A Comparison of Model Selection Metrics in DataShop

John C. Stamper
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA
john@stamper.org

Kenneth R. Koedinger
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA
koedinger@cmu.edu

Elizabeth A. McLaughlin
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA
mimim@cs.cmu.edu

ABSTRACT

Variations of cognitive models drive many instructional decisions that intelligent tutoring systems currently make. A better knowledge component model will yield better instruction, but how do we identify better cognitive models? One answer has been to create a latent variable version of a cognitive model or a so-called knowledge component (KC) model, then compare different models by how well they predict student performance data. In this research we analyze 1,943 proposed KC models that exist in DataShop (http://pslcdatashop.org) and compare and contrast the different metrics used to measure the quality of predictive fit to the data. All these metrics are designed to avoid over-fitting to the data, including AIC, BIC, and cross validation. We find that AIC is the metric most consistent with all the others and corresponds better with cross validation results than BIC.

Keywords

Model Selection, AIC, BIC, Cross-validation, KC Modeling.

1. INTRODUCTION

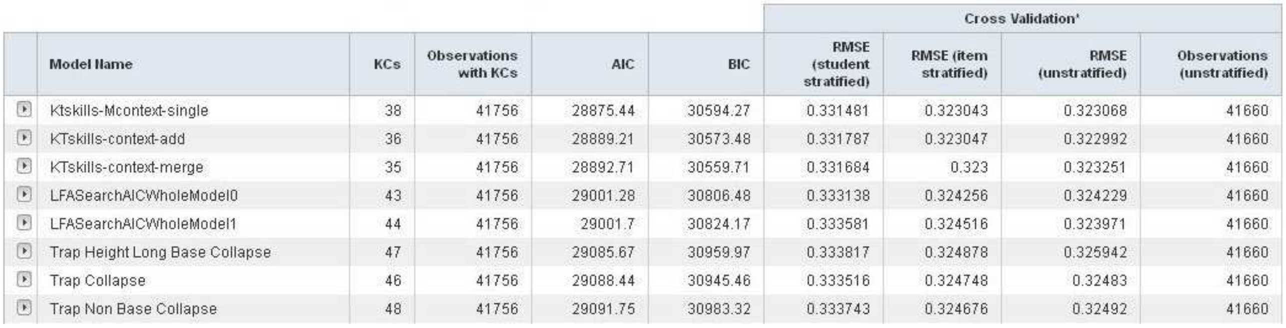
An important area of Educational Data Mining (EDM) is the building and improvement of models of student knowledge. Creating good models are important in the design of adaptive feedback, assessment of student knowledge, and predicting student outcomes [9]. A correct model of student knowledge is consistent with student behavior, such that, it predicts task difficulty and transfer between prior opportunities to practice and learn (via positive or negative feedback and next-step hints) and future opportunities to demonstrate learning (by correct performance). These models are evaluated by how well they predict the student performance on actual student data. To prevent selecting models that overfit the data (and would thus not work well in new contexts), prediction fit is measured using a number of techniques including cross validation, the Akaike information criterion (AIC), and the Bayesian information criterion (BIC). Cross validation is the gold standard for evaluating prediction fit and avoiding over-fitting, but it can take substantial time to run making it undesirable for searching for new models. AIC and BIC are metrics that can be calculated quickly, which makes them desirable when comparing a large number of proposed models, but how adequate are they and which one is better at anticipating cross validation results? This research explores comparisons of AIC, BIC, and various cross validations that are available in DataShop.

DataShop is the world's largest open data repository of transactional educational data collected from a wide variety of online learning environments [10]. The data is fine-grained, with student actions recorded roughly every 10 seconds on average, and it is longitudinal, spanning semester or yearlong courses. As of May 2013, over 400 datasets are stored including over 100 million student actions, which equates to over 250,000 student hours of data. Most student actions are "coded" meaning they are not only graded as correct or incorrect, but are categorized in terms of the hypothesized skills or knowledge components (KCs) needed to perform that action. DataShop stores a widespread selection of educational data from assorted technologies, domains and researchers. STEM subjects are well represented as are languages such as Chinese, English and French. There are also accessible datasets in miscellaneous content areas like reading, psychology, logic and handwriting. The acquisition of student log-data comes from a multitude of sources including intelligent tutors, online-courses and internet games and simulations. The collection methodologies include random controlled experiments, longitudinal studies, and anonymous on-line game playing.

Given the accessibility of data and diversity of applications stored in DataShop's repository, we were interested in exploring the metrics commonly used for model selection and prediction (i.e., AIC, BIC and Cross-validation). In particular, we examined correlations and rank order correlations between the various metrics to determine if one metric stands apart as a better predictor. Next, we examined best model selections for AIC and BIC and how they compared to cross validation best model selection.

A KC is defined as a piece of knowledge that can be applied to solve a specific task. Practically, KCs can be considered generalizations of skills or concepts that form the basis of a cognitive model of student knowledge. A typical step in a problem that a student will solve will include one or more KCs that describe the knowledge that the student is applying. A mapping of KCs to problem steps in a set of instruction forms a KC Model. Multiple mappings can be fit to the same set of student instruction based on the granularity of the KCs that make up each model. Figure 1 shows a screen shot from DataShop listing the KC Models and their evaluation metrics for a dataset called "Cog Model Discovery Experiment Spring 2010."

A KC model can be used to track individual student knowledge or predict student responses based on a statistical representation of the KC Model. In DataShop, the model used to evaluate student learning is called the Additive Factors Model (AFM) [3; 11]. AFM is an extension of item response theory that incorporates a growth or learning term [cf.,6]. AFM is shown in Figure 2. The discrete portion of the student model is represented by q_{jk}, the so-called "Q matrix" [13], which maps hypothesized difficulty or learning factors (the knowledge components or skills) to steps in problems. These factors are hypothesized causes for difficulty (β_k) or for learning improvement as students practice (γ_k). AFM gives a probability that a student i will get a problem step j correct based on the student's baseline proficiency (θ_i), the baseline difficulty (β_k) of the required KCs (q_{jk}), and the improvement (γ_k) in those KCs as the student gets practice opportunities (T_{ik}).



						Cross Validation'			
	Model Name	KCs	Observations with KCs	AIC	BIC	RMSE (student stratified)	RMSE (item stratified)	RMSE (unstratified)	Observations (unstratified)
▸	Ktskills-Mcontext-single	38	41756	28875.44	30594.27	0.331481	0.323043	0.323068	41660
▸	KTskills-context-add	36	41756	28889.21	30573.48	0.331787	0.323047	0.322992	41660
▸	KTskills-context-merge	35	41756	28892.71	30559.71	0.331684	0.323	0.323251	41660
▸	LFASearchAICWholeModel0	43	41756	29001.28	30806.48	0.333138	0.324256	0.324229	41660
▸	LFASearchAICWholeModel1	44	41756	29001.7	30824.17	0.333581	0.324516	0.323971	41660
▸	Trap Height Long Base Collapse	47	41756	29085.67	30959.97	0.333817	0.324878	0.325942	41660
▸	Trap Collapse	46	41756	29088.44	30945.46	0.333516	0.324748	0.32483	41660
▸	Trap Non Base Collapse	48	41756	29091.75	30983.32	0.333743	0.324676	0.32492	41660

Figure 1. Screenshot of the KC Models page in DataShop (http://pslcdatashop.org) for the dataset Cog Model Discovery Experiment Spring 2010. Here we can see named models with a different number of KCs in each. Note that all the models with the same number of observations with KCs (41,756 for example) are comparable with each other. DataShop also allows for the user to select the metric on which to sort the models. In this case, the models are sorted by AIC where a lower value is better.

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \theta_i + \sum_{k=1}^{K} q_{jk}\beta_k + \sum_{k=1}^{K} q_{jk}\gamma_k T_{ik}$$

Figure 2. In the Additive Factors Model (AFM), the probability student i gets step j correct (p_{ij}) is proportional to the overall proficiency of student i (θ_i) plus for each factor or knowledge component k present for this step j (indicated by q_{jk}), add the base difficulty of that factor (β_k) and the product of the number of practice opportunities this student (i) has had to learn this factor (T_{ik}) and the amount gained for each opportunity (γ_k).

The AFM model can be used to evaluate and predict learning, which can be visualized in DataShop with the use of learning curves. Fig 3 shows a learning curve for a KC called "Find circle circumference." The red line represents the actual student error rate from data collected in the dataset over each opportunity students have to apply the KC. The blue line represents the predicted model derived through AFM. DataShop allows for visual inspection of the KCs and their predicted fit with AFM, which can be used to help identify potential improvements in the KC model when the data and AFM curves do not match [12].

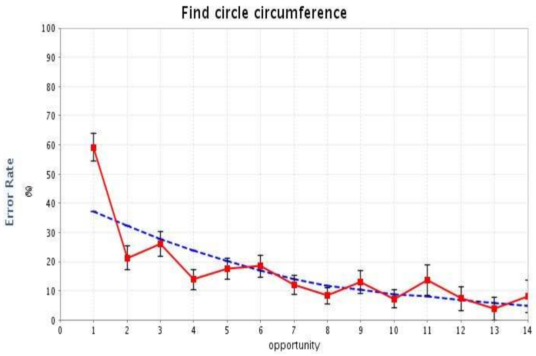


Figure 3. Learning Curve visualization from DataShop showing the KC "Find circle circumference". The y-axis is the error rate and the x-axis is each opportunity students have to apply the KC. The red is actual data and blue line is the predicted value.

When a potential improvement is found and a new model is proposed, it can be imported into DataShop and the system will automatically evaluate the new model against five metrics: AIC, BIC, Student Stratified Cross Validation (SSCV), Item Stratified Cross Validation (ISCV), and Non-Stratified Cross Validation (NSCV). Using these metrics the researcher can make a judgment as to whether the potential model leads to a better fit on the data.

When time is not an issue, cross validation is considered by most to be the best way to score models, but there is no consensus on how the cross validation should be done for educational transaction data. DataShop provides three cross validation measures that are each 10 fold and provides a value for the root mean squared error (RMSE). One measure stratifies the data by student, another by item, and the third is not stratified. While cross validation is considered the best method to score models, it is more time consuming and computationally expensive for large datasets than AIC or BIC. For this reason, when comparing many models we use AIC or BIC to score the models.

One active research area where many models are compared and evaluated against each other is the automated search for improved models. Using the AFM model and datasets in DataShop we have previously implemented an automated search algorithm, Learning Factors Analysis (LFA), for discovering better cognitive models [8]. This algorithm has been successfully applied to DataShop datasets and succeeded in improving existing models. Figure 1 includes two models that were automatically generated and are named "LFASearch…" The LFA search algorithm uses existing KC Models to complete a directed search, which results in labeled models that are easily interpretable by researchers.

AIC and BIC are measures for the goodness of predictive fit of a statistical model. They extend the log-likelihood measure of fit by penalizing less parsimonious models. Unlike the RMSE calculation from cross validation, the values of AIC and BIC have no meaning for an individual model, and are only useful when comparing alternative models built on the same dataset. Within DataShop, this means that models must have the same number of observations tagged with KCs to be comparable. DataShop also has a Model Values page under the Learning Curve tool that has more detailed information on the model metrics (AIC, BIC, and the cross validations), and the inputs used to calculate them (log likelihood, number of parameters, etc.). AIC is a metric for model comparison that trades off the complexity of the estimated model against how well the model fits the data [1]. In this way, it penalizes the model based on its complexity (the number of parameters). The equation for calculating AIC is ***AIC*= 2k – 2 ln(L)**, where k is the number of parameters and L is the likelihood. The equation for BIC is ***BIC*=k ln(n) – 2 ln(L)**, where n is the number of observations, k is the number of parameters, and L is the likelihood. BIC is similar to AIC, but BIC penalizes free parameters more strongly than AIC as can be seen by the formulas and noting that the coefficient of the number of parameters (k) is much larger for BIC (ln(n) for n observations)

Table 1. AIC and BIC correlations against each other and Cross-validation

KC set name	# students	# models	# obs	AIC-BIC correl	AIC-correlation			BIC-correlation		
					SSCV	ISCV	NSCV	SSCV	ISCV	NSCV
Assistments Math 2008-2009 Symb-DFA (302 Students)	302	31	8181	0.666	0.936	0.994	0.989	0.438	0.662	0.630
Assistments Math 2008-2009 Symb-DFA (302 Students)	302	23	4957	0.986	0.956	0.977	0.973	0.961	0.961	0.961
OLI Engineering Statics - Fall 2011 - CMU (74 students)	74	4	71805	0.973	0.967	1.000	0.999	0.882	0.976	0.979
OLI Engineering Statics - Fall 2011 - CMU (74 students)	74	5	37423	0.983	0.989	0.650	0.996	0.999	0.568	0.972
IWT Self-Explanation Study 2 (Fall 2009) (tutors only)	99	13	7094	0.200	0.822	0.945	0.916	0.538	0.198	0.064

than for AIC (2) for any non-trivially sized data set. In general, this means that BIC favors models with less parameters (again more strongly the AIC), and converges to the “true” or correct model [1], however, this does not mean that for BIC to be useful that the “true” model must exist in the set of possible models [2]. Both reduce the chance of over-fitting the data by penalizing for increasing the number of parameters in the model. They are much faster to compute than cross validation and are believed to reasonably predict the results of cross validation, though no systematic investigation of that has been performed, at least, for the kinds of EDM models investigated here. Given that AIC is more lenient, one might suspect it would be more susceptible to favoring models that over-fitted the data. On the other hand, BIC might over penalize more complex models that indeed do capture true variability in the data. Many of the previous efforts to evaluate knowledge component models in EDM have used BIC as the evaluation criteria including Learning Factors Analysis (LFA) [3], Performance Factors Analysis (PFA) [11], and Instructional Factors Analysis (IFA) [4].

2. DATA AND METHOD

DataShop has grown to include almost 400 datasets as of February 2013. One of the fundamental features available in DataShop is the ability to fit different KC Models to a dataset. There are a number of ways KC Models can be generated with DataShop.

1) KC Models can be imported with log data of an initial dataset.
2) KC Models can be exported, modified, and re-imported through DataShop’s intuitive user interface (Examples of this can be seen in the DataShop tutorial channel on Youtube [5]).
3) KC Models can be automatically generated by automated search algorithms such as LFA Search[8].
4) Every imported dataset automatically gets 2 models generated by DataShop- the Unique Step Model, which includes a KC for every step, and the Single KC model, which applies the same KC to every step.

Currently, there are 1,943 proposed KC Models in DataShop that were used for this analysis. Two conditions were established for a dataset to be included in the analysis: (1) three or more models with an equal number of observations were required and (2) the number of observations had to be greater than 800. We found 50 datasets within DataShop that met the conditions and 12 of them had more than one grouping of models (10 had 2 sets; 1 had 3 sets and 1 had 4 sets) for a total of 65 comparable KC sets. In addition to the aforementioned diversity of content and technology, the 65 KC sets have a broad range of the number of parameters (9 to 654), models (3 to 48), students (7 to 510), knowledge components (1 to 287), and observations (884 to 95,512). Such variation provides a rich environment for a deep analysis into what might be the best measure for model selection. As shown in Figure 1, DataShop provides a leaderboard of commonly used metrics across models within a dataset. We examined the correlations and rank order correlations for AIC, BIC, and Cross-validation across the 65 KC sets. We chose to report rank order correlations in addition to correlations because it is less sensitive to outliers that may excessively inflate (or, less frequently, lower) a correlation.

3. RESULTS AND DISCUSSION

After running the correlations between the metrics, we found that for the majority of KC sets (44 of 65), AIC and BIC do not agree on which model best fits the data. More importantly, AIC is overwhelmingly the better predictor when compared with cross validation best models (an average 94% match vs. BIC’s 33%). To be more precise, of the 44 comparable KC sets, 41 of AIC best models match with SSCV best models vs. 13 for BIC, for ISCV - 41 AIC best models match vs. 14 for BIC, and for NSCV - 42 AIC best models match vs. 16 for BIC. It is noteworthy that the three AIC best models that do not match with at least one cross validation best model have a substantially lower average number of KCs (12) and number of observations (5,941) than the 41 models with a match (average of 53 KCs and 17,374 observations). This appears to be because the AIC implementation in DataShop does not take into account second order Akaike Information Criterion (AIC_C) which has an adjustment for smaller sample sizes in relation to number of parameters [1]. As an example, Table 1 shows a small subset of the 65 comparable KC sets illustrating a strong positive correlation between AIC and

Table 2. Correlations and rank order correlations across the five metrics provided in DataShop (AIC, BIC, SSCV,ISCV and NSCV).

	AIC-BIC	AIC-SSCV	AIC-ISCV	AIC-NSCV	BIC-SSCV	BIC-ISCV	BIC-NSCV	SSCV-ISCV	SSCV-NSCV	ISCV-NSCV
Correlations	0.574	0.824	0.891	0.890	0.522	0.464	0.446	0.812	0.777	0.919
Rank Corr.	0.532	0.817	0.852	0.847	0.478	0.403	0.420	0.760	0.735	0.868

each of the three cross validations regardless of whether the AIC-BIC correlation is strong (rows 2-4), weak (row 5) or average (row 1). In all but one instance, the AIC correlations with cross validation are better than BIC. Table 2 shows the average correlations and rank correlations between AIC, BIC, and the Cross-validations (as stated earlier, three types of ten-fold cross validations are reported in DataShop: student stratified cross validation (SSCV), item stratified cross validation (ISCV) and non-stratified cross validation (NSCV)). From these averages in Table 2, AIC and BIC have correlations with each other of just over 0.5, which makes sense since they often do not agree on the best fitting model. More importantly, AIC is a better predictor than BIC of all three kinds of cross validation. Interestingly, table 2 shows SSCV is better indicated by AIC than the other CV metrics.

Thus, on those grounds, it seems as though AIC is the best single measure. In general, AIC best models average more knowledge components (53 vs. 34) and more parameters (205 vs. 166) than BIC best models. It is not surprising, then, that there is a high level of disagreement between best model selections for AIC and BIC (68% do not match). When comparing the best models of AIC and BIC to the best models of all three types of cross validation, AIC again matches better than BIC (approximately 70% to 10%). This better matching of best models is another strong argument that AIC is a better metric for model selection.

4. CONCLUSION AND FUTURE WORK

Although cross validation is the gold-standard for model selection, it is not a reasonable metric to use for computationally expensive processes, such as inside the LFA search, as it is too time consuming. Efficiency concerns together with uncertainty about which is a better heuristic led us to a detailed comparison of AIC/BIC across datasets and many models. Our evidence points toward AIC as the better predictor of cross validation results.

A possible reason may follow from the fact that AIC favors greater complexity within models than BIC. While the KC models in DataShop are a good approximation of student cognitive processing, it is quite likely that they significantly under-represent the true complexity of student thinking. Thus, rather than the higher bias toward simplicity that is implicit in BIC, it may be that higher complexity is a better prior belief. The true (more rich and complex) cognitive model is most likely outside the space of models that we are searching within and AIC is claimed to be better than BIC in such circumstances [14]. On the one hand, it is a positive sign of maturity of the field of Educational Data Mining that we now have so many datasets and so many alternative KC models that a comparison like this one is possible. On the other hand, it is clear that more and better research is needed to better uncover the true complexity and richness of student thinking.

It is also important to note that AIC and BIC are not the only model selection metrics available, and in the future we hope to explore alternatives for possible inclusion in DataShop. Further, the only statistical model used in this analysis was AFM. While we expect that the results would be similar with other regression based statistical models (such as PFA or IFA), we have implemented a facility in DataShop to accept external analyses, and we plan to score additional statistical models across the DataShop metrics using the external analyses support.

We thank the DataShop team for providing custom reports of the KC Model metrics and Hui Cheng for help with the LFA Search.

5. REFERENCES

[1] Burnham, K., P., Anderson, D., R. Model selection and multimodel inference. a practical information-theoretic approach. New York: Springer; 2002.

[2] Burnham, K. P.; Anderson, D. R. (2004), "Multimodel inference: understanding AIC and BIC in Model Selection", Sociological Methods and Research 33: 261–304.

[3] Cen, H., Koedinger, K. R., & Junker, B. (2006). Learning Factors Analysis: A general method for cognitive model evaluation and improvement. In *Proceedings of the 8th International Conference on ITS*, 164-175. Springer-Verlag.

[4] Chi, M., Koedinger, K., Gordon, G., Jordan, P., and VanLehn, K. (2011). Instructional factors analysis: A cognitive model for multiple instructional interventions. In *Proceedings of the 4th International Conference on Educational Data Mining*. Eindhoven, the Netherlands

[5] DataShop Tutorial 2: Exploring an alternative skill model. In DataShop Youtube Channel. Retrieved 2/25/13. From www.youtube.com/user/datashoptutorials

[6] Draney, K.L., Pirolli, P., & Wilson, M. (1995). A measurement model for complex cognitive skill. In P. Nichols, S.F. Chipman, & R.L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 103–126). Hillsdale: Erlbaum.

[7] Koedinger, K.R. & McLaughlin, E.A. (2010). Seeing language learning inside the math: Cognitive analysis yields transfer. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Conference of the Cognitive Science Society*. (pp. 471-476.) Austin, TX: Cognitive Science Society.

[8] Koedinger, K. R., McLaughlin, E. A., & Stamper, J. C. (2012). Automated Student Model Improvement. *Proceedings of the 5th International Conference on Educational Data Mining*. (pp. 17-24) Chania, Greece.

[9] Koedinger, K. R., Stamper, J. C., McLaughlin, E. A., & Nixon, T. (2013). Using data-driven discovery of better student models to improve student learning. (Submitted). *AIED 2013 - The 16th International Conference on AIED.*

[10] Koedinger, K.R., Baker, R.., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J., (2011) A Data Repository for the EDM community: The PSLC DataShop. In *Handbook of Educational Data Mining*. CRC Press.

[11] Pavlik Jr., P.I., Cen, H., Koedinger, K.R.: Learning Factors Transfer Analysis: Using Learning Curve Analysis to Automatically Generate Domain Models. In *Proceedings of the the 2nd International Conference on Educational Data Mining*, Cordoba, Spain, pp. 121-130 (2009).

[12] Stamper, J. & Koedinger, K.R. (2011). Human-machine student model discovery and improvement using data. In J. Kay, S. Bull & G. Biswas (Eds.), *Proceedings of the 15th International Conference on AIED*, pp. 353-360. Springer.

[13] Tatsuoka, K.K. (1983) Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20, 345-354.

[14] Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. Biometrika 92: 937–950.