Model-Based Collaborative Filtering Analysis of Student Response Data: Machine-Learning Item Response Theory

Yoav Bergner[*], Stefan Dröschler[†], Gerd Kortemeyer[‡], Saif Rayyan, Daniel Seaton, and David E. Pritchard
Massachusetts Institute of Technology
77 Massachusetts Ave.
Cambridge, MA 02139

ABSTRACT

We apply collaborative filtering (CF) to dichotomously scored student response data (right, wrong, or no interaction), finding optimal parameters for each student and item based on cross-validated prediction accuracy. The approach is naturally suited to comparing different models, both unidimensional and multidimensional in ability, including a widely used subset of Item Response Theory (IRT) models which obtain as specific instances of the CF: the one-parameter logistic (Rasch) model, Birnbaum's 2PL model, and Reckase's multidimensional generalization M2PL. We find that IRT models perform well relative to generalized alternatives, and thus this method offers a fast and stable alternate approach to IRT parameter estimation. Using both real and simulated data we examine cases where one- or two-dimensional IRT models prevail and are not improved by increasing the number of features. Model selection is based on prediction accuracy of the CF, though it is shown to be consistent with factor analysis. In multidimensional cases the item parameterizations can be used in conjunction with cluster analysis to identify groups of items which measure different ability dimensions.

1. INTRODUCTION

Online courses offer the prospect of large data sets of student responses to assessment activities that occur over time and under varying conditions (e.g. training, practice, graded homework, and tests). These present a more complex analysis task than test data recorded under constrained circumstances, but they offer the opportunity to learn about learning (e.g. over a semester, or from a specific intervening instructional activity) in the spirit of evidence-centered design [1]. Analyzing such data will require extensions of standard assessment methods such as Item Response Theory (IRT), for example when multiple attempts are allowed [2].

In the context of educational measurement, item response models have numerous advantages over classical test theory, and their use is widespread [3]. Despite a variety of available software packages, IRT parameter estimation is still technical and goodness-of-fit analysis continues to be a subject of research [4; 5]. In this paper we describe an alternate approach to IRT parameter estimation and goodness-of-fit motivated by machine learning. Our approach springs from an operationalist interpretation of the goals of IRT as stated by Lord [6]: "to describe the items by item parameters and the examinees by examinee parameters in such a way that we can predict probabilistically the response of any examinee to any item, even if similar examinees have never taken similar items before."

Collaborative filtering (CF) is commonly used in recommender systems with the goal of recommending unfamiliar items to a user based on ratings of those items by other users and prior rating information by the user in question [7]. The Netflix prize, for example, drew much attention to the problem of movie recommendations [8]. The idea behind any collaborative filter is that when multiple users interact with overlapping subsets of items, information from the interactions can be extracted and used to make probabilistic inferences about potential future interactions. Memory-based CFs attempt to do this by exploiting similarity between users based on a vector of their prior interactions. A naive algorithm might predict that user J will have identical interactions to those of the most similar user K (or to cluster of similar users). This descriptive approach does not attempt to model, causally or otherwise, the nature of the individual interactions. By contrast, model-based CF uses the partial interaction information to model a set of parameters for the users and the items which, taken together, can reconstruct probabilistic predictions about the missing interactions. In this aspect, CF and IRT have the same end.

The structural similarity between IRT and logistic regression has been noted in [9; 10]. Beck and Woolf [11] have applied a linear regression machine learning algorithm to an intelligent arithmetic tutor to predict when a student would answer a problem correctly (and in how much time). Desmarais and Pu [12] have compared Bayesian modeling of knowledge spaces to IRT in pursuit of examinee ability estimation. Whereas Bayesian knowledge tracing requires identification of subject-specific knowledge components, multidimensional IRT is a general framework for measuring ability along multiple dimensions.

This paper explores the application of model-based collaborative filtering (CF) to the analysis of student responses with similar goals to IRT, i.e. finding parameters for students

[*]bergner@mit.edu
[†]also: Ostfalia University of Applied Sciences
[‡]permanent affiliation: Michigan State University

and items that combine to predict student performance on an item by item basis. From machine learning, we borrow the notion of learning the model from the data. Rather than assign an item response model *a priori*, we use the CF to train a class of log-linear models on the data and select the one which performs the best in terms of prediction accuracy. The model is selected for capturing maximal information from a student response matrix, with no prior knowledge about the data assumed. We show that several standard IRT models emerge naturally as special cases.

In the remaining sections, we describe the numerical protocol for parameter estimation as well as an approach to goodness-of-fit based on prediction accuracy and cross-validation techniques standard in machine learning. The approach is naturally suited to comparing different IRT models, both unidimensional and multidimensional. We apply the CF to two sets of student response data. One of the two, containing roughly 120 online homework responses in a General Chemistry course with 2000 students, hints strongly at two dimensions of skill and discrimination for students and items respectively. We demonstrate that the items, thus parametrized by the CF, cluster into the same groupings that are suggested by principal component analysis.

2. REGULARIZED LOGISTIC REGRESSION AS A COLLABORATIVE FILTER

2.1 Parameter Estimation

We describe the collaborative filtering approach for dichotomously scored responses using regularized logistic regression. Particular IRT models obtain as a special case.

A binary classifier of individual responses is built *ab initio* around a logistic function

$$P = \frac{1}{1+e^{-Z}} \tag{1}$$

which provides a mapping from the real line to the probability interval [0,1]. We are given a response matrix U_{si} whose rows represent the response vector of student s to each item i. Each student is to be parametrized by a vector θ_k and each item by a vector X_k. The vectors are by design of commensurate dimension (known as the number of features n_f) such that a scalar product can be constructed, the logit, or inverse of the logistic function,

$$Z = \theta \cdot X = \sum_k \theta_k X_k \tag{2}$$

Although student and item indices have been suppressed, Z is a matrix product of θ ($N_s \times n_f$) and X ($n_f \times N_i$). It is useful to modify the description slightly to include a bias component (fixed, equal to 1) on either the student side or the item side, or both, by considering generalizations such as

$$\theta^* = \begin{pmatrix} 1 \\ \theta \end{pmatrix} \qquad X^* = \begin{pmatrix} X \\ 1 \end{pmatrix} \tag{3}$$

in which case

$$Z = \theta^* \cdot X^* = X_0 + \sum_k \theta_k X_k + \theta_0 \tag{4}$$

where we have taken the liberty of relabeling the indices for simplicity of presentation. The bias component in a student or item vector does not add parameter information but importantly allows the logit to be a function of the difference between student and item parameters. (Nothing is gained by having more than one bias component since a sum of student or item parameters defines a single alternate parameter with the same information). The logistic function now generates a probability (or expectation) matrix with the dimensions $N_s \times N_i$ of the response matrix U_{si},

$$P_{si} = \frac{1}{1+e^{-Z_{si}}} \tag{5}$$

The likelihood function for the observed response matrix U given the parameters θ and X is given by the product

$$L(U|\theta, X) = \prod_s \prod_i P_{si}^{U_{si}} (1 - P_{si})^{(1-U_{si})} \tag{6}$$

and remains to be maximized by suitable assignment of student and item parameters. For computational benefit, one typically uses the logarithm of the likelihood function. If we multiply the log likelihood by -1 (turning the maximum into a minimum), we can relabel the result in the convention of machine learning as the "cost function"

$$J(\theta, X) = -\sum_s \sum_i [U_{si} \log P_{si} + (1 - U_{si}) \log(1 - P_{si})] \tag{7}$$

Numerically maximizing the likelihood function L or (equivalently) minimizing the cost function J is quite fast on a modern desktop with off-the-shelf optimization packages (in our R implementation, we use `optim` with method "BFGS"). Typically these min/max finders take as arguments one long parameter vector (formed by unrolling the X and θ matrices) and a definition of the cost function and its gradient. As of this writing, a response matrix of 2000 students and 50 items takes about 10 seconds to process on a 3.4 GHz Intel i7 machine. This approach to Joint Maximum Likelihood Estimation (JMLE) no longer necessitates a stepwise update of item and student parameters as was once standard [13; 14; 15].

As the number of model features n_f is increased in any data fitting scenario, it becomes possible to minimize the cost function with parameters that do not generalize well to new data, i.e. to over-fit the data. Regularization terms may be introduced in the cost function to reduce over-fitting. To equation 7 we add the terms (sums exclude any bias components)

$$\lambda \sum_{k=1} \theta_k^2 + \lambda \sum_{k=1} X_k^2 \tag{8}$$

where the optimal regularization parameter λ can be determined from cross-validation as discussed is section 2.3.

2.2 IRT Recovered as Special Cases of the CF

It is now possible to show explicitly how IRT models emerge from this framework. To keep track of the absence or presence of the optional bias component, we label the dimensionality of the student or item vector as an ordered pair. The first component refers to the number of information-carrying

parameters while the second (either a zero or a one) indicates whether or not a bias component is used. Thus the Rasch model (operationally equivalent to the 1PL model) obtains under the arrangement

$$\left.\begin{array}{ll}\dim(\theta)=(1,1) & \theta^*=(1\ \theta)\\ \dim(X)=(1,1) & X^*=(X\ 1)\end{array}\right\}\rightarrow Z_{\mathrm{Rasch}}=X+\theta \quad (9)$$

where we have used the generalized form of the logit constructed in equation 4. The scalars θ and X here are identified with the student ability and item easiness parameters in the Rasch model.

The Birnbaum 2PL model, still unidimensional in skill, is obtained as

$$\left.\begin{array}{ll}\dim(\theta)=(1,1) & \theta^*=(1\ \theta)\\ \dim(X)=(2,0) & X=(X_1\ X_2)\end{array}\right\}\rightarrow Z_{\mathrm{2PL}}=X_1+\theta X_2 \quad (10)$$

Although the slope-intercept form of the logit appears in the literature, it is common to map X_1 and X_2 to the discrimination and difficulty parameters α and β, where $\alpha = X_2$ and $\beta = -X_1/X_2$, such that $Z = \alpha(\theta - \beta)$.

As a final example, Reckase and McKinley [15; 16; 17] have defined as M2PL the multidimensional extension of the 2PL model for m skill dimensions, which emerges here when

$$\left.\begin{array}{ll}\dim(\theta)=(m,1) & \theta^*=(1\ \theta_1\ldots\theta_m)\\ \dim(X)=(m+1,0) & X=(X_0\ X_1\ldots X_m)\end{array}\right\}\rightarrow Z_{\mathrm{M2PL}}=X_0+\sum_{i=1}^{m}\theta_m X_m \quad (11)$$

This is a *compensatory* multidimensional model to the extent that high values of one component of θ may compensate for low values in another component. However the model is still capable of describing items which have very low discrimination along one or more skills. The X_m item parameters for $m > 1$ should be seen as "discrimination-like" parameters whereas a "difficulty-like" parameter along each axis could be constructed by analogy with the 2PL model as the ratio $-X_0/X_m$.

2.3 Evaluating the Model, or Goodness-of-Fit

The CF minimization procedure results in a set of parameters for each student and item. These can be used to construct item response curves (or surfaces or hyper surfaces) as a prelude to studying model-data fit. An alternate approach however, common to machine learning algorithms, is to sequester a portion of the response matrix as a test set which is not considered during parameter estimation. Once parameters are estimated using the remaining "training" data, these same parameters are used to predict the values in the test set (where a probability value of greater than 0.5 results in the prediction of a correct item response). The percentage of correctly classified elements is the accuracy score. An intermediate test-set can be used for cross-validation, for example to adjust the regularization parameter to avoid over-fitting the training set. Moreover by subsampling multiple times (either with disjoint partitions or random subsamples) and averaging the accuracy score, subsampling variability can be controlled.

In the following section, we present results of data analyzed using this recipe, with 30% of the response matrix randomly subsampled for use as a test set, repeating 100 times. In section 4, we discuss interpretation of the accuracy score as a goodness of fit statistic.

3. SAMPLE RESULTS OF CF ANALYSIS

We analyze three data sets, two real and one simulated. The first set comes from a pre-test administration of a physics instrument, the Mechanics Baseline Test (MBT) [18] at the Massachusetts Institute of Technology over multiple years from 2005-2009 (26 items and 2300 examinees). The MBT is a standard instrument used to gauge student learning gains on and competencies with essential concepts in introductory physics. A superset of these data has been described and analyzed by Cardamone et al. using (unidimensional) IRT [19].

To test whether the collaborative filter would indeed "discover" multidimensionality of skills in student response data, we constructed a second data set of simulated responses to a two-skill test, assuming correlated skill-components but unidimensional items. In other words, 2000 skill-pairs were sampled from a multivariate Gaussian distribution and a response matrix for 60 items simulated based on a 2PL unidimensional model. Responses to the first 30 items depended only on the first skill component, while responses to the last 30 items depended only on the second component. The two skills over the sampled population were correlated with a Pearson coefficient $r = 0.58$.

The third data set comes from online homework data using LON-CAPA for a General Chemistry class at Michigan State University (MSU). The class was selected for study because it had a large student enrollment in a typical year (N = 2162), and because the 120 items were repeatedly administered over several years between 2003-2009. Although students were allowed multiple attempts on homework problems, the responses were scored correct/incorrect on first try for this analysis. No prescreening of the items was performed, and the data analysis was completely blind to the content of this course.

When the dichotomously scored response matrix contained omitted responses (up to 40% in the General Chemistry homework) the sum over matrix elements in Equation 7 and the computation of the accuracy score both excluded omitted responses.

For each data set, the model space was scanned by starting with $\dim(X) = (1\ 0), \dim(\theta) = (1\ 0)$ and proceeding incrementally subject to the commensurability constraint (i.e. to construct a scalar product of θ and X). In the figures below we denote each model by combining the dimensions of θ and X into one compact string $(\dim(\theta)\ \dim(X))$, i.e. (1010). In this notation, the model (2130) is read as containing two skill parameters plus a bias parameter and three item parameters (no bias). The apportionment of bias parameters means that both skill parameters multiply an item parameter, but there is one item parameter that remains as a term by itself in the logit.

Figures 1-3 display the accuracy scores of the CF models as the dimensionality is varied. For reference, we indicate with shaded regions the separation of the model space by the dimensionality of student skills. We also indicate with vertical dashed lines the CF models corresponding to particular IRT models. We observe that for the MBT data set, accuracy increases up to the unidimensional 2PL model, but

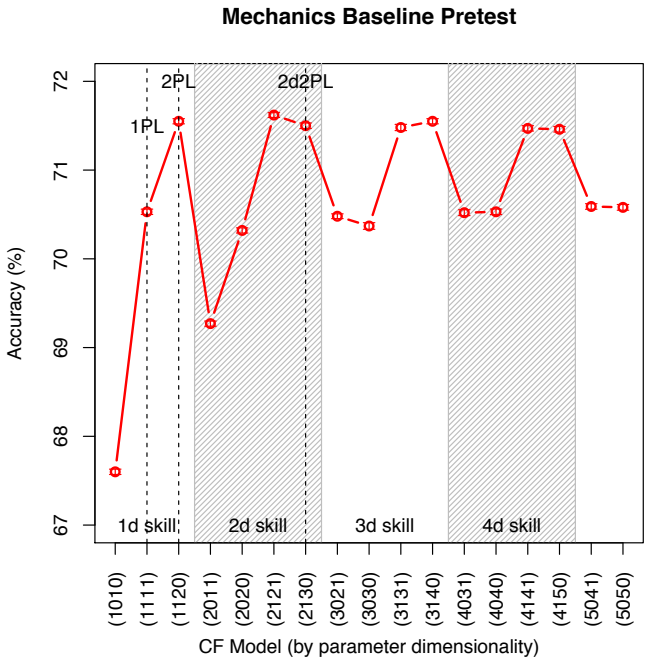


Figure 1: Model by model accuracy scores using the Mechanics Baseline Test data. Performance peaks at the 2PL model and is not improved by additional features.

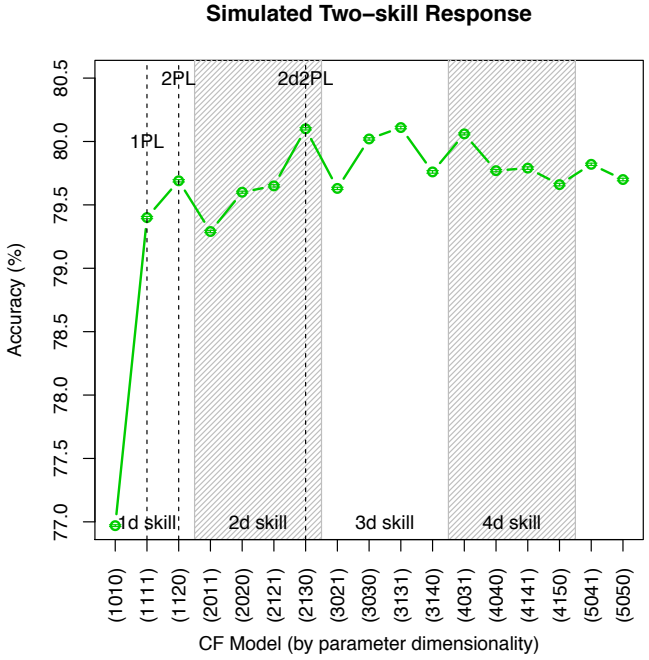


Figure 2: Model by model accuracy scores using the simulated two-skill responses. Two-dimensional models (and the 2d-2PL model in particular) perform optimally.

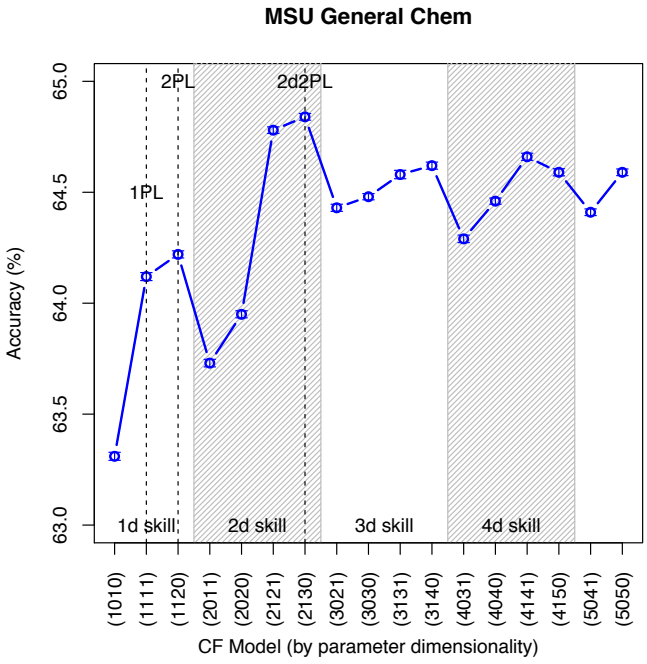
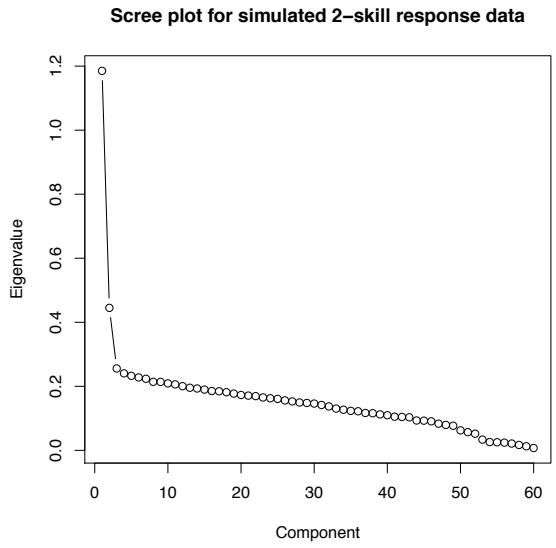


Figure 3: Model by model accuracy scores using online General Chemistry homework. Two-dimensional models (and the 2d-2PL model in particular) outperform unidimensional models.

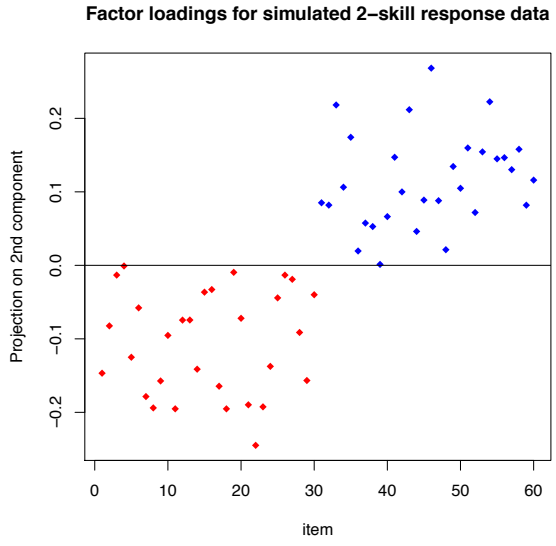
no significant gains are achieved by going to higher dimensional models. In the simulated response data based on a two-part, two-skill test, an accuracy improvement is realized by going to two-dimensional models (and the 2D-generalized 2PL MIRT model in particular), but again this asymptotic limit is not exceeded by higher-dimensional models. This is not surprising given that the simulated data were devised using two skills, but it serves as confirmation that the CF is capable of learning this feature of the data. The substantive result is that the General Chemistry analysis (Fig. 3) follows the pattern of the two-dimensional simulated data and not the unidimensional MBT data.

We note that among the four possible models representing m skill dimensions (for $m > 1$) the latter two models appear to outperform the first two (except in the case of the simulated data). The better performing models are the Reckase M2PL model (m 1 m+1 0) and a hybrid model (m 1 m 1) which could be thought of as M2PL along all but one skill component and 1PL for the remaining skill. Models with higher dimensionality require larger regularization parameters to avoid over-fitting. The apparent degradation of performance for increasing dimensionality is most likely due to over-fitting/sub-optimal choice of regularization parameter (the choice was suitable for the MBT data).

To understand the structure of the simulated two-dimensional data set and calibrate our perceptions for the General Chemistry data, we perform an exploratory factor analysis of the simulated response matrix and show the scree plot in figure 4(a). We plot the projection of each item (factor loading) onto the second principal component in figure 4(b). Whereas

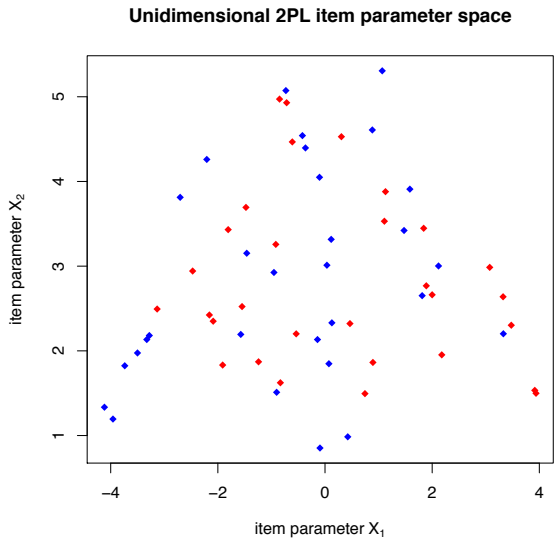


(a)

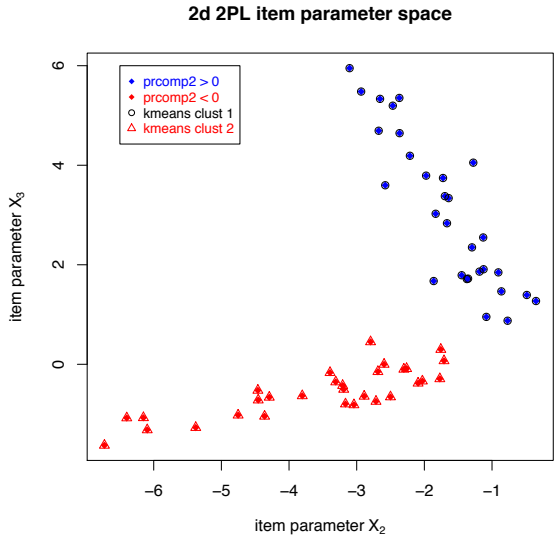


(b)

Figure 4: Simulated two-skill analysis: (a) scree plot and (b) projection of items onto second principal component for simulated data set. Color is added in (b) to identify points in later figures.



(a)



(b)

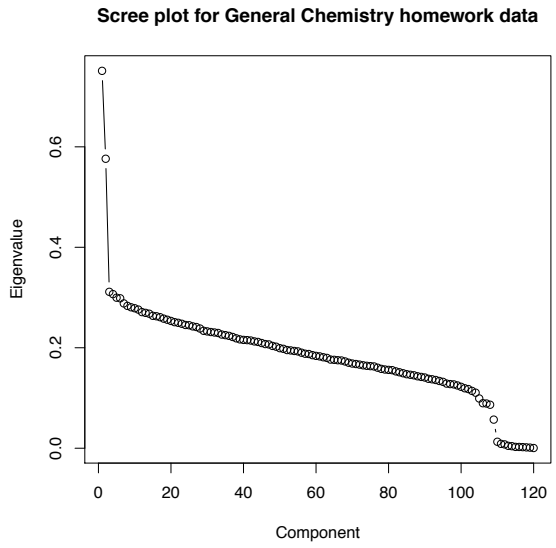
Figure 5: Simulated two-skill analysis: Item parameter space in (a) 1d- and (b) 2d-2PL IRT. Color coding is based on second principal component loading, and k-means cluster analysis is superimposed using shapes in (b).

the first principal component captures the variance in overall score or (unidimensional) skill, the second component will differentiate between students who may have the same overall score but perform proportionately better or worse on different groups of items.

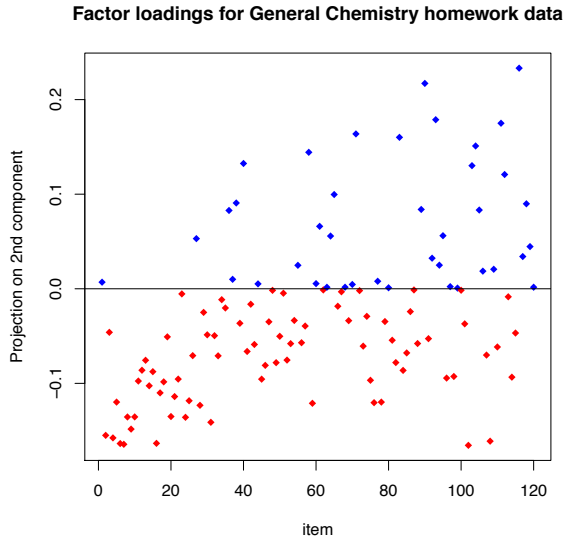
The second principal component loadings in figure 4(b) clearly distinguish two different subsets of items in the simulated data, the first and second half of the item set *by design*.

In figures 5(a) and figure 5(b), we now plot the items as points in the item parameter space generated by two CF models: the (1120) CF model corresponding to unidimensional 2PL (the full item-parameter space is 2-dimensional, spanned by X_1 and X_2) and the (2130) CF model, corresponding to 2d-2PL IRT. There are three item parameters in the latter model, and we examine the reduced parameter space spanned by the two discrimination-like parameters X_2 and X_3. The parameters plotted here come from a single run of the CF algorithm.

The unidimensional model blurs any distinction between the two known groups of items, but this distinction is manifest in the 2d-2PL model. The roughly orthogonal arms in figure 5(b) reflect the fact that in our simulated responses, each

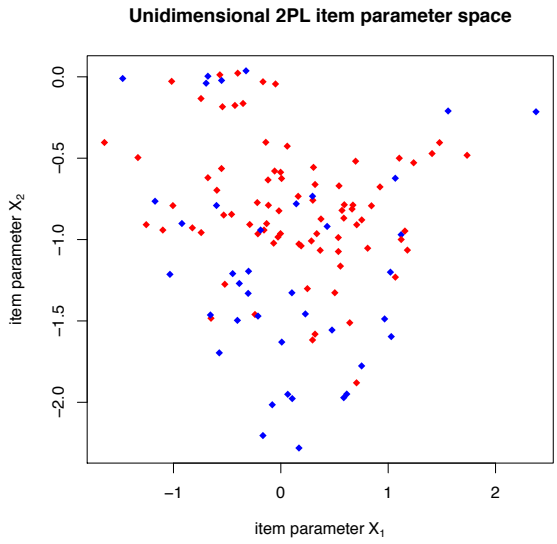


(a)

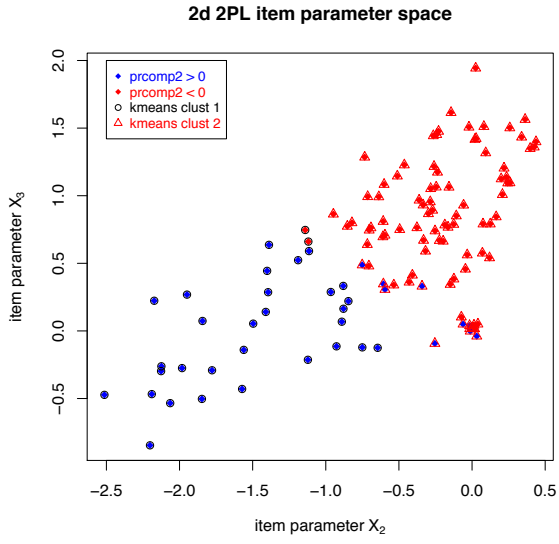


(b)

Figure 6: General Chemistry homework data justifies two-dimensional models. Scree plot (a) showing two significant eigenvalues and (b) projection of items onto second principal component. Color is added in (b) to identify item points in later figures.



(a)



(b)

Figure 7: General Chemistry: (a) unidimensional 2PL item parameter space shows little separation of colors (corresponding to loading onto the second principal component). Conversely (b) 2d-2PL IRT clearly separates color-coded items in the space of two discrimination-like parameters; moreover cluster analysis in this space identifies the border.

item was truly unidimensional in skill-dependence and thus does not discriminate at all with respect to the complementary second skill. We superimpose the results of a k-means (with $k = 2$) clustering analysis indicated by shape on the plotted points in figure 5(b). All the red (blue) points are overlapped with triangles (circles), showing that the clustering algorithm finds the same two groups that were identified by the factor loading in figure 4(b). We have verified that this clustering is manifest in three dimensions as well using a 3d-2PL model on simulated data.

We repeat the procedure to visualize the results of the General Chemistry data in figures 6(a)-7(b) with similar results. The cluster analysis in the 2d-2PL parameter space identifies the same two groups as the principal component analysis for over 90% of the items (discrepant items are those that fall very close to the zero line in figure 6(b)).

We emphasize that the choice of model (2130) was driven by the accuracy score of the CF algorithm. Given the model, the two clusters of items emerge from the assignment of

discrimination-like item parameters which best predict the response matrix. We have added no information about the items nor offered any interpretation of the meaning of the two clusters in this case, though we are working with domain experts on identifying the significance. For the simulated data, the two clusters emerge as expected from simulated responses predicated on the assumption that the different item groups test different, though correlated, abilities of the examinees.

4. INTERPRETATION OF THE ACCURACY SCORE STATISTIC

It may be noted that the overall accuracy scale differs in each of the figures 1-3, that the scores sometimes seem rather unimpressively low ($\approx$ 65%), and that in some cases the model scores for a given data set differ by only a fraction of one percent. Since we claim that this score provides a basis for preferring one model over another, it behooves us to discuss the meaning of the score value itself.

Beck and Woolf also observed that in any probabilistic binary classifier, the maximum expected accuracy score depends on the distribution of values in the probability (or expectation) matrix [11]. For example, if all probabilities (for each student-item pairing) are equal to 0.75, then all responses would be predicted by the binary classifier to be correct, though of course only 75% should be expected. Perhaps less intuitive, if the values in the probability matrix are distributed uniformly over all values in the interval $[0, 1]$, the expected accuracy score will also be 75%.

A workaround suggested in [11] is to bin the matrix elements into probability bins before comparing with the observed responses. This indeed results in a visible one-to-one correspondence between expected bin-fractions and observed bin-fractions, but bin-based statistics inevitably raise several concerns about the binning procedure itself. Certainly binning choice is not a characteristic of the model. Instead, we probe the accuracy score formally as follows. If the distribution of p values in the expectation matrix is given by a distribution function $g(p)$, then the expected accuracy score is given by the following "average"

$$S = \int_0^{0.5} (1-p)g(p)dp + \int_{0.5}^1 pg(p)dp \qquad (12)$$

where the first term accounts for predicted-to-be-wrong and the second term for predicted-to-be-right matrix elements. The shape of $g(p)$ in turn depends on the distribution of the student and item parameters and the function that is used to model the probability. As an explicit example, for the Rasch or 1PL model, the probability of a correct response when the student skill is θ and the problem difficulty is β is given by

$$p = \frac{1}{1+e^{-(\theta-\beta)}} \qquad (13)$$

If student skills are distributed as $g_\theta(\theta)$ and item difficulties as $g_\beta(\beta)$, then $g(p)$ can be shown to be the convolution

$$g(p) = \frac{1}{p(1-p)} \int_{-\infty}^{\infty} g_\theta(\theta) g_\beta \left(\theta + \ln\frac{1-p}{p}\right) d\theta \qquad (14)$$

Although the model dependence has been folded into equation 14, the dependence on the distribution of item difficulties is explicit. The accuracy score thus cannot be meaningfully compared for two different data sets unless the examinees and items are drawn from very similar distributions. For 2PL and M2PL models, the best score will also be a function of the distribution of item discriminations. In fact, we have observed that after removing two MBT items with pathological item response curves found in [19], prediction accuracy on the remaining data increased by 2 percentage points, while this gain was not observed when two randomly selected problems were removed.

In view of the model dependence of equation 14, a cautionary flag might be raised in using the accuracy score to compare different models on a given data set. However since the models are designed to predict the data, we argue that this model-dependence is justly accounted for in using the accuracy score as a goodness-of-fit statistic.

In practice it is much easier to calculate the expected score in equation 12 numerically from the expectation matrix without any integrals. Simply replace all probabilities less than 0.5 by one minus the probability and average over the resulting matrix.

5. SUMMARY AND CONCLUSIONS

We have applied a model-based collaborative filter, i.e. a numerical method for analyzing a dichotomous student response matrix with the goal of predicting the observed responses. Relying on readily available optimization code, the CF is fast, flexible and stable. We showed that CF naturally parameterizes a series of models with increasing dimensionality and that this family contains several common unidimensional and multidimensional IRT models.

We showed with sample data that the CF can aid in model-selection and that the multidimensional-model capability can result in improved prediction accuracy and easy investigation of whether the data are better fit by alternate models. Practitioners of IRT will be pleased to learn that, at least in the cases considered here, CF was not able to improve significantly on the quality of fit achieved using standard, but in two cases multidimensional, IRT models. Moreover, the dimensionality of models suggested by the CF and the clustering of items in the ensuing parameterizations are consistent with results from exploratory factor analysis.

Finally, the stability, speed, close connection with IRT, and easy generalizability of CF recommends it very highly for use in analyzing student response data of all sorts.

Acknowledgements

We are grateful to MIT and NSF Grant No. DUE-1044294 for supporting this work. Additionally, we thank Bob Field and Sarah Seaton for their help in diagnosing chemistry questions.

6. REFERENCES

[1] Robert J. Mislevy and G. D. Haertel. Implications of Evidence-Centered Design for Educational Testing. *Educational Measurement: Issues and Practice*, 25(4):6–20, 2006.

[2] Young-Jin Lee, David J Palazzo, Rasil Warnakulasooriya, and David E. Pritchard. Measuring student

learning with item response theory. *Physical Review Special Topics - Physics Education Research*, 4(1):1–6, January 2008.

[3] Ronald K. Hambleton and Russell W. Jones. Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. *Educational Measurement: Issues and Practice*, 12(3):38–47, October 1993.

[4] Maria Orlando and D Thissen. Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models. *Applied Psychological Measurement*, 24(1):50–64, 2000.

[5] Maria Orlando and David Thissen. Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, 27(4):289–298, 2003.

[6] F.M. Lord. *Applications of item response theory to practical testing problems*. Erlbaum Associates, 1980.

[7] Xiaoyuan Su and Taghi M. Khoshgoftaar. A Survey of Collaborative Filtering Techniques. *Advances in Artificial Intelligence*, 2009(Section 3):1–19, 2009.

[8] J Bennett and S Lanning. The Netflix Prize. *Proceedings of KDD Cup and Workshop*, 2007(10):8, 2007.

[9] A.D. Wu and B.D. Zumbo. Thinking About Item Response Theory from a Logistic Regression Prespective. In Shlomo Sawilowsky, editor, *Real Data Analysis*, pages 241–269. Information Age Publishing, 2007.

[10] Patrick Mair, Steven P. Reise, and Peter M. Bentler. IRT Goodness-of-Fit Using Approaches from Logistic Regression. *Department of Statistics Papers, UCLA*, 2008.

[11] Joseph E Beck and Beverly Park Woolf. High-level student modeling with machine learning. In *Intelligent tutoring systems*, pages 584–593. Springer, 2000.

[12] Michel C. Desmarais and Xiaoming Pu. A Bayesian student model without hidden nodes and its comparison with item response theory. *International Journal of Artificial Intelligence in*, 2005.

[13] Robert J. Mislevy and Martha L. Stocking. A consumer's guide to LOGIST and BILOG. *Applied psychological measurement*, 13(1):57, 1989.

[14] Frank B Baker and Seock-Ho Kim. *Item Response Theory: Parameter Estimation Techniques*, volume 176 of *Statistics*. Marcel Dekker, 2004.

[15] R. L. McKinley and M. D. Reckase. MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods and Instrumentation*, 15:389–390, 1983.

[16] M. D. Reckase and R. L. McKinley. The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement*, 15(4):361, 1991.

[17] M. D. Reckase. *Multidimensional Item Response Theory*. Springer, 2009.

[18] David Hestenes and Malcolm Wells. A mechanics baseline test. *The physics teacher*, 30(3):159–166, 1992.

[19] Caroline N. Cardamone, Jonathan E. Abbott, Saif Rayyan, Daniel T. Seaton, Andrew Pawl, David E. Pritchard, N. Sanjay Rebello, Paula V. Engelhardt, and Chandralekha Singh. Item response theory analysis of the mechanics baseline test. In *AIP Conference Proceedings*, volume 1413, pages 135–138, February 2012.