Using Dirichlet priors to improve model parameter plausibility

Dovan Rai, Yue Gong, and Joseph E. Beck
{dovan, ygong, josephbeck}@wpi.edu
Computer Science Department, Worcester Polytechnic Institute

Abstract. Student modeling is a widely used approach to make inference about a student's attributes like knowledge, learning, etc. If we wish to use these models to analyze and better understand student learning there are two problems. First, a model's ability to predict student performance is at best weakly related to the accuracy of any one of its parameters. Second, a commonly used student modeling technique, knowledge tracing, suffers from having multiple sets of parameters providing equally good model fits. Furthermore, common methods for estimating parameters, including conjugate gradient descent and expectation maximization, suffer from finding local maxima that are heavily dependent on their starting values. We propose a technique that estimates Dirichlet priors directly from the data, and show that using those priors produces model parameters that provide a more plausible picture of student knowledge. Although plausibility is difficult to quantify, we employed external measures to show the parameter estimates were indeed improved, even if our model did not predict student behavior any more accurately.

1 Introduction

The goal of student modeling is to take observations of a student's performance and use those to estimate the student's knowledge, goals, preferences, and other latent characteristics. In general, student models are used to adapt instruction and are evaluated by how well they predict the student's behavior. However, with the advent of educational data mining, it is becoming more common to use model parameters to answer scientific questions (e.g. [1]). Unfortunately, just because a model is an accurate predictor of student behavior, that does not mean we are justified in interpreting the model's parameters to make claims about student learning. This paper focuses on examining this issue, investigates techniques for finding more plausible model parameters, and proposes methods for evaluating parameters for plausibility. First, we provide some background into our student modeling framework, knowledge tracing, and the statistical approach we use to bias model fitting, Dirichlet priors.

1.1 Knowledge tracing model

Knowledge tracing [2], shown in Figure 1, is an approach for taking student observations and using those to estimate the student's level of knowledge. There are two parameters *slip* and *guess*, which mediate student knowledge and student performance. These two parameters are called the performance parameters in the model. An assumption of the model is that even if a student knows a skill, there is a chance he might still respond incorrectly to a question that utilizes that skill. This probability is the *slip* parameter.

There are a variety of reasons for an incorrect response, for example, the student could have made a simple typo (e.g. typed '12' instead of '21' for "7 x 3").

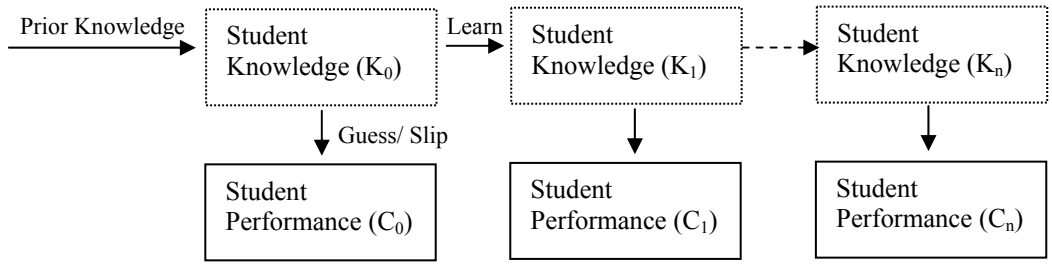

Figure 1. Knowledge tracing model

Prior Knowledge = Pr (K_0=True)
Guess = Pr (C_n=True | K_n=False)
Slip = Pr (C_n=False | K_n =True)
Learning rate = Pr (K_n =True | K_{n-1}=False)
Conversely, a student who does not know the skill might still be able to generate a correct response. This probability is referred to as the *guess* parameter. A guess could occur either through blind chance (e.g. in a 4- choice multiple choice test there is a ¼ chance of getting a question right even if one does not understand it), or the student being able to utilize a weaker version of the correct rule that only applies in certain circumstances.

In addition to the two performance parameters, there are two learning parameters. The first is *prior knowledge* (K_0), the likelihood the student knows the skill when he first uses the tutor. The second learning parameter is *learning*, the probability a student will acquire a skill as a result of an opportunity to practice it. Every skill to be tracked has these four parameters, *slip*, *guess, K_0,* and *learning*, associated with it.

1.2 The problem

One issue is how to estimate the model parameters. One approach is to use the expectation maximization (EM) algorithm to find parameters that maximize the data likelihood (i.e. the probability of observing our student performance data). However, in EM, we have to start with some initial value of the parameter, and final parameter estimations are sensitive to those initial values. Furthermore, one flaw of a knowledge tracing model is that it has multiple global maxima. That is to say, there can be more than one set of learning/performance parameters that fit the data equally well.

Consider the three sets of hypothetical knowledge tracing parameters shown in Table 1, the *knowledge* model reflects a set of model parameters where students rarely guess. The *guess* model assumes that 30% of correct responses are due to randomness. This limit of 30% is the maximum allowed in the knowledge tracing code used by the Cognitive Tutors [2]. The third model has parameters similar to data from Project Listen's Reading Tutor [3].

By using the four parameters and the knowledge tracing equations, we can compute the theoretic learning and performance curves for each model. Specifically, we initialize P(know) to be K_0. After each practice opportunity, we use formula I to update P(know) as the new likelihood of the student knows the skill after the previous practice. Also we compute P(correct), the probability of the student will respond correctly in the current practice opportunity, by using the knowledge tracing formula to combine the estimated knowledge with the *slip* and *guess* parameters shown in formula II.

$$P(know) = P(know) + (1 - P(know)) * learning \quad \text{(I)}$$

$$P(correct) = P(know) * (1\text{-}slip) + (1 - P(know)) * guess. \quad \text{(II)}$$

For example, the knowledge model's prior knowledge (K_0) is 0.56. At the second practice opportunity the knowledge model would have a P(know) of 0.56 + (1 – 0.56) * 0.1 = 0.604. Furthermore, the likelihood for the student making a correct response would be 0.604 *(1-0.05) + (1-0.604) * 0.00 = 0.574. As seen in Figure 2, the three models have identical student performance (in the left graph), but their estimates of student knowledge (right graph in Figure 2) are very different.

Table 1. Parameters for three hypothetical knowledge tracing models

Parameter	Model		
	Knowledge	***Guess***	***Reading Tutor***
Prior Knowledge	0.56	0.36	0.01
Learning	0.1	0.1	0.1
Guess	0.00	0.3	0.53
Slip	0.05	0.05	0.05

Given the same set of performance data, we have presented three knowledge tracing models that fit the data equally well, i.e. all three sets of estimated parameters have equally good predictive power. Unfortunately, for drawing conclusions about student learning, they make very different claims. Statistically there is no justification for preferring one model over the others, since all three of the sets of parameters fit the observed data equally well. This problem of multiple (differing) sets of parameter values that make identical predictions is known as identifiability [4].

1.3 Proposed solution: Dirichlet priors

Dirichlet prior is an approach used to initialize conditional probability tables when training a Dynamic Bayesian network. Dirichlet distributions are specified by a pair of numbers (α,β). Figure 3 shows an example (the dashed line) of the Dirichlet distribution for (9,6). If this sample distribution were of K_0, it would suggest that few skills have particularly high or low knowledge, and we expect students to have a moderate probability of mastering most skills. Conceptually, one can think of the conditional probability table of the graphical model being as seeded with 9 instances of the student knowing the skill initially and 6 instances of him not. If there is substantial training data, the parameter estimation procedure is willing to move away from an estimate of 0.6. If

there are few observations, the priors dominate the process. The distribution has a mean of $\alpha/(\alpha+\beta)$. Note that if both α and β increase, as in the solid curve in Figure 3, the mean of the distribution is unchanged (since both numerator and denominator are multiplied by 3) but the variance is reduced. Thus, Dirichlets enable researchers to not only specify the most likely value for a parameter but the confidence in the estimate.

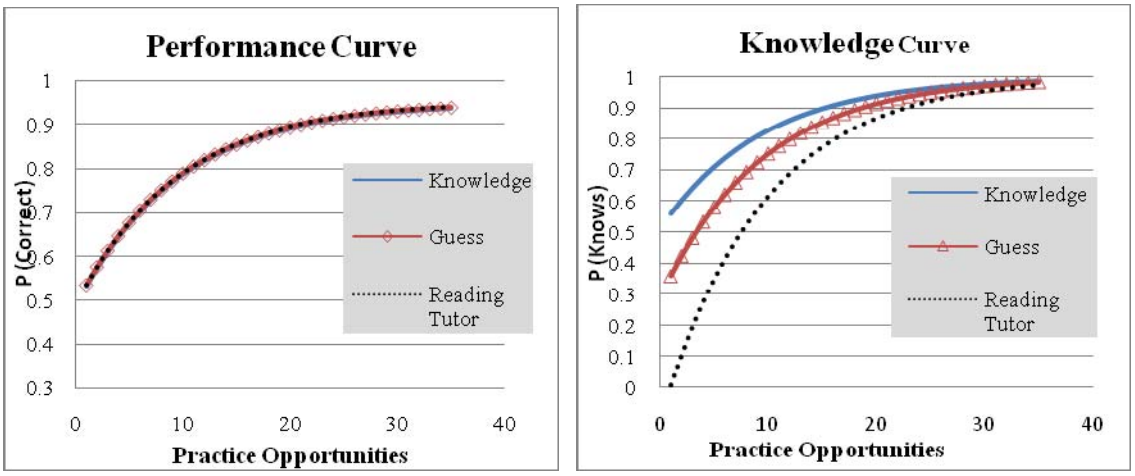

Figure 2 performance & learning curve

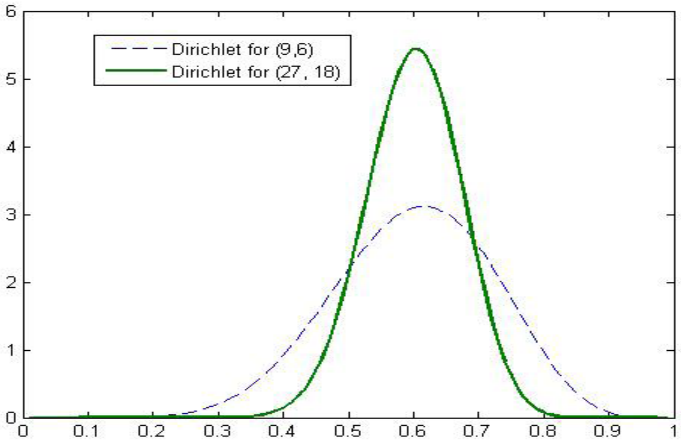

Figure 3. Sample Dirichlet Distributions demonstrating decreasing variance

Dirichlets provide bias towards the mean of the distribution. Since we estimated a set of parameters for each skill, for models with few training data, the parameter estimates can get wacky, since sparse data provide few constraints on the parameters. Hence, those parameters are sometimes estimated as extreme values. In this situation, we prefer to have parameters which are more similar to other, better-estimated, skills. With Dirichlet priors, the observations for each case are weighted against prior α, β values, i.e. models with few data are more influenced by the priors towards the mean. Therefore, we expect those estimates will be become more reasonable.

It is important to note that researchers can use Dirichlets to set confidence on priors. If the variance is less, we are surer about the priors, whereas if the variance is high, we are

less sure about the priors. Each of the four parameters will not only have different mean values, but different degrees of certainty. Suppose, in a group of students if they start with similar incoming knowledge but have variable learning. Then Dirichlet prior will set higher confidence in students' prior knowledge (e.g.: $\alpha, \beta = 20, 34$) but lower confidence in students' learning (e.g.: $\alpha, \beta = 1, 4$). As a result, prior knowledge parameter estimation will be more biased towards prior or distribution's mean whereas learning will have more tendency to move away from prior value.

2 Methodology

There are several sources of setting Dirichlet prior values. One approach is using knowledge of the domain [e.g. 4]. If someone knows how quickly students tend to master a skill or the likelihood of knowing a skill, that knowledge can be used to set the priors. One complaint is that such an approach is not necessarily replicable as for different domains and different subjects, different experts may give different answers.

2.1 An automatic approach for selecting priors

To compare estimations from fixed and Dirichlet prior models, we trained two KT models initialized with fixed and with Dirichlet priors. We used the following approach:

1. Initialize EM with fixed priors from our rough estimates of the domain. Then use EM to estimate the model parameters for each skill in the domain
2. For all four parameters (guess, slip, K_0, learning)
   - Compute the mean (μ) and variance (σ^2) of the parameter estimates
   - Weight the mean and variance by the number of cases (n) of each skill. Specifically, for each parameter P of skill i,
     - $weight_i = \sqrt{n_i}$
     - $\mu' = \sum Pi * weight_i / \sum weight$
     - $\sigma^{2\prime} = \sum weight_i * (P_i - \mu_p)^2 / \sum weight$
   - Select α and β to generate a Dirichlet with the same mean and variance as the estimates Specifically, solve for α and β such that:
     - $\alpha = (\mu'^2 / \sigma^{2\prime}) * (1- \mu') - \mu'$
     - $\beta = \alpha *((1/ \mu')-1)$
3. We now have one Dirichlet distribution described by (α, β) for each of the four parameters
4. Reestimate two kinds of knowledge tracing models: a fixed prior model with initial value of μ' and Dirichlet prior model using the (α, β) pairs.

We calculated the mean and variance of the data. Based on those two values, we calculated α, β parameters (using the equations in step #2). However, simply calculating the mean gives all data points equal weight. This can be problematic, since as we mentioned earlier, skills with few cases are susceptible to error: going to extreme values such as getting 0 as student's learning parameter. Therefore, we weight each estimate by the square root of the number of cases used to generate the estimate, since $\sqrt{N}$ is how the standard error decreases.

2.2 Iterating the algorithm

Rather than just stopping after step #4, it is possible to loop back to step #2. We were interested to see how the parameter estimates change by iterating the algorithm with new prior values. We ran a number of iterations on both fixed and Dirichlet prior.

Table 2. Results of iterating automatic process approach for K_0 and slip parameters

		Prior Knowledge (K_0)			Slip		
		Iteration 1	**Iteration 2**	**Iteration 3**	**Iteration 1**	**Iteration 2**	**Iteration 3**
Fixed Prior	Mean, Variance	0.473, 0.025	0.471, 0.025	0.468, 0.025	0.205, 0.006	0.205, 0.006	0.203, 0.005
Dirichlet Prior	Mean, Variance	0.478, 0.019	0.477, 0.017	0.476, 0.016	0.207, 0.003	0.208, 0.002	0.208, 0.002
	α, β	5.76, 6.3	6.66,7.3	6.86,7.55	11.21, 42.82	14.5, 55.2	16.63, 63.31

As shown from Table 2 , the parameters do not change much across iterations, although the variance decreases. The amount of bias towards the mean is proportional to how large α and β are, which is inversely related to the population variance. That is, if the population has a high variance then there is a small bias. Conversely, if a parameter value is already tightly clustered, there will be a strong bias towards the mean. Therefore, at each iteration estimates will move towards the mean, and the values of α, β will increase. We discuss this problem further in the future work section.

3 Validating the models

For this study, we used data from ASSISTment, a web-based math tutoring system. The data are from 199 twelve- through fourteen- year old 8th grade students in urban school districts of the Northeast United States. They were from two classes, each of which only lasted one month. These data consisted of 92,319 log records of ASSISTment during January 2009 to February 2009. Performance records of each student were logged across time slices for 106 skills (e.g. area of polygons, Venn diagram, division, etc). We split our data into training set and test set with the proportion of 2:1.

Using our approach, we ran the fixed prior model and the Dirichlet prior model for a number of successive iterations and compared their predictive accuracy and parameter plausibility.

3.1 Predictive Accuracy

Predictive accuracy is the measure of how well the instantiated model fits the data. We used two metrics to examine the model performance on test set: AUC (Area Under Curve) and Summed Squared Error (SSE).

As seen in We also computed the SSE = Σ (observed performance – P (correct)) $^{2.}$ We found the first iteration of the Dirichlet prior model shows a slightly better, but not meaningfully better SSE than the first iteration of fixed prior model: 8008 vs. 8016. With more iteration, SSE marginally decreases for fixed prior whereas it increases in Dirichlet.

Table 3, the AUC values don't show any difference in performance of fixed prior model and Dirichlet prior model. The values remain unchanged even for successive iterations. We also computed the SSE = Σ (observed performance – P (correct)) ^{2}. We found the first iteration of the Dirichlet prior model shows a slightly better, but not meaningfully better SSE than the first iteration of fixed prior model: 8008 vs. 8016. With more iteration, SSE marginally decreases for fixed prior whereas it increases in Dirichlet.

Table 3. Comparison of SSE and AUC

	AUC		SSE	
	Fixed	Dirichlet	Fixed	Dirichlet
iteration #1	0.66	0.66	8016	8008
iteration #2	0.66	0.66	8015	8010
iteration #3	0.66	0.66	8015	8012

These results show that predictive accuracy is not meaningfully better with Dirichlet priors and the accuracy does not seem to be improving with successive iterations.

3.2 Parameter plausibility

Predictive accuracy is a desired property, but EDM is also about interpreting models to make scientific claims. Therefore, we prefer models with more plausible parameters when we want to use those for scientific study. Unfortunately, quantifying parameter plausibility is difficult since there are no well-established means of evaluation. In our study, we explored two metrics for this analysis.

For our first metric, we inspected the number of practice opportunities required to master each skill in the domain. We assume that skills in the curriculum are designed to neither be so easy to be mastered in three or fewer opportunities nor too hard as to take more than 50 opportunities. We define mastery as the same way as was done for the mastery learning criterion in the LISP tutor [5]: students have mastered a skill if their estimated knowledge is greater than 0.95. Based on students' prior knowledge and learning parameters and knowledge tracing equations described before, we calculated the number of practice opportunities required until the predicted value of P(know) exceeds 0.95. Then, we compared the number of skills with unreliable extreme values in both cases (fewer than 3 and more than 50).

As seen in Table 4, fixed priors result in more extreme cases than Dirichlet priors. This result implies that Dirichlet prior model estimates more plausible parameters. . With more iteration, the extreme cases remain constant with fixed prior whereas the number slightly decreases with Dirichlet priors. The skills that are found implausible by Dirichlet are a subset of those found by fixed priors. Hence, Dirichlet is fixing the implausibility of fixed priors and is not introducing new problems of its own.

Along with this method, we had tried to make an evaluation based on the correlation between estimated the model's K_0 and the skill difficulty. We consulted two domain experts to rate skill difficulties. But their ratings were not consistent (correlation <0.4) with each other and so we abandoned this approach.

Table 4. Comparison of extreme number of practice until mastery

	# of skills with # of practices >=50		# of skills with # of practices <=3	
	Fixed	Dirichlet	Fixed	Dirichlet
iteration #1	29	17	2	0
iteration #2	29	16	2	0
iteration #3	29	15	2	0

Next, we tried to model students instead of skills since we it is easier to objectively rate characteristics of students rather than skills. We trained KT model per student by observing his responses in all questions across skills. The model then estimated a set of parameters (prior knowledge, guess, slip and learning) for each *student* (rather than for each skill) which represents his aggregate performance across all skills.

The students in our study had taken a 33-item algebra pre-test just before using the tutor. Taking the pre-test as external measure of incoming knowledge, we calculated the correlation between students' prior knowledge (K_0) as estimated by KT models and their pretest scores. In Table 5, we can see that the Dirichlet prior model produces slightly stronger, but not reliably so, correlations than the fixed prior. Neither method improves with more iterations.

Table 5 Comparison of correlation between prior knowledge and pretest

	Fixed prior model	Dirichlet prior model
iteration #1	0.76	0.80
iteration #2	0.73	0.81
iteration #3	0.73	0.81
iteration #4	0.72	0.81

4 Contributions

This paper extends prior work in automatically generating Dirichlet priors [6] in several ways. First, this study has been scaled up both in terms of more students and more skills. Prior work found a small positive, but non-reliable, gain in predictive accuracy from using Dirichlets. This paper provides evidence that the improvement was illusory. We have also improved the estimation of the α and β parameters by weighting the parameter estimates by the number of observations we have for the skill. In this way we reduce the effect of skills that only have few estimates of skewing the mean and increasing the variance.

This paper also presents a new method for evaluating student models for parameter plausibility. Although prior work [4,6] in this area proposed and used a variety of metrics, there is still a need for additional methods. Our new method was to essentially swap the knowledge tracing problem, and estimate a set of model parameters for the students rather than the skills. We then correlated the K_0 parameter for each student with his pretest score. There are many ways of estimating how much knowledge students have, and many research efforts will have approaches for doing this. Therefore, we expect this technique to have broad applicability.

Finally, we are able to extend the result that EM produces more predictive models than Conjugate Gradient Descent [8], the approach used to estimate parameter in the CMU cognitive tutors. We are now able to say that EM + Dirichlet priors is better than EM alone. Using Dirichlets we are not able to predict student behavior any better, but the parameters are generally more plausible than with fixed priors.

5 Future work and Conclusions

There are several interesting open issues regarding the estimation of Dirichlet priors. First, our method of weighting the parameter estimation process by $\sqrt{N}$, although inspired by the relative standard error of each skill's parameters, could use more theoretic grounding. Second, neither the current nor past attempt [6] at automatically extracting α and β values from the data have shown improvements in model predictive performance. However, the single attempt at human-generated Dirichlet priors [4] did show such gains. Perhaps people have useful knowledge to bring to bear on this task? Some means of incorporating human experts, and perhaps combining their insight with computer-suggested priors could be a positive step.

The notion of iterating our process of fitting the data, estimating α and β, and refitting the data seems like it should work, and was in fact inspired by the expectation maximization recipe. That it did not work was something of a disappointment, but we think we understand why: at each iteration the population variance shrinks, increasing α and β, which further shrinks the population variance on the next iteration. We need some mechanism of preventing α and β from increasing arbitrarily high, or some better metric that suggests what a "good" value of those parameters would look like.

Finally, the assumption that we can estimate the shape of the Dirichlet distribution from which the parameters were drawn is certainly more relaxed than the standard assumption that we can correctly estimate the parameter values for each skill, however it is still somewhat naïve. For example, consider the initial knowledge of a skill. It is plausible that some skills will not have been covered in class by the students: those skills could be described by a Dirichlet with a low average. Other skills, that were covered in class, could be well described by a Dirichlet with a high average. There is no single distribution that would handle both cases. Therefore, it might be productive to consider mixtures of Dirichlets.

This paper has shown that automatically generated Dirichlets are a method for generating more plausible parameters. We found that, with Dirichlets, fewer skills were estimated to

require too many or too few practice opportunities to master. We have also introduced a new evaluation technique for evaluating parameter plausibility, and expect this technique to be widely applicable.

Acknowledgements

We would like to thank all of the people associated with creating the ASSISTment system listed at www.ASSISTment.org. We would also like to acknowledge funding from the National Science Foundation, the Fulbright Program for funding the first author and the US Department of Education and the Office of Naval Research for funding the second and third authors. All of the opinions expressed in this paper are those solely of the authors and not those of our funding organizations.

References

[1] Joseph E. Beck, Kai-min Chang, Jack Mostow, Albert T. Corbett, *Does Help Help? Introducing the Bayesian Evaluation and Assessment Methodology. Intelligent Tutoring Systems* 2008: 383-394.

[2] Corbett, A. and J. Anderson, *Knowledge tracing: Modeling the acquisition of procedural knowledge*. User modeling and user-adapted interaction, 1995. 4: p. 253-278.

[3] Mostow, J. and G. Aist, Evaluating tutors that listen: An overview of Project LISTEN, in Smart Machines in Education, K. Forbus and P. Feltovich, Editors. 2001, MIT/AAAI Press: Menlo Park, CA. p. 169-234.

[4] Beck, J. E., & Chang, K.-m. (2007, June 25-29). *Identifiability: A Fundamental Problem of Student Modeling.* Proceedings of the 11th International Conference on User Modeling (UM 2007), Corfu, Greece.

[5] Corbett, A.T. Cognitive computer tutors: *Solving the two-sigma problem. in International Conference on User Modeling.* 2001. p. 137-147.

[6] Beck, J. E. (2007, July 9). *Difficulties in inferring student knowledge from observations (and why you should care).* Proceedings of the AIED2007 Workshop on Educational Data Mining, Marina del Rey, CA, 21-30.

[7] Kimberly Ferguson, Ivon Arroyo, Sridhar Mahadevan, Beverly Woolf and Andy Barto: *Improving Intelligent Tutoring Systems: Using Expectation Maximization to Learn Student Skill Level.* Intelligent Tutoring Systems: Volume 4053/2006

[8] Kai-min Chang, Joseph Beck, Jack Mostow and Albert Corbett *: A Bayes Net Toolkit for Student Modeling in Intelligent Tutoring Systems* : Intelligent Tutoring Systems: Intelligent Tutoring SystemsVolume 4053/2006