

Multidimensional Student Skills with Collaborative Filtering

Yoav Bergner, Saif Rayyan, Daniel Seaton and David E. Pritchard

Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139.

Abstract. Despite the fact that a physics course typically culminates in one final grade for the student, many instructors and researchers believe that there are multiple skills that students acquire to achieve mastery. Assessment validation and data analysis in general may thus benefit from extension to multidimensional ability. This paper introduces an approach for model determination and dimensionality analysis using collaborative filtering (CF), which is related to factor analysis and item response theory (IRT). Model selection is guided by machine learning perspectives, seeking to maximize the accuracy in predicting which students will answer which items correctly. We apply the CF to response data for the Mechanics Baseline Test and combine the results with prior analysis using unidimensional IRT.

Keywords: formative assessment, testing, item response theory, collaborative filtering
PACS: 01.40.Fk, 01.40.gf, 01.50.Kw

INTRODUCTION

The field of physics education research (PER) has its own measurement problem: “knowing what students know” is key to probing learning gains, comparing pedagogical strategies, and ultimately making decisions aimed at improving education [1]. The development within PER of widely-used instruments like the Force Concept Inventory has had great impact by making manifest the shortcomings of traditional instruction for fostering conceptual learning in mechanics and, as remediation of that gap, the value of pedagogical shifts towards interactive engagement [2]. Mining the comparative gains in pre-post testing has opened up many frontiers in both pedagogical and sociological aspects of education reform, for example in exploring the gender gap [3].

In recent years new attention has been paid to validation studies using tools from or inspired by the field of psychometrics and educational measurement, e.g. [4, 5]. Practitioners are hearing more about Item Response Theory (IRT) and unidimensionality—the idea that the likelihood of a correct response to any of the items in a set depends on only a single latent ability in the student. To be fair, this idea is implicit in any test whose outcome is characterized by a single score.

Ding and Beichner reviewed a number of methods of analysis for multiple-choice assessments including IRT, factor analysis and cluster analysis [6]. Wang and Bao applied IRT analysis to the FCI and argued, for example, that it satisfies the assumption of unidimensionality [7]. Wallace and Bailey applied IRT to a concept inventory used in astronomy education research, finding problematic items via poor model-data fits as well as evidence for violations of unidimensionality using test-subtest comparisons [8]. And Cardamone *et al.* used IRT to discover two problematic test items (which could be understood

in terms of the unfortunate double-negation effect of ambiguous presentation and a common student misconception) in the Mechanics Baseline Test (MBT) [9].

Item response models have numerous advantages over classical test theory [10], but despite a variety of software packages, IRT parameter estimation and goodness-of-fit testing is still rather technical. PER practitioners are not by and large compelled by the concerns within the psychometric community for rigorous, sample-independent statistics as much as they are interested in knowing what their students know. Thus violations of unidimensionality are not likely to be perceived as a defect in the design of a physics instrument. On the contrary, the broad research on differentiated learning, conceptual learning, sense-making and problem-solving would seem to suggest that multidimensional ability would be the rule of the day. The PER community thus stands to benefit from an approach to IRT parameter estimation that is more intuitive and accommodating to a varying number of underlying ability dimensions.

One such approach, motivated by ideas from machine-learning, is the subject of this paper. It springs from an operationalist interpretation of the goals of IRT as stated by Lord [11]: “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 [12]. The Netflix prize, for example, drew much attention to the problem of movie recommendations [13]. The idea behind any collaborative filter is that when many users interact with overlapping subsets of

items, extracted information can be used to make inferences about potential new interactions. A model-based CF uses the interaction data to model parameters for the users and the items which, taken together, can reconstruct probabilistic predictions about the missing interactions.

The principle reason for the label of collaborative filtering to an IRT-type analysis is that we run logistic regression on all of the student and item parameters *simultaneously*. In this sense, our approach would be classified as joint (as opposed to marginal) maximum likelihood estimation, or as finding the item parameters *conditional* on the student parameters, not unconditionally. This approach, not uncommon in the early days of IRT, has fallen out of favor because it does not offer the kind of statistical guarantees which are critical for implementations on high stakes tests. We maintain however that the parameters lead to better fits and are thus better representations of the information content of a responses matrix *vis a vis those particular students*, i.e. even if the student abilities are not normally distributed.

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. We show that several standard IRT models emerge naturally.

Given the space constraints of these proceedings, we defer mathematical details, including a discussion of goodness-of-fit, to a lengthier paper. We include here enough information about the approach to demonstrate the parsimony of the assumptions. We then apply the CF to the same MBT data set in [9] and discuss the evidence for multidimensionality and the clustering of items.

CF PARAMETER ESTIMATION

A binary classifier of individual responses (i.e. a predictor of correct/incorrect) is built *ab initio* around any function which provides a mapping from the real line (continuous parameter space) to the probability interval [0,1]. A convenient choice is the logistic function,

$$P_{si} = \frac{1}{1 + e^{-z_{si}}}. \quad (1)$$

We observe a response matrix U_{si} whose rows represent the response vector of student s to each item i . P_{si} is a parameter-dependent expectation matrix. Each student is to be parametrized by a vector θ_k and each item by a vector X_k of dimension m , such that a scalar product can be constructed, the logit, or inverse of the logistic function,

$$z_{si} = \theta_s \cdot X_i = \sum_k \theta_{sk} X_{ik} \quad (2)$$

z is the matrix product of θ ($N_s \times m$) and X ($m \times N_i$). It is useful to modify the description slightly to include a

bias component on the student or item vector, or both, by considering generalizations such as

$$\theta_s^* = (1 \ \theta_s) \quad X_i^* = \begin{pmatrix} X_i \\ 1 \end{pmatrix} \quad (3)$$

in which case

$$z_{si} = \theta_s^* \cdot X_i^* = X_{i,1} + \sum_k \theta_{s,k} X_{i,k+1} + \theta_{s,k+1}. \quad (4)$$

The bias component does not add parameter information but importantly allows the logit to be a function of the difference between student and item parameters. The likelihood function for the observed response matrix U given the parameters θ and X is given by the product

$$L(U|\theta, X) = \prod_s \prod_i P_{si}^{U_{si}} (1 - P_{si})^{(1-U_{si})} \quad (5)$$

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, which we relabel as a ‘‘cost’’ function by negation:

$$J(\theta, X) = - \sum_s \sum_i [U_{si} \log P_{si} + (1 - U_{si}) \log(1 - P_{si})] \quad (6)$$

Numerically minimizing the cost function J is very fast on a modern desktop with off-the-shelf optimization packages (in our R implementation, we use `optim` with method ‘‘L-BFGS-B’’).

As the number of model features m is increased in any data fitting scenario, numerical algorithms will over-fit the data. Regularization terms may be added to Eq. 6 to reduce over fitting (sums exclude bias components):

$$\lambda \sum_{k=1} \theta_k^2 + \lambda \sum_{k=1} X_k^2 \quad (7)$$

where the optimal regularization parameter λ is determined by cross-validation.

The CF minimization procedure outputs a set of parameters for each student and item with as many dimensions as contained in the model. An approach to model-selection, common to machine learning algorithms, is to sequester a portion of the response matrix as a test set which is withheld 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. The percentage of correctly classified elements is the accuracy score (i.e. a probability of greater than 0.5 results in the prediction of a correct item response; alternately, root mean squared error is computed on the raw probabilities). An intermediate test-set can be used for cross-validation, for example to adjust the regularization parameter (7) to avoid over-fitting the training

set. By subsampling multiple times (either with disjoint partitions or random subsamples) and averaging the accuracy score, subsampling variability is controlled.

IRT Models from the CF

Several IRT models emerge from this framework. We first label the dimensionality of the student or item vector as an ordered pair, where the first component refers to the number of information-carrying parameters and the second (binary) indicates whether or not a bias component is used. We summarize the correspondence between CF labels and IRT models in Tables 1-2 below.

TABLE 1. Correspondence between CF labels and standard unidimensional IRT models

dim(θ)	dim(X)	logit	IRT model	CF label
(1, 1)	(1, 1)	$X + \theta$	Rasch (1PL)	(1111)
(1, 1)	(2, 0)	$X_1 + \theta X_2$	2PL	(1120)

(The abbreviations 1PL and 2PL refer to one- and two-parameter logistic models, respectively.) 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 $z = \alpha(\theta - \beta)$.

TABLE 2. Correspondence between CF labels and generalized multidimensional IRT models

logit	IRT model	CF label
$\sum \theta_m X_m$	non-standard	(m 0 m 0)
$X_1 + \sum \theta_m X_{m+1} + \theta_{m+1}$	non-standard	(m 1 m 1)
$\theta_1 + \sum \theta_{m+1} X_m$	non-standard	(m+1 0 m 1)
$X_1 + \sum \theta_m X_{m+1}$	M2PL	(m 1 m+1 0)

The final example in Table 2 is Reckase and McKinley’s multidimensional extension of the 2PL model, M2PL [14]. This is a *compensatory* 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 $-X_1/X_m$.

CF ANALYSIS OF THE MBT

Details about the Mechanics Baseline Test (MBT) [15] and the data set from the Massachusetts Institute of Technology ($N = 4700$) are described in [9].

The CF analysis proceeds by scanning the model space incrementally in search of the best accuracy score (see Fig. 1). When the best model is identified, the student and item parameters corresponding to that model are ob-

Mechanics Baseline Test – model performance

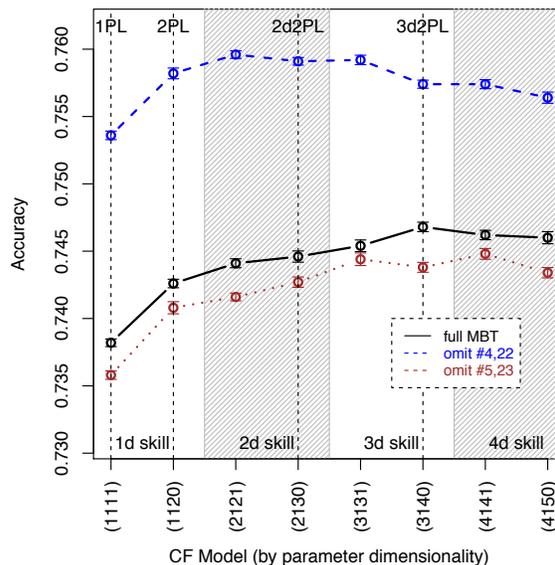


FIGURE 1. Model accuracy scores using the MBT data. Performance peaks at the 3-d 2PL model. Dashed (dotted) lines show performance if two items are removed. Removing the two pathological items from [9] increases accuracy overall and appears to obviate the need for 3 skill dimensions (dashed line), whereas removing two arbitrary items simply reduces model performance (dotted line).

tained (the item parameters are represented in Fig. 2). In Fig. 1, we denote each model by the convention in Table 2. For clarity, we plot only the two best of the four possible models for each value of m (feature dimension). Models with higher dimensionality require larger regularization parameters to avoid over-fitting, effects of which can be observed if higher-dimensional models perform worse than the models they subsume. 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 full MBT data set, shown in the solid line in Fig. 1, accuracy increases up to the 3-dimensional 2PL model (3140), and no significant gains are achieved by going to higher dimensional models. Student responses on the MBT thus suggest three student-ability parameters coupled to three item discrimination-like factors. The items, thus parametrized, are plotted in Fig. 2, where we have used an expectation-maximization (EM) clustering algorithm and shape-coded the items by the four clusters thus obtained.

We note the following features: the first two questions on the MBT form their own cluster (the points and labels lie almost on top of each other in Fig. 2), as do

