new slipping models have the highest accuracy on 10-fold cross validation. We also find evidence that the slip parameters better enable the logistic models to fit steep learning rates, rather than better fitting the tail of learning curves as we expected. Lastly, we explore the use of high slip values as an indicator of skills that might benefit from skill label refinement. We find that after refining the skill model for one dataset using this approach the traditional model fit improved to be on par with the slip model.

Keywords

Cognitive Modeling, Statistical Models of Learning, Additive Factors Model, Performance Factors Analysis, Knowledge Tracing

1. INTRODUCTION

Statistical models of student learning make it possible for Intelligent Tutoring Systems [18] to be adaptive. These models estimate students' latent skill knowledge, so that tutors can use these estimates to intelligently select problems that give students more practice on skills that need it. Prior work has shown that even minor improvements in the predictive fit of latent knowledge models can result in less "wasted" student time, with more time on effective practice [22].

Two popular models of student learning are the Additive Factors Model (AFM) [4] and Performance Factors Analysis (PFA) [16]. Both are extensions of traditional Item Response Theory models [8]. While the two models differ in their parameterization of student learning, they both utilize logistic regression to estimate parameters and predict student performance. These models stand in contrast to other popular approaches like Bayesian Knowledge Tracing (BKT) [7], which uses Hidden Markov Modeling.

The BKT model is used both for "online" knowledge estimation within Intelligent Tutoring Systems (e.g., in Carnegie Learning's Cognitive tutor) to adaptively selecting practice items and for "offline" educational data modeling. The logistic models, on the other hand, have mainly been used in the context of offline data modeling. For example, DataShop, the largest open repository of educational data [12], uses AFM to fit student performance within existing datasets and to generate predicted learning curves. Data-driven cognitive task analyses, i.e., discovering and testing new mappings of tutor items to skills (or knowledge components), have used AFM as the core statistical model [17]. Novel knowledge component models can be discovered, evaluated in conjunction with AFM as a statistical model, validated on novel datasets [14], and used to guide tutor redesign efforts [13].

Despite the success of approaches like AFM, its lack of slip parameters has been emphasized as a key reason for favoring knowledge tracing over logistic models [10]. But knowledge tracing models tend to suffer from identifiability problems [1, 2]; e.g., the same performance data can be fit equally well by different parameters values, with different implications for system behavior. Furthermore, the actual effect of slip parameters on model predictions is complicated. The guess and slip parameters in BKT serve the dual purpose of modeling both noise, and the upper and lower bounds, in student performance. Without slip parameters, if a student gets an answer wrong, then BKT must assume that the student has not yet learned the skill. In contrast, the logistic models just model noise in the observations, so as long as the average student success rate converges to 100% then both models should perform similarly (assuming all other parameters are comparable across models). These approaches should only differ in situations where student performance converges to lower probabilities at higher opportunities; i.e., where false negatives such as slipping are actually occurring.

To investigate false negative phenomena, we augmented the logistic regression formalism to support slipping parameters. Using this new approach, which we call Bounded Logistic Regression, we produce two new student learning models: Additive Factors Model + Slip (AFM+Slip) and Performance Factors Analysis + Slip (PFA+Slip). These models are identical to their traditional counterparts but have additional parameters to model the false negative rates for each skill. We compare these models to their traditional counterparts and to BKT on five datasets across the domains of Geometry, Equation Solving, Writing, and Number Line Estimation. In all cases, the slip models have higher predictive accuracy (based on 10-fold cross validation) than their traditional counterparts.

We then move beyond comparing the predictive accuracies of the models to investigate how these parameters affect the predictions of the models and *why* these models are more accurate. Our analyses suggest that slipping parameters are not used to capture actual student "slipping" behavior (i.e., non-zero base rates for true student errors) but, rather, make the logistic models more flexible and allow better modeling of steeper learning rates while still predicting performance accurately at high opportunity counts (in the learning curve tail).

Lastly, we use AFM+Slip to perform data-driven refinement of the knowledge component (KC) model for a Geometry dataset. We identified a KC with a high false negative, or slip, rate and searched for ways to refine it. Using domain expertise, we refined the underlying KC model and showed that the traditional model (AFM) with the new KC model performed as well as the comparable slip model (AFM+Slip) did with the original KC model. This suggests that slip parameters allow the model to compensate for, and identify, an underspecified KC model.

2. STATISTICAL MODELS OF LEARNING

2.1 Logistic Models

The models in this class use logistic regression to estimate student and item parameters and to predict student performance. Thus, they model the probability that a student will get an step i correct using the following logistic function:

$$p_i = \frac{1}{1+e^{-z_i}}$$

where z_i is some linear function of student and item parameters for step i. The likelihood function for these models has been shown to be convex (i.e., no local maximums), so optimal parameter values can be efficiently computed and issues of identifiability only occur when there are limited amounts of data for each parameter. There are many possible logistic student learning models; in fact, most Item Response Theory models are in this class. For this paper, we will focus on two popular models in the educational data mining community: Additive Factors Model [4] and Performance Factors Analysis [16].

2.1.1 Additive Factors Model

This model utilizes individual parameters for each student's baseline ability level, each knowledge component's baseline difficulty, and the learning rate for each knowledge component (i.e., how much improvement occurs with each additional practice opportunity). The standard equation for this model is shown here:

$$z_i = \alpha_{student(i)} + \sum_{k \in KCs(i)} (\beta_k + \gamma_k \times opp(k,i))$$

where $\alpha_{student(i)}$ represents the prior knowledge of the student performing step i, the βs and γs represents the difficulty and learning rate of the KCs needed to solve step i, and $opp(k,i)$ represents the number of prior opportunities a student has had to practice skill k before step i. In the traditional formulation, the learning rates (γs) are bounded to be positive, so practicing KCs never decreases performance. To prevent the model from overfitting, the student parameters (αs) are typically L_2 regularized; i.e., they are given a normal prior with mean 0. Regularization decreases the model fit to the training data (i.e., the log-likelihood, AIC, and BIC) but improves the predictive accuracy on unseen data. Thus, when comparing regularized models to other approaches it should primarily be compared on measures that use held out data, such as cross validation.

2.1.2 Performance Factors Analysis

There are two key differences between this model and AFM. First, PFA does not have individual student parameters [16] (later variants have explored the addition of student parameters [6], but we base our current analysis on the original formulation). This usually substantially reduces the number of parameters of the model relative to AFM, particularly in datasets with a large number of unique students. Second, the model takes into account students' actual performance (not just opportunities completed) by splitting the learning rate for each skill into two learning rates: a rate for successful practice and a rate for unsuccessful practice. The standard equation based on these changes is the following:

$$z_i = \sum_{k \in KCs(i)} (\beta_k + \gamma_k success(i,k) + \rho_k failure(i,k))$$

where the βs represent the difficulty of the KCs, γs and ρs represent the learning rates for successful and unsuccessful practice on the KCs, $success(i,k)$ represents the number of successful applications of a skill k for the given student prior to step i, and $failure(i,k)$ represents the number of unsuccessful applications of a skill k for the given student prior to step i. Similar to AFM it is typical to restrict the learning rates (i.e., γs and ρs) to be positive [9]. One complication when comparing this model to other approaches using held out data (i.e., cross validation) is that the success and failure counts potentially contain additional information about the test data (i.e., performance on held out practice opportunities). Thus, we argue that comparing AFM to PFA using cross validation is usually not a fair comparison. Bearing this in mind, in the current analysis we were more interested in comparing AFM+Slip and PFA+Slip to their respective baseline models than to each other. To this end, we utilized cross validation as the primary measure of predictive accuracy for reasons previously discussed.

2.2 Bayesian Knowledge Tracing

There are many different models in the knowledge tracing family [10], but for this paper we focus on traditional 4-parameter BKT [7]. In contrast to the logistic approaches,

BKT utilizes a Hidden Markov Model to estimate latent parameters and predict student performance. This model has four parameters for each skill: the initial probability that the skill is known $p(L_0)$, the probability that the skill will transition from an unlearned to a learned state $p(T)$, the probability of an error given that the skill is learned $p(Slip)$, and the probability of a success when the skill is not learned $p(Guess)$. Unlike the logistic models, the estimation of these parameters can sometimes be difficult due to issues of identifiability [2] (e.g., there are many parameter values that yield the same likelihood) so these parameters are typically bounded to be within reasonable ranges; e.g., guess is typically bounded to be between 0 and 0.3 and slip is bounded to be between 0 and 0.1 [1]. Prior research has produced toolkits that can efficiently estimate these parameters using different approaches. For the comparisons in this paper we use the toolkit created by Yudelson et al. [23] and we use the gradient descent method.

One of the core differences between the logistic models and BKT is how they parameterize false negative student actions (i.e., slipping behavior). The logistic models do not have slip parameters and so they model student success as converging monotonically to 100% success (i.e., learning rates are bounded to be positive). In contrast, the BKT model explicitly models false negatives and allows monotonic convergence (under the typical assumption that the probability of forgetting is zero) to lower success rates. The slip parameters in BKT also serve the purpose of accounting for noise in student performance, and it is unclear whether these parameters account for true slipping behavior (i.e., non-zero base rate error) or just general noise in the student actions. Since the logistic models can already handle noise in the data, it remains to be seen what would happen if slip parameters were added to these models. That is the focus of this papers' investigation.

3. BOUNDED LOGISTIC REGRESSION

There is no trivial approach to incorporating explicit slip parameters into the logistic models; e.g., the prediction probability cannot be bounded by an additional linear term to the logistic function. In order to add these parameters we modified the underlying logistic model to have the following form:

$$p_i = \frac{1}{1+e^{-s_i}} \times \frac{1}{1+e^{-z_i}}$$

where z_i is the same as that used in standard logistic regression and s_i is a linear function of the parameters that impose an upper bound on the success probability for the step i. For modeling a slip rate for each skill we use the following equation:

$$s_i = \tau + \sum_{k \in KCs(i)} \delta_k$$

where τ is the parameter corresponding to the average slip rate across all items and students and δ_k is the change in the average slip rate for each skill k. We also apply an L_2 regularization to the δ parameters to prevent overfitting. To fit the parameters we used the sequential quadratic programming package in Octave, which uses an approach similar to Newton-Raphson but properly accounts for parameter constraints (e.g., positive learning rates). For details on parameter estimation see Appendix A.

This formulation is a generalization of Item Response Theory approaches that model item slip (e.g., [21]). In particular, it supports slipping with multiple KC labels per an item by using a logistic function to map the sum of slip parameters to a value between 0 and 1. For items with a single KC label, the $\frac{1}{1+e^{-s_i}}$ term reduces to the slip probability for that KC. For multi-KC items, this term models slipping as the linear combination of the individual KC slipping parameters in logit space. This approach mirrors that taken by AFM and PFA for modeling KC difficulty and learning rates in situations with multiple KC labels. In these situations, prior work has shown that the logit approach gives a good approximations of both conjunctive and disjunctive KC behavior [4].

During early model exploration we used Markov Chain Monte Carlo methods to compare this formulation with a more complex formulation that had parameters for both guessing and slipping. Our preliminary results showed that AFM with slip parameters outperformed the guess-and-slip variation for the 'Geometry Area (1996-97)' [11] and the 'Self Explanation sch_a3329ee9 Winter 2008 (CL)' [3] datasets (accessed via DataShop [12]) in terms of deviance information criterion (a generalization of AIC for sampled data). Further analysis showed that there was little data to estimate the guessing portion of the logistic curve. This is because the average student error rate in these datasets starts off at less than 50% and only gets lower with practice. This is typical of many of the available tutor datasets, so for our Bounded Logistic Regression approach we decided it would be sufficient to model the slipping parameters.

4. EVALUATION

4.1 Method

We used bounded logistic regression to add slip parameters to AFM and PFA, thus creating two new student learning models: AFM + Slip and PFA + Slip. We were interested in how these approaches compared with their traditional counterparts and to Bayesian Knowledge Tracing, which parameterizes guess and slip. Furthermore, we were interested in how these different approaches compared across different datasets spanning distinct domains. To perform this evaluation we fit each of the five models to five datasets we downloaded from DataShop [12]: Geometry Area (1996-97) [11], Self Explanation sch_a3329ee9 Winter 2008 (CL)[3], IWT Self-Explanation Study 1 (Spring 2009) (tutors only) [19], IWT Self-Explanation Study 2 (Fall 2009) (tutors only) [20], and Digital Games for Improving Number Sense - Study 1 [15]. These datasets cover the domains of geometry, equation solving, writing, and number line estimation. We selected these datasets because they have undergone extensive KC model refinement, including both manually created models by domain experts and automatically-refined models using Learning Factors Analysis [5]. For all datasets we used the best fitting KC model, based on unstratified cross validation.

In addition to comparing the different statistical models' predictive accuracies, we were interested in understanding

Table 1: In all five datasets the slip models outperform their non-slip counterparts in terms of log-likelihood and cross validation. In four out of the five datasets, the PFA+Slip model outperforms the AFM+Slip model in terms of log-likelihood and cross validation performance. In this table "Par." represents the number of parameters in the model and the CV RMSE values are the averages of 10 runs of 10-fold un-stratified cross validation.

Dataset	Model	Par.	LL	AIC	BIC	CV RMSE
Geometry						
	AFM	95	-2399.7	4989.4	5610.5	0.396
	AFM+Slip	114	-2377.0	4982.0	5727.3	0.395
	PFA	54	-2374.9	4857.8	**5210.8**	0.389
	PFA+Slip	73	**-2298.3**	**4742.6**	5219.8	**0.383**
	BKT	72	-2460.8	5065.7	5536.5	0.396
Equation Solving						
	AFM	106	3011.6	6235.2	6953.9	0.390
	AFM+Slip	125	**-2992.5**	**6235.0**	7082.54	**0.388**
	PFA	48	-3205.2	6506.4	6831.8	0.400
	PFA+Slip	67	-3088.9	6311.8	**6766.0**	0.392
	BKT	72	-3202.7	6549.5	7037.7	0.426
Writing 1						
	AFM	169	-3214.6	6767.2	7916.1	0.406
	AFM+Slip	196	-3214.6	6821.2	8153.6	0.406
	PFA	72	-3212.0	6568.0	**7057.4**	0.401
	PFA+Slip	99	**-3158.0**	**6514.0**	7187.0	**0.398**
	BKT	104	-3480.2	7168.5	7875.6	0.419
Writing 2						
	AFM	129	-2976.4	6210.8	7096.6	0.375
	AFM+Slip	145	-2962.8	6215.6	7211.3	0.373
	PFA	45	-2994.7	6079.4	**6388.4**	0.373
	PFA+Slip	61	**2965.7**	**6053.4**	6472.2	**0.371**
	BKT	60	-3177.1	6474.2	6886.2	0.384
Number Line						
	AFM	93	-2352.7	4891.4	5484.0	0.433
	AFM+Slip	115	-2356.3	4942.6	5675.4	0.432
	PFA	62	-2337.5	**4799.0**	**5194.1**	0.430
	PFA+Slip	84	**-2318.9**	4805.8	5341.1	**0.428**
	BKT	84	-2548.7	5265.4	5800.7	0.451

and interpreting the situations in which slip parameters improve model fit. Prior to analysis we hypothesized that slipping parameters might have three potential effects on the model fit: (1) enabling the model to capture true student slipping behavior; i.e., KCs that have a non-zero base-rate error, (2) enabling the model to fit steeper initial learning rates while still making correct predictions at higher opportunity counts, and (3) enabling the model to compensate for an underspecified knowledge component model. We focused in on one dataset, Geometry Area (1996-97), to explore these possibilities. Within this dataset we conducted a residual analysis to explore possibilities (1) and (2). We then refined the geometry KC model for a specific KC that the slip model identified as having a high false negative rate (i.e., slip value) to explore possibility (3). For brevity we only show the results of AFM and AFM+Slip in these analyses, but similar trends hold for PFA and PFA+Slip.

4.2 Results

4.2.1 Model Fits for Five Datasets

We fit each of the five models to the five datasets. Table 1 shows the resulting model fit statistics and the number of parameters used in each model. AFM has 1 parameter per student and 2 parameters per skill, PFA has 3 parameters for each skill, and BKT has 4 parameters for each skill. The slip variations have an additional parameter for each skill, plus a parameter for the average slip rate. When using the PFA models in practice many of the KCs never had any unsuccessful practice (i.e., their failure count was always 0). In these situations we removed the parameters for the failure learning rates because they have no effect on the model behavior. Thus, in some situations, the number of parameters in each model might differ from the general trends. All of the cross validation results are the average of 10 runs of 10-fold unstratified cross validation, where the cross validated RMSE was computed using the predicted probability of a correct response (rather than discrete correct/incorrect predictions).

All of the slip models have better log-likelihood and cross validation performance than their respective baseline models (AFM and PFM). Furthermore, in four out of the five datasets, PFA+Slip has better cross validation performance than AFM+Slip, even though it does not have individual student parameters. Finally, all of the logistic models outperformed traditional four-parameter BKT. Based on prior work [16] we expected this last result, but we included BKT as a comparison model that supports slipping. In particular,

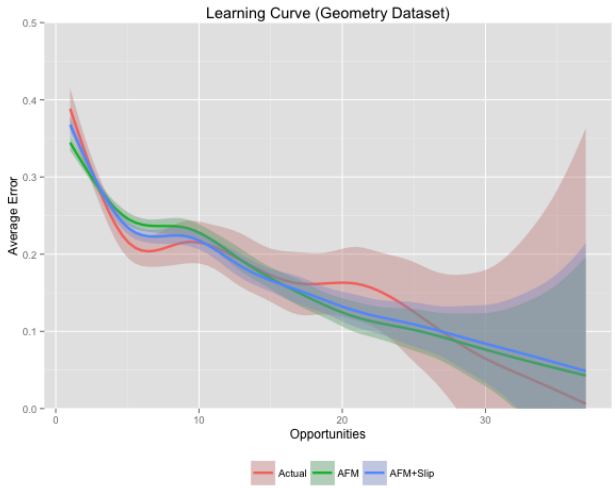

Figure 1: The AFM+Slip model better fits the steeper learning rate of the Geometry dataset than the AFM model, but both models fit the tail of the learning curve reasonably well and the actual student error appears to be converging to 0%. The shaded regions denote the 95% confidence intervals for the respective values.

Figure 2: The 95% confidence intervals (shaded regions) for the residuals of the AFM model do not include zero for lower opportunity counts, the model first overpredicts and then underpredicts success. In contrast the 95% confidence intervals for residuals of the AFM+Slip model always include zero indicating a better model fit.

Figure 3 shows an example of how the AFM+Slip model fits the data more like the BKT model than the AFM model for a KC with a high slip rate.

4.2.2 *Residual Analysis*

To investigate how the predictions of the slip models differ from that of the traditional models we analyzed the residuals for the AFM and AFM+Slip models on the Geometry dataset. Figure 1 shows the actual and predicted error rates for the two models on this dataset and Figure 2 shows the model residuals plotted by opportunity count. Investigating patterns in residual error across opportunity counts is a useful way of assessing systematic discrepancies between a given model's predicted learning curves and students' actual learning curves.

Although both models fit the data reasonably well, the slip model better models the steepness at the beginning of the learning curve. At low opportunity counts, AFM without slip typically predicts a substantially flatter learning curve compared to the actual data. The residual plot mirrors this finding; the 95% confidence interval for the AFM residuals does not include zero for earlier opportunities and the model flips from over-predicting success to under-predicting it. The AFM+Slip model, in contrast, better models the initial steepness of the learning curve. The 95% confidence interval for the AFM+Slip model residuals always includes zero. Finally, we see no evidence of actual slipping behavior in the tail of the learning curve: the 95% confidence intervals for residuals in both models include zero for higher opportunity counts. If true student slipping were occurring, we would expect the traditional AFM model to overpredict success in the tail, but we do not observe this.

4.2.3 *KC Refinement Based on False Negatives*

In order to explore the hypothesis that a high false negative, or slip, rate on a skill is indicative of a underspecified knowledge component model, we analyzed a KC on which the slip parameter was high and on which AFM and AFM+Slip differed substantially in their predictions. One KC, "geometry*compose-by-multiplication," fit this criteria. Figure 3 shows the learning curve with model predictions for this KC. AFM+Slip makes predictions that are nearly identical to BKT and seems to better fit the actual student learning curve. Upon further investigation, we found that many of the items labeled with this skill were on the same problems. Within these problems, we noticed that the later problem steps (items) might actually have been solved by applying the "arithmetic" skill to the result of an earlier application of the "compose-by-multiplication" skill. We generated a new knowledge component model to reflect these findings and re-fit the model using AFM. The predictions of this new model (AFM-New-KC) are also shown in Figure 3. For the AFM-New-KC plot, we plotted the observations with the opportunity counts from the original KC model (x-axis) but with predicted errors from the new KC model (y-axis). This was necessary for the purposes of comparison to the original KC model predictions. Once the knowledge component model was refined based on the insights provided by fitting AFM+Slip, standard AFM improved. Furthermore, based on this change the overall AFM model fit improved to be on par with AFM+Slip in terms of log-likelihood, AIC, and cross validation (LL = -2378.8, AIC = 4947.6, BIC = 5568.6, and CV RMSE = 0.395).

5. DISCUSSION

Our model fit results show that the slip models have better predictive accuracy (i.e., cross validation performance) and

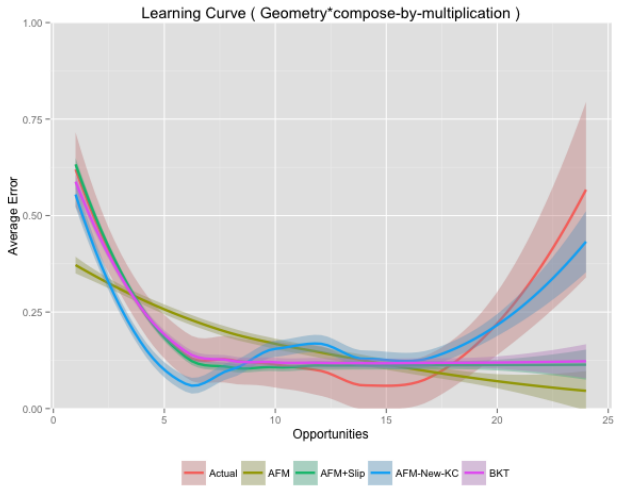

Figure 3: AFM+Slip looks much more like BKT for this KC and seems to model the data better (the overlapping purple and green lines). We took the high false negative rate (i.e., the sharp floor in the predicted error at approx. 11%) as an indicator that the KC model might benefit from refinement. Refitting the regular AFM model with a refined KC model (AFM-new-KC) shows a better fit to the actual data. Shaded regions denote the 95% confidence intervals for the respective values.

log-likelihood fits than their traditional counterparts across all five datasets. Furthermore, the AIC scores generally mirror this finding. These results suggest that the addition of the slip parameters to the logistic model formalism results in an improved model fit and an increased ability to predict behavior on unseen data.

In four of the five datasets, PFA + Slip best fit the data in terms of both log-likelihood and cross validation. In one sense, its superior cross-validation performance is surprising because the PFA models (as implemented here) have no student intercept parameters. However, they have an advantage for the cross validation statistic because they get success and failure counts that include information about performance on held out data, essentially giving these models an advantage over the other models. The better log-likelihood (and often AIC) scores are indicative of a better ability to fit the data that doesn't suffer from this discrepancy. However, PFA models have an advantage over AFM for this statistic because AFM uses regularization, which intentionally worsens the fit of the model to the data in an effort to improve predictive accuracy. To test if regularizing student parameters was causing PFA and PFA + Slip to outperform AFM and AFM + Slip we refit the AFM models to the Geometry dataset with student parameter regularization disabled and found that, at least for the Geometry dataset, the PFA models still outperforms the AFM models in terms of log-likelihood, AIC, BIC, and CV RMSE. These findings suggest that the PFA models better fits the data than the AFM models, but more work is needed to explore how best to compare these two approaches and to determine when one approach is preferable to another.

Lastly, the logistic models consistently outperform traditional four-parameter BKT. This is somewhat unsurprising because BKT does not have individual student parameters or separate learning rates for success and failure. However, we still included traditional BKT as a baseline model that is widely used and has explicit parameters for guess and slip. In particular, Figure 3 shows that for a KCs with high slip rate the AFM+Slip model performs more like BKT than AFM, suggesting that the new model is able to fit slipping and other false negative student behavior.

Given the finding that the slip models have better predictive accuracy and log-likelihood fits than their traditional counterparts, we investigated how the addition of slip parameters changed the model predictions. Residual analyses on the Geometry dataset showed that both AFM and AFM+Slip had similar fits to the data, but AFM+Slip better fit the initial steepness of the learning curve while maintaining a good fit in the tail. This intuition is confirmed in the residual by opportunity plot, which shows that the 95% confidence intervals for the residuals from AFM exclude zero at low opportunity counts, first overpredicting success and then underpredicting it. In contrast, the 95% confidence interval for the residuals from AFM+Slip include zero at these same low opportunity counts. This evidence supports the hypothesis that adding slip parameters enables the model to better accommodate steeper learning rates. In contrast, we find no evidence to support the hypothesis that adding slipping parameters enables the model to better fit non-zero base rate error; i.e., true student slipping. If this were the case, then we would expect AFM to overpredict success in the tail (i.e., for the residuals to be non-zero at higher opportunity counts), but we found no evidence that this occurred.

Finally, we demonstrated that high false negative, or slip, rates can serve as detectors of KCs that might benefit from further refinement. We identified a KC in the Geometry dataset that had a high slip rate and that differed from the traditional model: the "geometry*compose-by-multiplication" KC. We found that this KC could be further refined and showed that AFM with the refined KC model performed on par with AFM+Slip in terms of log-likelihood and cross validation. This suggests that adding slip parameters to a model can enable it to compensate for an underspecified KC model but, more importantly, can help identify these poorly specified KCs. The newly discovered KC model better fit the student data than the previous best model, which was the result of years of hand and automated KC model refinement.

6. CONCLUSIONS

Logistic models of learning, such as AFM and PFA, are popular approaches for modeling educational data. However, unlike models in the knowledge tracing family, they do not have the ability to explicitly model guessing and slipping rates on KCs. In this work we augmented traditional logistic regression to support slipping rates using an approach that we call Bounded Logistic Regression. We then used this approach to create two new student models: AFM + Slip and PFA + Slip. We then compared the performance of these new models in relation to their traditional counterparts. Furthermore, for AFM we explored how the addi-

tion of slip parameters changed the predictions made by the model. We explored three possibilities: (1) they might enable the model to capture true student slipping behavior (i.e, non-zero base-rate error), (2) they might enable the model to accommodate steeper learning rates while still effectively predicting performance at higher opportunity counts, and (3) they might enable the model to compensate for an underspecified knowledge component model.

To explore the first two possibilities, we conducted a residual analysis and found that the slip parameters appear to help the model fit steeper learning rates, rather than improving model fit in the tail. To explore the third possibility, we used a high false negative, or slip, rate as an indicator of where the given KC model might benefit from refinement. We found that after refining a KC model using this approach AFM performance (e.g., CV, LL, AIC) improved to be on par with AFM-Slip. This suggests that the slip parameters enable the model to compensate for underspecified KC models and that high slip values can be used to identify KCs that might benefit from further KC label refinement.

7. LIMITATIONS AND FUTURE WORK

One key limitation of the current work is that we did not explore issues of identifiability in the Bounded Logistic Regression model. In particular, we have not yet demonstrated that the log-likelihood for models using this formalism are convex. In the current formulation we only model slip parameters (not guess parameters), so we expect identifiability to be less of an issues. In line with this intuition we found that the current approach returned reasonable parameter values and consistently improved model fit across the five data sets we explored. However, we recognize that the model would benefit from a more rigorous analysis of the quality of estimated parameters and acknowledge that this would be an important direction for future work.

Finally, the current work focuses on comparing the slip models to their traditional counterparts, but future work might explore how different models (e.g., AFM+Slip, PFA+Slip, and BKT) compare to one another. In the current work we purposefully avoided making conclusions about how these models compare because there is some ambiguity in how different approaches are evaluated. For example, Yudelson's Bayesian Knowledge Tracing toolkit [23] performs incremental prediction during cross validation (i.e., predicting student performance on a step and then "showing" the model the actual performance before moving on to the next step). While this approach aligns well with the actual use of the BKT model it gives an unfair advantage when comparing it to cross validated AFM, which gets no information about test data when making predictions. A similar complication exists for PFA, which gets information about the performance of unseen steps from the success and failure counts. A more equivalent comparison would be to perform an incremental prediction using AFM and PFA, but this was beyond the scope of the current paper and represents an open area for future work.

8. ACKNOWLEDGEMENTS

We thank Erik Harpstead, Michael Yudelson, and Rony Patel for their thoughts and comments when developing this work. This work was supported in part by the Department of Education (#R305B090023 and #R305B110003) and by the National Science Foundation (#SBE-0836012). Finally, we thank Carnegie Learning and all other data providers for making their data available on DataShop.

9. REFERENCES

[1] 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. Aimeur, R. Nkambou, and L. S, editors, *ITS '08*, pages 406–415, 2008.

[2] J. E. Beck and K.-M. Chang. Identifiability: A Fundamental Problem of Student Modeling. In C. Conati, K. McCoy, and G. Paliouras, editors, *UM '07*, pages 137–146, 2007.

[3] J. Booth and S. Ritter. Self Explanation sch_a3329ee9 Winter 2008 (CL). `pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=293`.

[4] H. Cen. *Generalized Learning Factors Analysis: Improving cognitive Models with Machine Learning*. PhD thesis, Carnegie Mellon University, 2009.

[5] H. Cen, K. R. Koedinger, and B. Junker. Learning Factors Analysis - A General Method for Cognitive Model Evaluation and Improvement. In M. Ikeda, K. Ashlay, and T.-W. Chan, editors, *ITS '06*, pages 164–175, 2006.

[6] M. Chi, K. R. Koedinger, G. Gordon, P. Jordan, and K. Vanlehn. Instructional Factors Analysis: A Cognitive Model For Multiple Instructional Interventions. In M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero, and J. Stamper, editors, *EDM '11*, pages 61–70, 2011.

[7] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *UMUAI*, 4(4):253–278, 1995.

[8] K. L. Draney, P. Pirolli, and M. Wilson. A measurement model for a complex cognitive skill. In P. N, S. Chipman, and R. Brennan, editors, *Cognitively diagnostic assessment*, pages 103–125. Lawrence Erlbaum Associates Inc., 1995.

[9] Y. Gong, J. E. Beck, and N. T. Heffernan. Comparing Knowledge Tracing and Performance Factor Analysis by Using Multiple Model Fitting Procedures. In V. Aleven, J. Kay, and J. Mostow, editors, *ITS '10*, pages 35–44, 2010.

[10] G.-B. Jose, H. Yun, and P. Brusilovsky. General Features in Knowledge Tracing: Applications to Multiple Subskills,Temporal Item Response Theory, and Expert Knowledge. In J. Stamper, Z. Pardos, M. Mavrikis, and B. M. McLaren, editors, *EDM '14*, pages 84–91, 2014.

[11] K. Koedinger. Geometry Area 1996-1997. `pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=76`.

[12] K. R. Koedinger, R. S. J. d. Baker, K. Cunningham, A. Skogsholm, B. Leber, and J. Stamper. A Data Repository for the EDM community: The PSLC DataShop. In C. Romero, S. Ventura, M. Pechenizkiy, and R. S. J. d. Baker, editors, *Handbook of Educational Data Mining*. CRC Press, 2010.

[13] K. R. Koedinger, J. Stamper, E. McLaughlin, and

T. Nixon. Using Data-Driven Discovery of Better Student Models to Improve Student Learning. In H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik, editors, *AIED '13*, pages 421–430, 2013.

[14] R. Liu, K. R. Koedinger, and E. A. McLaughlin. Interpreting Model Discovery and Testing Generalization to a New Dataset. In *EDM '14*, pages 107–113, 2014.

[15] D. Lomas. Digital Games for Improving Number Sense - Study 1. `pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=445`.

[16] P. I. Pavlik, H. Cen, and K. R. Koedinger. Performance Factors Analysis –A New Alternative to Knowledge Tracing. In V. Dimitrova and R. Mizoguchi, editors, *AIED '09*, pages 531–538, 2009.

[17] J. C. Stamper and K. R. Koedinger. Human-machine student model discovery and improvement using Data. In J. Kay, S. Bull, and G. Biswas, editors, *AIED '11*, pages 353–360, 2011.

[18] K. Vanlehn. The Behavior of Tutoring Systems. *IJAIED*, 16(3):227–265, 2006.

[19] R. Wylie. IWT Self-Explanation Study 1 (Spring 2009) (tutors only). `pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=313`.

[20] R. Wylie. IWT Self-Explanation Study 2 (Spring 2009) (tutors only). `pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=372`.

[21] Y. C. Yen, R. G. Ho, W. W. Laio, L. J. Chen, and C. C. Kuo. An Empirical Evaluation of the Slip Correction in the Four Parameter Logistic Models With Computerized Adaptive Testing. *APM*, 36(2):75–87, 2012.

[22] M. V. Yudelson and K. R. Koedinger. Estimating the benefits of student model improvements on a substantive scale. In S. K. D'Mello, R. A. Calvo, and A. Olney, editors, *EDM '13*, 2013.

[23] M. V. Yudelson, K. R. Koedinger, and G. J. Gordon. Individualized Bayesian Knowledge Tracing Models. In H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik, editors, *AIED '13*, pages 171–180, 2013.

APPENDIX

A. PARAMETER ESTIMATION

Similar to standard logistic regression we assume the data follows a binomial distribution. Thus, the likelihood and log-likelihood are as follows:

$$\begin{aligned} Likelihood(data) &= \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{(1-y_i)} \\ ll(data) &= \sum_{i=1}^{n} y_i log(p_i) + (1-y_i) log(1-p_i) \end{aligned}$$

where y_i is 0 or 1 depending on if the given step i was correct. As mentioned earlier, p_i is defined as:

$$p_i = \frac{1}{1+e^{-s_i}} \times \frac{1}{1+e^{-z_i}}$$

where s_i is the linear combination of the slip parameters and z_i is the linear combination of the student and item parameters.

To estimate the parameters values for bounded logistic regression, we maximize the conditional maximum likelihood of the data using sequential quadratic programming (specifically the sqp package in Octave). This approach reduces to applying the Newton-Raphson method, but properly accounts for situations when the parameter values are constrained, such as the positive bound for the learning rates in AFM and PFA. To apply this method, we needed to compute the gradient and hessian for the likelihood of the data given the model.

To compute the gradient we took the derivative with respect to the student and item parameters (w's) and slip parameters (sp's). For the student and item parameters the gradient is the following:

$$\frac{dll}{dw_a} = \sum_{i=1}^{n} \frac{x_{ia}}{1+e^{z_i}} \frac{(y_i - p_i)}{(1-p_i)}$$

where x_{ia} is the value of the student or item feature that is being weighted by parameter w_a for step i.

Similarly, for the slip parameters the gradient is the following:

$$\frac{dll}{dsp_a} = \sum_{i=1}^{n} \frac{q_{ia}}{1+e^{s_i}} \frac{(y_i - p_i)}{(1-p_i)}$$

where q_{ia} is the value of the slip feature (in AFM and PFA these are the 0 or 1 entries from the Q-matrix) that is being weighted by parameter sp_a for step i.

Given these gradients we have a hessian matrix with values for the interactions of the ws with each other, the ws with the sps, and the sps with each other. These values are defined as the following:

$$\begin{aligned} \frac{d^2ll}{dw_a dw_b} &= \sum_{i=1}^{n} \frac{x_{ia}x_{ib}}{(1+e^{z_i})^2(1-p_i)^2}[p_i(y_i-1) \\ &\quad + e^{z_i}(p_i - y_i)(1-p_i)] \\ \frac{d^2ll}{dsp_a dsp_b} &= \sum_{i=1}^{n} \frac{q_{ia}q_{ib}}{(1+e^{s_i})^2(1-p_i)^2}[p_i(y_i-1) \\ &\quad + e^{s_i}(p_i - y_i)(1-p_i)] \\ \frac{d^2ll}{dw_a dsp_b} &= \sum_{i=1}^{n} \frac{x_{ia}}{1+e^{z_i}} \left[\frac{(p_i-1)+(y_i-p_i)}{(1-p_i)^2}\right] \end{aligned}$$

Finally, in our formulation we applied an L_2 regularization to all of the parameter values (i.e., a normal prior with mean 0), where the λ parameter of the regularization could be set individually for each model parameter. For the AFM models we set λ to 1 for the student parameters. For all of the slip models we λ to 1 for the KC slip parameters (i.e., δs). For all other parameters we turned regularization off ($\lambda = 0$).