Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Challenges to Applying Performance Factor Analysis to Existing Learning Systems

Cristina MAIER[a*], Ryan S. BAKER[b] & Steve STALZER[c]
[a]*McGraw Hill Education, USA*
[b]*University of Pennsylvania, USA*
[c]*McGraw Hill Education, USA*
*cristina.maier@mheducation.com

Abstract: The last decade has seen a wide variety of new algorithms proposed for knowledge tracing in adaptive learning. However, with the exception of Bayesian Knowledge Tracing (BKT), most of these algorithms' properties for real-world usage have not been thoroughly studied. In this paper, we consider real-world practical concerns around the scaled use of Performance Factors Analysis (PFA), another widely researched algorithm: developing models that work for skills that were rare or unavailable in initial data sets, skills encountered by many students but only in one or two items, content tagged with both common and rare skills, and whether skills are compensatory or conjunctive. We map these limitations to the problem of model degeneracy, not yet explored in detail for PFA. We discuss the scope and properties of each challenge, and then discuss potential solutions.

Keywords: Knowledge Tracing, Performance Factor Analysis, Adaptive Learning

1. Introduction

In recent years, a proliferation of ways to model student knowledge have emerged for adaptive learning environments. Early research focused on variants of a single algorithm, Bayesian Knowledge Tracing (BKT; Corbett & Anderson, 1995). More recently, dozens of algorithms – many of them variants on logistic regression or neural networks – have become popular. However, most commercial adaptive learning systems still continue to either use BKT (Sales & Pane, 2019) or simpler heuristics such as three-in-a-row correct. Part of the reason for this choice is the speed and ease of implementing BKT – the code to incorporate BKT into a system can fit into a single page, and there are several public packages for fitting BKT, which can run through large data sets in under a day. However, another reason is the relatively deep understanding of BKT's properties in real-world usage (e.g. Baker et al., 2008; van de Sande, 2013; Pelánek et al., 2016; Sales & Pane, 2019). For example, researchers have studied the phenomenon of degenerate parameters in BKT -- parameter values where the algorithm's behavior does not match the intuitive interpretation of what the parameters should mean (Baker et al., 2008; van de Sande, 2013). The ELO algorithm, used in many adaptive learning systems, has also been formally studied (e.g. Pelánek et al., 2016; Yudelson, 2019).

The real-world properties of neural network variants is also being studied. However, despite excellent predictive performance (e.g. Piech et al., 2015; Yeung, 2018), researchers found that the first variant (Deep Knowledge Tracing, DKT; Piech et al., 2015), had limitations for real-world usage, such as predicted performance going down after correct answers and large fluctuations in prediction from action to action (Yeung, 2018). Concerns have also been raised about the interpretability and usefulness of this model's predictions for teachers and other end users (Zhang et al., 2017). Research continues into the properties of these types of algorithms (Yeung, 2018; Lee et al., 2021).

In this paper, we consider another algorithm, Performance Factors Analysis (PFA; Pavlik et al., 2009). PFA performs competitively in predicting performance (Gong et al., 2010; Scruggs et al., 2020), though more poorly than DKT variants for large data sets (Gervet et al., 2020). However, its properties for real-world use have been less thoroughly studied.

2. Knowledge Tracing in Adaptive Learning

Knowledge Tracing is commonly posed as the problem of trying to use student performance on a set of items to predict performance on future items (e.g. Piech et al., 2015). However, some have argued instead that it should be defined as the problem of trying to use student performance on a set of items to predict latent knowledge that leads to better performance outside the learning system (Corbett & Anderson, 1995; Scruggs et al., 2020). In this section, we will define PFA and mention a few other algorithms that will be discussed within the paper.

PFA, in its typical formulation, predicts performance on a given item, at a given time, using a student's past number of successes, multiplied by a weight γ fit to each of the item's skills, a student's past number of failures, multiplied by a weight ρ fit to each of the item's skills, a weight β which, depending on the variant of PFA, is applied across all contexts, across all items linked to the current skill (the most common approach and what we will use here), across all items of the same "item-type", or for individual items**.** These features are inputted into a logistic function to obtain a prediction, p(m), between 0 and 1 for success for a given student on a given (future) item.

$$\mathbf{m(i, j \in KCs, s, f)} = \sum_{\mathbf{j \in KCs}} \left(\boldsymbol{\beta_j} + \boldsymbol{\gamma_j s_{i,j}} + \boldsymbol{\rho_j f_{i,j}}\right); \qquad \mathbf{p(m)} = \frac{\mathbf{1}}{\mathbf{1 + e^{-m}}}$$

Within this equation, KCs are the Knowledge Components (i.e. skills, benchmarks) linked to the item. Parameter $\boldsymbol{i}$ represents the current learner. Parameters $\boldsymbol{\beta_j}, \boldsymbol{\gamma_j}$ and $\boldsymbol{\rho_j}$ represent the learned parameters for skill $\boldsymbol{j}$, $\boldsymbol{s_{i,j}}$ represents the number of successful practices made by learner $\boldsymbol{i}$ on skill $\boldsymbol{j}$ thus far, and $\boldsymbol{f_{i,j}}$ is the number of failed practices made by learner $\boldsymbol{i}$ on skill $\boldsymbol{j}$ thus far. PFA is therefore able to model learning in common real-world situations where multiple skills are associated with an item.

Despite PFA's widespread use in research and its competitive success on predicting future student performance in some studies (e.g. Gong & Beck, 2011; Gong, Beck, & Heffernan, 2010; Scruggs et al., 2020), there has been relatively little study of what factors impact PFA's behavior in real-world learning settings. This paper's goal is to study the factors that have emerged for other algorithms, to better understand the use of PFA in real-world learning. In section 3, we discuss our data set and PFA's baseline performance. In section 4, we focus on four challenges: insufficient data for a student and skill, degenerate parameters, rare benchmarks, and compensatory vs. conjunctive skill relationships. We conclude the paper with a brief discussion of other potential challenges for PFA.

3. Dataset and the Baseline PFA Results

The experiments provided within this paper use data from Reveal Math Course 1, a digital core math product for 6th grade. The platform provides instructional material, such as lessons and units, and assessments with question items. The items in this courseware are tagged with metadata that represent a set of skills. For these experiments we used a type of skill called a benchmark that corresponds to mathematics standards. Items can be tagged with multiple benchmarks. Throughout the paper the terms skill and benchmark will be used interchangeably. The data we used for the experiments come from three U.S. school districts that use NGA Center/CCSSO Common Core Standards.

We obtained data from 3073 students in 46 schools, who used the system between August 2019 to January 2021. We extracted information about scored items, and normalized the scores to values between 0 and 1. Though the data had partial credit, only 1.35% of responses received a partial score. Therefore, we only used binary scores, assigning any normalized partial score < 1 a final score of 0.

For analysis of model quality, we split our dataset into training and testing sets. We randomly selected 20% of the students (614 students; 52,516 data points; 64 benchmarks) to the testing dataset, leaving the remaining 80% of the students (2,459 students; 208,553 data points; 62 benchmarks) in the training dataset. The two benchmarks missing from the testing dataset were very rare; each had only one data point. The dataset contains a mix of multi-skill and single-skill items. 52.77% of the training set involved multi-skill items, and 53.81% of the testing set involved multi-skill items.

We analyzed our models' performance across the testing dataset. For some experiments, we

analyzed the results for four subcategories of datapoints linked to common and/or rare benchmarks. We consider a benchmark common if the training dataset has at least 200 students that have at least 3 data points (practices, items) linked to that benchmark (cf. Slater & Baker, 2018), and rare if it does not satisfy this condition. 27 benchmarks were common and 37 were rare. The four testing data subcategories are: datapoints linked to at least one common benchmark (50,251, out of which 56.03% are multi-skill), datapoints linked to only common benchmarks (49,371, out of which 52.25% are multi-skill), datapoints linked to at least one rare benchmark (3,145, out of which 31.22% are multi-skill), and datapoints linked only to rare benchmarks (2,265, out of which 4.5% are multi-skill).

Baseline PFA Results

We trained a "baseline" model using the original PFA formula on all 64 benchmarks. In the baseline model, β parameters were bounded from –3 to 3, and γ and ρ parameters were bounded from –1 to 1. Bounds were used to keep parameters within a reasonable range and speed the training process (training with no bounds had minimal impact on AUC and RMSE). Overall, AUC is in the 0.78-0.80 range (Table 1), except (unexpectedly) for rare benchmarks, where AUC reaches 0.83. These values are somewhat higher than typically seen for PFA in other papers (Gervet et al., 2020; Gong et al., 2010; Scruggs et al., 2020). RMSE values, in the 0.42-0.44 range, are more in line with values seen in past papers.

Table 1. Baseline unmodified PFA model - Validation Results (Testing Dataset)

Category Data Points	AUC (bounds)	AUC (no bounds)	RMSE (bounds)	RMSE (no bounds)
All	0.7818	0.7810	0.4245	0.4207
At least one common benchmark	0.7809	0.7801	0.4244	0.4203
Only common benchmarks	0.7797	0.7787	0.4235	0.4193
At least one rare benchmark	0.7954	0.7930	0.4408	0.4420
Only rare benchmarks	0.8320	0.8302	0.4268	0.4297

4. Challenges, Potential Solutions and Experimental Results

Insufficient Number of Practices

One challenge with using PFA appears when students in the data set have insufficient practice with a skill. Effective estimation for PFA, like all algorithms, depends on having sufficient data, and sufficient data set size for knowledge tracing depends on the number of data points per student and skill as well as the overall data set size, and insufficient data set size can lead both to poorer prediction and extreme parameter values (Slater & Baker, 2018). Training a PFA model with data that does not contain enough students with a sufficient number of practices for a skill might yield unsatisfactory performance, and create degenerate or extreme learned parameters. The question of how much practice is needed per student before knowledge tracing estimation can yield reliable parameters has been studied for BKT (Slater & Baker, 2018), but not yet for PFA, although one comparison across data sets suggests that PFA may need less data than DKT (Gervet et al., 2020).

Using the baseline PFA approach, we ran several experiments with different thresholds for the maximum number of practices, to simulate the impact of having less data per student. Specifically, we filtered the training data points to retain only the data points that are linked to at least one skill for which the student had at most 'threshold' practices. We trained multiple models using the filtered training data points for different thresholds (2 to 7), and then we validated each model against the entire testing dataset. We found major improvement from 2 to 3 practices (AUC increased from ~0.753 to ~0.772), with continued improvement up to practice 6 (AUC ~0.78). For the trained models with fewer than 5 practices (i.e. threshold < 5), we observed more degenerate learned parameters. We will talk more about this issue in the next subsection.

We also investigated how insufficient number of practices with a skill impacted prediction, by training a PFA model using the entire training dataset, and validating across different subsets of the testing dataset. We filtered the testing dataset such that we retained only the datapoints which were linked to at least one skill for which this datapoint represented for the student a practice number less or equal to a given threshold (2-20). The higher the threshold, the more practices allowed. With more practices, the validation results improved – most substantially from the second to fifth practices (AUC increased from ~0.725 to ~0.755), followed by a slower increase up to around practice 12, after which the performance flattens out (AUC ~0.78).

Degenerate Parameters

Another challenge is degenerate learned parameters. Model degeneracy has been discussed relatively thoroughly for BKT (Baker, Corbett, & Aleven, 2008; Pardos & Heffernan, 2010; van de Sande, 2013). A model is degenerate "where parameter values violate the model's conceptual meaning (such as a student being more likely to get a correct answer if he/she does not know a skill than if he/she does)." (Baker, Corbett, & Aleven, 2008, p. 406). There have been reports of degenerate behavior for DKT as well (Yeung, 2018). We believe there are three cases where a PFA model is degenerate. First, when $\gamma < 0$ -- this indicates that if a student obtains a correct answer, they are likely to do more poorly in the future. Second, when $\gamma < \rho$ -- this indicates that a student's future performance will be better if they get the item wrong now than if they get the item right now. Third, when γ and ρ are both zero, no matter what the student does, the predictions will not change. It is worth noting that a fourth case when $\rho > 0$ -- is not degenerate, due to the multiple functions the parameters perform in PFA. In this case, the rate of learning the skill may outweigh the evidence of lack of student knowledge that an incorrect answer provides. So long as $\gamma > \rho$, a positive ρ is conceptually acceptable.

Several causes could produce degenerate PFA parameters. For example, if β is used at the skill or item type level (as in Pavlik et al., 2009), then a learning system that moves students from easier items to harder items, within the same nominal skill, will produce $\gamma < 0$. In cases where items are tagged with multiple skills, collinearity between skills could produce degenerate parameters.

In practice, within our full test data set, a baseline PFA model resulted in 6 skills (9% of skills) that had type 1 degeneracy, and 7 skills (11% of all skills) had type 3 degeneracy. No skill showed type 2 degeneracy. All degenerate learned parameters were linked to rare benchmarks**.** However, type 2 degeneracy was observed when training models that only used data from 4 or fewer practices.

This problem can be addressed as in (Corbett & Anderson, 1995), by constraining parameter values to be non-degenerate. When we do this with the Reveal Math Course 1 data and a baseline PFA model, constraining γ to be within [0,1] and ρ within [-1,0], we get slightly better test set performance than for the baseline PFA with looser bounds, with an AUC ROC of 0.7834 and an RMSE of 0.4219. However, even with these constraints, just as many (7) skills still had type 3 degeneracy.

It is noteworthy that baseline PFA had excellent AUC ROC on rare skills despite having degenerate parameters for some of those skills. This may be due to the fact that these skills generally had at most 1 or 2 practices for a student, which means that γ and ρ values either did not count at all or played a minimal role. Overall, then, it seems like degeneracy is a challenge for PFA, but primarily so when data is limited. The *merged-rare model* proposed next offers a potential solution to this problem.

Rare vs Common Skills

Another challenge is the handling of rare skills, which may occur in several situations, including when instructors frequently choose not to assign some concepts, or some items are tagged with skills that are not directly taught within the courseware, such as prerequisite skills. It is also possible that authors may under/over-tag items. Thus, there may be many cases where tags differ in granularity and frequency.

A learning system needs to be able to handle both common and rare skills. Depending on how rare a skill is, there might not be enough data points for precise parameter estimation when training for those skills, or there might not be enough data points per student. Simulation-based research has been conducted to distill guidelines for minimum data set size for BKT (Slater & Baker, 2018), but not yet for PFA, which would be more complicated due to multi-skill items. There is also the case when we have to deal with an item tagged with a skill with no past data. For both these cases, we need to decide

which skills to train our model on, and then need a way to handle performance prediction for items tagged with skills the model was not trained for. Hence, we propose to adjust the original PFA formula to differentiate between rare and common skills. While different β, γ and ρ parameters are fit for common skills, a common set of parameters $(\beta_d, \gamma_d, \rho_d)$ are used for rare skills:

$$\mathbf{m(i, j \in KCs, s, f)} = \sum_{\mathbf{j \in common\ KCs}} \left(\boldsymbol{\beta_j} + \boldsymbol{\gamma_j s_{i,j}} + \boldsymbol{\rho_j f_{i,j}}\right) + \sum_{\mathbf{j \in rare\ KCs}} \left(\boldsymbol{\beta_d} + \boldsymbol{\gamma_d s_{i,j}} + \boldsymbol{\rho_d f_{i,j}}\right)$$

As discussed above, our data set has 27 common skills, and 37 rare skills. We trained our *merged-rare* model with 28 sets of parameters: 27 for the common benchmarks, and one default set that applied for every rare benchmark. No degenerate parameters of any type were observed. Validating the model against our testing dataset resulted in an AUC of 0.7836 and an RMSE of 0.4223 for the entire testing dataset, slightly better than the results of our baseline PFA model.

An alternative approach to handling rare benchmarks is to use the average parameters of the common skills for the rare skills. This model obtains an AUC of 0.7834 and an RMSE of 0.4223 for the entire test set, essentially the same results – though again, no degeneracy. This suggests that using a set of parameters that is calculated as an average across the trained learned parameters might be a reasonable option to handle entirely new or very rare skills.

Overall, then, combining rare skills in PFA can reduce degeneracy when data is limited.

Compensatory vs Conjunctive Skills in PFA

Another question is whether skills are conjunctive or compensatory. To succeed when solving an item with conjunctive skills, a student needs to know all the skills tagged to that item. For an item that has compensatory skills, knowing only a subset of the tagged skills could be enough.

PFA's original formula is inherently compensatory. PFA can be adjusted to be conjunctive by multiplying together the skills rather than adding them (i.e. replacing the $\sum$ with $\prod$ in the m function). A third, in-between assumption is possible – that performance depends on each skill evenly. This *even-skill* model, uses averaging instead of $\sum$ or $\prod$ in the function. When fitting our data with the compensatory model we see an AUC of 0.7818 and a RMSE of 0.4245 on the testing dataset, whereas when using a conjunctive model the performance drops significantly, resulting in an AUC of 0.6725 and a RMSE of 0.4666. The performance results for the even-skill approach was slightly better than the compensatory approach, with an AUC of 0.7849 and an RMSE of 0.4171. This result contradicts past findings for BKT, where conjunctive models performed best in other data (i.e. Pardos et al, 2008).

5. Discussion and Conclusions

In this paper, we have discussed some of the challenges in real-world use of PFA. We focus on four potential areas of challenge: insufficient number of practices, degenerate parameters, rare benchmarks, and compensatory vs. conjunctive skill relationships. Overall, we find that degenerate parameters are an issue for PFA, particularly for benchmarks where there is limited initial data per student, and that degeneracy can be hidden by overall high performance (much like BKT – cf. Baker et al., 2008).

Other challenges may emerge in using PFA in real-world settings, depending on the design of the learning system. For one thing, the tagging of skills to items – the knowledge structure -- may have flaws, and there may be benefits to refitting the knowledge structure. Of course, the cost to arbitrarily refitting item to skill mappings is interpretability, and other challenges listed above – such as model degeneracy -- may also become more prominent with more complex mappings. These challenges may magnify if the skills have pre-requisite structure, but the order of the content does not respect the pre-requisite structure. In that case, a skill may be encountered both before and after a student has mastered its prerequisites, leading to sudden spikes in learning which PFA may be unable to capture, in contrast to BKT (which assumes sudden shifts in knowledge) or DKT-family algorithms (which can capture arbitrarily complex relationships between items). Another potential opportunity is seen in cases (like this data set) where tags are at the level of standards or benchmarks rather than fine-grained skills – in

these cases, making the tagging more fine-grained may increase the precision of estimation.

Even if items are tagged well to skills, items may have different levels of difficulty due to issues such as cognitive load or calculation time. PFA has a natural way of handling this issue – shifting the β parameter to the item level, or adding multiple β parameters (one at the skill level, one at the item level). Differences in item guess and slip may also play a role. Unlike BKT, which incorporates guess and slip probabilities for each skill, the original version of PFA does not take that into account (see MacLellan, Liu, & Koedinger, 2015 for a variant that does). While there is a 50% chance for a student to answer a true/false question correctly by guessing, the chances are smaller for a multiple choice question, and far smaller for a fill in the blank question, suggesting that the item type could be useful to model.

Overall, this article demonstrates that there are several considerations that must be taken into account in using PFA in the real-world, as previously demonstrated for other learning algorithms. However, for our specific data set, PFA's limitations seem very feasible to address with only minor adjustments. Given PFA's high interpretability, predictable behavior, and competitive performance in terms of AUC ROC/RMSE, PFA is a very reasonable choice for real-world student modeling.

References

Baker, R.S.J.d., Corbett, A.T., Aleven, V. (2008) More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (pp. 406-415). Montreal, Canada: Springer.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, *4*(4), 253-278.

Gervet, T., Koedinger, K., Schneider, J., & Mitchell, T. (2020). When is Deep Learning the Best Approach to Knowledge Tracing?. *Journal of Educational Data Mining, 12* (3), 31-54.

Gong, Y., & Beck, J. E. (2011). Looking beyond transfer models: finding other sources of power for student models. *Proc. of the Int.l Conf. on User Modeling, Adaptation, and Personalization* (pp. 135-146).

Gong, Y., Beck, J. E., & Heffernan, N. T. (2010). Comparing knowledge tracing and performance factor analysis by using multiple model fitting procedures. *Proc. of the International conference on intelligent tutoring systems* (pp. 35-44). Pittsburgh, PA, USA: Springer.

Lee, S., Choi, Y., Park, J., Kim, B., Shin, J. (2021) Consistency and monotonicity regularization for neural knowledge tracing, Preprint. Retrieved 3/27/2021 from https://openreview.net/forum?id=4P35MfnBQIY.

MacLellan, C. J., Liu, R., & Koedinger, K. R. (2015). Accounting for Slipping and Other False Negatives in Logistic Models of Student Learning. *Proc. of the Int'l Conf on Educational Data Mining*. Madrid, Spain.

Pardos, Z. A., Beck, J., Ruiz, C., Heffernan, N. T. (2008) The Composition Effect: Conjunctive or Compensatory? An Analysis of Multi-Skill Math Questions in ITS. *Proc. of the Int Conf. on Educational Data Mining.*

Pardos, Z., & Heffernan, N. (2010). Navigating the parameter space of Bayesian Knowledge Tracing models: Visualizations of the convergence of the Expectation Maximization algorithm. In *Proceedings of the International Conference on Educational Data Mining* (pp. 161-170). Pittsburgh, PA, USA.

Pavlik, P.I., Cen, H., Koedinger, K.R. (2009) Performance Factors Analysis – A New Alternative to Knowledge Tracing. *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 531-538).

Pelánek, R., Rihák, J., & Papoušek, J. (2016). Impact of data collection on interpretation and evaluation of student models. *Proc. of the 6th int. conf. on learning analytics & knowledge* (pp. 40-47).

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. *In Proc. of the 28th Int. Conf. on Neural Information Processing Systems* (pp. 505-513). Montreal, Canada: ACM Press.

Sales, A. C., & Pane, J. F. (2019). The role of mastery learning in an intelligent tutoring system: Principal stratification on a latent variable. *The Annals of Applied Statistics*, *13* (1), 420-443.

Scruggs, R., Baker, R.S., McLaren, B.M. (2020) Extending Deep Knowledge Tracing: Inferring Interpretable Knowledge and Predicting Post System Performance. *Proc. of the 28th Int. Conf. on Computers in Ed*.

Slater, S., Baker, R.S. (2018) Degree of Error in Bayesian Knowledge Tracing Estimates From Differences in Sample Sizes. *Behaviormetrika*, *45* (2), 475-493.

Van de Sande, B. (2013). Properties of the Bayesian Knowledge Tracing Model. *Journal of Educational Data Mining*, *5* (2), 1-10.

Yeung, C. K. (2018). Improving deep knowledge tracing with prediction-consistent regularization. Doctoral dissertation, Hong Kong University of Science and Technology.

Yudelson, M. (2019). Elo, I Love You Won't You Tell Me Your K. *European Conference on Technology Enhanced Learning* (pp. 213-223). Delft, The Netherlands: Springer.

Zhang, J., Shi, X., King, I., & Yeung, D. Y. (2017). Dynamic key-value memory networks for knowledge tracing. *Proc. of the 26th int. conf. on World Wide Web* (pp. 765-774). Perth, Australia: ACM Press.