Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

MOOCs Meet Measurement Theory: A Topic-Modelling Approach

Jiazhen He,[†⋆] Benjamin I. P. Rubinstein, James Bailey,[†⋆] Rui Zhang[†]
Sandra Milligan,[‡] Jeffrey Chan[§]
[†]Dept. Computing and Information Systems, The University of Melbourne, Australia [⋆]National ICT Australia
[‡]Melbourne Graduate School of Education, The University of Melbourne, Australia
[§]School of Computer Science and Information Technology, RMIT University, Australia
{jiazhenh@student., baileyj, brubinstein, rui.zhang, s.milligan}@unimelb.edu.au, jeffrey.chan@rmit.edu.au

Abstract

This paper adapts topic models to the psychometric testing of MOOC students based on their online forum postings. Measurement theory from education and psychology provides statistical models for quantifying a person's attainment of intangible attributes such as attitudes, abilities or intelligence. Such models infer latent skill levels by relating them to individuals' observed responses on a series of items such as quiz questions. The set of items can be used to measure a latent skill if individuals' responses on them conform to a Guttman scale. Such well-scaled items differentiate between individuals and inferred levels span the entire range from most basic to the advanced. In practice, education researchers manually devise items (quiz questions) while optimising well-scaled conformance. Due to the costly nature and expert requirements of this process, psychometric testing has found limited use in everyday teaching. We aim to develop usable measurement models for highly-instrumented MOOC delivery platforms, by using participation in automatically-extracted online forum topics as items. The challenge is to formalise the Guttman scale educational constraint and incorporate it into topic models. To favour topics that automatically conform to a Guttman scale, we introduce a novel regularisation into non-negative matrix factorisation-based topic modelling. We demonstrate the suitability of our approach with both quantitative experiments on three Coursera MOOCs, and with a qualitative survey of topic interpretability on two MOOCs by domain expert interviews.

Introduction

Massive Open Online Courses (MOOCs) have recently been the subject of a number of studies within disciplines as varied as education, psychology and computer science (Ramesh et al. 2014c; Anderson et al. 2014; Kizilcec, Piech, and Schneider 2013; Dıez et al. 2013; Milligan 2015). With few studies taking a truly cross-disciplinary approach, this paper is the first to marry topic modelling with measurement theory from education and psychology.

Measurement in education and psychology is the process of assigning a number to an attribute of an individual in such a way that individuals can be compared to one another (Pedhazur and Schmelkin 1991). These attributes are often intangible such as attitudes, abilities or intelligence. Since the attribute to be measured is not directly observable, a set of items is often devised manually and individuals' responses on the items are collected. Based on a modelled correspondence with observed item responses, latent attribute levels of a cohort can be inferred. This process is called *scaling* (De Ayala 2013).

A *Guttman scale* (Guttman 1950) is one which induces a total ordering on items—an individual who successfully answers/agrees with a particular item also answers/agrees with items of lower rank-order. Table 1 depicts an example Guttman scale measuring mathematical ability (Abdi 2010), where the items are ordered in increasing latent difficulty, from *Counting* to *Division*. Here the total score corresponds to the persons' latent ability: the greater the higher.

In MOOCs, as in the traditional classroom, we may hypothesise that students possess a latent ability in the subject at hand. For example, in a MOOC on macroeconomics, students are expected to develop knowledge in introductory macroeconomics via videos, quizzes and forums. Students' latent abilities can be defined, validated and measured using indicators drawn from student responses to activities like interaction with videos, quiz results and forum participation. Unlike the traditional classroom, MOOCs create new challenges and opportunities for measurement through the multiple modes of student interaction online—all monitored at large scale. The education research community is broadly interested in whether and how latent complex patterns of engagement might evidence the possession of a latent skill, and not just explanatory variables (*e.g.*, visible quizzes and assignments) by themselves (Milligan 2015).

Table 1: An example of a perfect Guttman scale measuring mathematical ability(Abdi 2010) , where 1 means the person has mastered the item and 0 for not. Person 5 who has mastered the most difficult item *Division*, is expected to have mastered all easier items as well.

	Item 1	**Item 2**	**Item 3**	**Item 4**	**Item 5**	**Total**
	(Counting)	**(+)**	**(−)**	**(×)**	**(÷)**	**Score**
Person 1	1	0	0	0	0	1
Person 2	1	1	0	0	0	2
Person 3	1	1	1	0	0	3
Person 4	1	1	1	1	0	4
Person 5	1	1	1	1	1	5

This paper focuses on using the content of forum discussion in MOOCs for measurement, which is too time-consuming to analyse manually but that can provide a predictive indicator of achievement (Beaudoin 2002). We automatically generate items (topics) from unstructured forum data using topic modelling. Our goal is to discover items on which dichotomous (posting on a topic or not) student responses conform to a Guttman scale; where items are interpretable to subject-matter experts who could be teaching such MOOCs. For example, for a MOOC on discrete optimisation, our goal is to automatically discover topics such as *How to use platform/python*—the easiest which most students contribute to—and *How to design and tune simulated annealing and local search*—a more difficult topic which only a few students might post on. Such well-scaled items can be used for curriculum design and student assessment.

The challenge is to formalise the Guttman scale educational constraint and incorporate it into topic models. We opt to focus on non-negative matrix factorisation (NMF) approaches to topic modelling, as these admit natural integration of the Guttman scale educational constraint.

Contributions. The main contributions of this paper are:

- A first study of how a machine learning technique, NMF-based topic modelling, can be used for the education research topic of psychometric testing;
- A novel regularisation of NMF that incorporates the educational constraint that inferred topics form a Guttman scale; and accompanying training algorithm;
- Quantitative experiments on three Coursera MOOCs covering a broad swath of disciplines, establishing statistical effectiveness of our algorithm; and
- A carefully designed qualitative survey of experts in two MOOC subjects, which supports the interpretability of our results and suggests their applicability in education.

Related Work

Various studies have been conducted into MOOCs for tasks such as dropout prediction (Halawa, Greene, and Mitchell 2014; Yang et al. 2013; Ramesh et al. 2014b; Kloft et al. 2014; He et al. 2015b), characterising student engagement (Anderson et al. 2014; Kizilcec, Piech, and Schneider 2013; Ramesh et al. 2014b) and peer assessment (Dıez et al. 2013; Piech et al. 2013; Mi and Yeung 2015).

Forum discussions in MOOCs have been of interest recently, due to the availability of rich textual data and social behaviour. For example, Wen, Yang, and Rose (2014) use sentiment analysis to monitor students' trending opinions towards the course and to correlate sentiment with dropouts over time using survival analysis. Yang et al. (2015) predict students' confusion with learning activities as expressed in the discussion forums using discussion behaviour and clickstream data, and explore the impact of confusion on student dropout. Ramesh et al. (2015) predict sentiment in MOOC forums using hinge-loss Markov random fields. Yang, Adamson, and Rosé (2014) study question recommendation in discussion forums based on matrix factorisation. Gillani et al. (2014) find communities using Bayesian Non-Negative Matrix Factorisation. Despite this variety of works, no machine learning research has explored forum discussions for the purpose of measurement in MOOCs.

Topic modelling has been applied in MOOCs for tasks such as understanding key themes in forum discussions (Robinson 2015), predicting student survival (Ramesh et al. 2014a), study partner recommendation (Xu and Yang 2015) and course recommendation (Apaza et al. 2014). However, to our knowledge, no studies have leveraged topic modelling for measurement. More generally, psychometric models have enjoyed only fleeting attention by the machine learning community previously.

Preliminaries and Problem Formalisation

We choose NMF as the basic approach to discover forum topics due to the interpretability of topics produced, and the extensibility of its optimisation program. We begin with a brief overview of NMF and then define our problem.

Non-Negative Matrix Factorisation (NMF)

Given a non-negative matrix $\mathbf{V} \in \mathbb{R}^{m \times n}$ and a positive integer k, NMF factorises $\mathbf{V}$ into the product of a non-negative matrix $\mathbf{W} \in \mathbb{R}^{m \times k}$ and a non-negative matrix $\mathbf{H} \in \mathbb{R}^{k \times n}$

$$\mathbf{V} \approx \mathbf{WH}$$

A commonly-used measure for quantifying the quality of this approximation is the Frobenius norm between $\mathbf{V}$ and $\mathbf{WH}$. Thus, NMF involves solving the following optimisation problem,

$$\min_{\mathbf{W},\mathbf{H}} \|\mathbf{V} - \mathbf{WH}\|_F^2 \quad \text{s.t.} \quad \mathbf{W} \geq \mathbf{0}, \ \mathbf{H} \geq \mathbf{0} \ . \tag{1}$$

The objective function is convex in $\mathbf{W}$ and $\mathbf{H}$ separately, but not together. Therefore standard optimisers are not expected to find a global optimum. The multiplicative update algorithm (Lee and Seung 2001) is commonly used to find a local optimum, where $\mathbf{W}$ and $\mathbf{H}$ are updated by a multiplicative factor that depends on the quality of the approximation.

Problem Statement

We explore the automatic discovery of forum discussion topics for measurement in MOOCs. Our central tenet is that topics can be regarded as useful items for measuring a latent skill, if student responses to these items conform to a Guttman scale, and if the topics are semantically-meaningful to domain experts. As Guttman scale item responses are typically dichotomous, we consider item responses to be whether a student posts on the topic or not. Our goal is to generate a set of meaningful topics that yield a student-topic matrix conforming to the properties of a *Guttman scale*, *e.g.*, a near-triangular matrix (see Table 1). This process can be cast as optimisation. We apply such well-scaled topics to measure skill attainment—as level of forum participation is known to be predictive of learning outcomes (Beaudoin 2002).

Using NMF, a word-student matrix $\mathbf{V}$ can be factorised into two non-negative matrices: word-topic matrix $\mathbf{W}$ and topic-student matrix $\mathbf{H}$. Our application requires that the topic-student matrix $\mathbf{H}$ be **a) Binary** ensuring the response of a student to a topic is dichotomous; and **b) Guttman-scaled** ensuring the student responses to topics conform to a Guttman scale. NMF provides an elegant framework for incorporating these educational constraints via adding novel regularisation, as detailed in the next section. A glossary of important symbols used in this paper is given in Table 2.

Table 2: Glossary of symbols

Symbol	Description
m	the number of words
n	the number of students
k	the number of topics
$\mathbf{V} = (v_{ij})_{m\times n}$	word-student matrix
$\mathbf{W} = (w_{ij})_{m\times k}$	word-topic matrix
$\mathbf{H} = (h_{ij})_{k\times n}$	topic-student matrix
$\mathbf{H}_{ideal} = ((h_{ideal})_{ij})_{k\times n}$	exemplar topic-student matrix with ideal Guttman scale
$\lambda_0, \lambda_1, \lambda_2$	regularisation coefficients

NMF for Guttman scale (NMF-Guttman)

Primal Program

We introduce the following regularisation terms on $\mathbf{W}$ to prevent overfitting, and on $\mathbf{H}$ to encourage a binary solution and Guttman scaling:

- $\|\mathbf{W}\|_F^2$ to prevent overfitting;
- $\|\mathbf{H} - \mathbf{H}_{ideal}\|_F^2$ to encourage a Guttman-scaled $\mathbf{H}$, where $\mathbf{H}_{ideal}$ is a constant matrix with ideal Guttman scale;
- $\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|_F^2$ to encourage a binary solution $\mathbf{H}$, where operator $\circ$ denotes the Hadamard product.

Binary matrix factorisation (BMF) is a variation of NMF, where the input matrix and the two factorised matrices are all binary. Inspired by the approach of Zhang et al. (2007) and Zhang et al. (2010), we add regularisation term $\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|_F^2$. Noting this term equals $\|\mathbf{H} \circ (\mathbf{H} - \mathbf{1})\|_F^2$, it is clearly minimised by binary $\mathbf{H}$.

These terms together yield the objective function

$$f(\mathbf{W},\mathbf{H}) = \|\mathbf{V} - \mathbf{W}\mathbf{H}\|_F^2 + \lambda_0\|\mathbf{W}\|_F^2 + \lambda_1\|\mathbf{H} - \mathbf{H}_{ideal}\|_F^2 + \lambda_2\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|_F^2 \ , \tag{2}$$

where $\lambda_0, \lambda_1, \lambda_2 > 0$ are regularisation parameters; with primal program

$$\min_{\mathbf{W},\mathbf{H}} f(\mathbf{W},\mathbf{H}) \quad \text{s.t.} \quad \mathbf{W} \geq \mathbf{0},\ \mathbf{H} \geq \mathbf{0} \ . \tag{3}$$

Algorithm

A local optimum of program (3) is achieved via iteration

$$w_{ij} \leftarrow w_{ij} \frac{(\mathbf{V}\mathbf{H}^T)_{ij}}{(\mathbf{W}\mathbf{H}\mathbf{H}^T + \lambda_0\mathbf{W})_{ij}} \tag{4}$$

$$h_{ij} \leftarrow h_{ij} \frac{(\mathbf{W}^T\mathbf{V})_{ij} + 4\lambda_2 h_{ij}^3 + 3\lambda_2 h_{ij}^2 + \lambda_1 (h_{ideal})_{ij}}{(\mathbf{W}^T\mathbf{W}\mathbf{H})_{ij} + 6\lambda_2 h_{ij}^3 + (\lambda_1 + \lambda_2) h_{ij}} \tag{5}$$

These rules for the constrained program can be derived via the Karush-Kuhn-Tucker conditions necessary for local optimality. First we construct the unconstrained Lagrangian

$$\mathcal{L}(\mathbf{W},\mathbf{H},\boldsymbol{\alpha},\boldsymbol{\beta}) = f(\mathbf{W},\mathbf{H}) + \mathrm{tr}(\boldsymbol{\alpha}\mathbf{W}) + \mathrm{tr}(\boldsymbol{\beta}\mathbf{H}) \ ,$$

where $\alpha_{ij}, \beta_{ij} \leq 0$ are the Lagrangian dual variables for inequality constraints $w_{ij} \geq 0$ and $h_{ij} \geq 0$ respectively, and $\boldsymbol{\alpha} = [\alpha_{ij}]$, $\boldsymbol{\beta} = [\beta_{ij}]$ denote their corresponding matrices.

The KKT condition of stationarity requires that the derivative of $\mathcal{L}$ with respect to $\mathbf{W}, \mathbf{H}$ vanishes:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\mathbf{W}} =& 2\left(\mathbf{W}^\star\mathbf{H}^\star\mathbf{H}^{\star T} - \mathbf{V}\mathbf{H}^{\star T} + \lambda_0\mathbf{W}^\star\right) + \boldsymbol{\alpha}^\star = \mathbf{0} \ , \\ \frac{\partial\mathcal{L}}{\partial\mathbf{H}} =& 2\Big(\mathbf{W}^{\star T}\mathbf{W}^\star\mathbf{H}^\star - \mathbf{W}^{\star T}\mathbf{V} + (\lambda_1 + \lambda_2)\mathbf{H}^\star \\ & - \lambda_1\mathbf{H}_{ideal}\Big) + 4\lambda_2\mathbf{H}^\star \circ \mathbf{H}^\star \circ \mathbf{H}^\star \\ & - 6\lambda_2\mathbf{H}^\star \circ \mathbf{H}^\star + \boldsymbol{\beta}^\star = \mathbf{0} \ .\end{aligned}$$

Complementary slackness $\alpha_{ij}^\star w_{ij}^\star = \beta_{ij}^\star h_{ij}^\star = 0$, implies:

$$\begin{aligned}0 =& \Big(\mathbf{V}\mathbf{H}^{\star T} - \mathbf{W}^\star\mathbf{H}^\star\mathbf{H}^{\star T} - \lambda_0\mathbf{W}^\star\Big)_{ij} w_{ij}^\star \ , \\ 0 =& \Big(\mathbf{W}^{\star T}\mathbf{V} + 3\lambda_2\mathbf{H}^\star \circ \mathbf{H}^\star + \lambda_1\mathbf{H}_{ideal} - \mathbf{W}^{\star T}\mathbf{W}^\star\mathbf{H}^\star \\ & -2\lambda_2\mathbf{H}^\star \circ \mathbf{H}^\star \circ \mathbf{H}^\star - (\lambda_1 + \lambda_2)\mathbf{H}^\star \\ & +4\lambda_2\mathbf{H}^\star \circ \mathbf{H}^\star \circ \mathbf{H}^\star - 4\lambda_2\mathbf{H}^\star \circ \mathbf{H}^\star \circ \mathbf{H}^\star\Big)_{ij} h_{ij}^\star \ .\end{aligned}$$

These two equations lead to the updating rules (4), (5). Our next result proves that these rules improve the objective value.

Theorem 1. *The objective function* $f(\mathbf{W},\mathbf{H})$ *of program* (3) *is non-increasing under update rules* (4) *and* (5).

The proof of Theorem 1 is given in He et al. (2015a). Our overall approach is described as Algorithm 1. $\mathbf{W}$ and $\mathbf{H}$ are initialised using plain NMF (Lee and Seung 1999; 2001), then normalised (Zhang et al. 2007; 2010).

Algorithm 1 NMF-Guttman

```
Input:
    $\mathbf{V}$, $\mathbf{H}_{ideal}$, $\lambda_0$, $\lambda_1$, $\lambda_2$, $k$;
Output:
    A topic-student matrix, $\mathbf{H}$;
Initialise $\mathbf{W}$, $\mathbf{H}$ using NMF;
Normalise $\mathbf{W}$, $\mathbf{H}$ following (Zhang et al. 2007; 2010);
repeat
    Update $\mathbf{W}, \mathbf{H}$ iteratively based on Eq. (4) and Eq. (5);
until converged
return $\mathbf{H}$;
```

Selection of $\mathbf{H}_{ideal}$ Topic-student matrix $\mathbf{H}_{ideal}$ is an ideal target where students' topic responses conform to a perfect Guttman scale. $\mathbf{H}_{ideal}$ can be obtained in different ways depending on the attribute of interest to be measured. In this paper, we are interested in measuring students' latent skill in MOOCs. We envision measurement at the completion of a first offering, with scaled items applied in subsequent offerings for measuring students or curriculum design;

alternatively within one offering after a mid-term. Thus, $\mathbf{H}_{ideal}$ can be obtained using assessment, *which need not be based on Guttman-scaled items*. For each student j, his/her responses to the topics given by column $(h_{ideal})_{\cdot j}$ are selected based on his/her grade $g_j \in [0, 100]$, as

$$(h_{ideal})_{\cdot j} = (\underbrace{1 \cdots 1}_{b} \underbrace{0 \cdots 0}_{k-b})$$

$$\text{where} \qquad b = \min \left\{ \left\lfloor \frac{g_j + width}{width} \right\rfloor, k \right\}, width = \frac{100}{k} \quad .$$

For example, student j with $g_j = 35$ has response pattern on $k = 10$ topics $(h_{ideal})_{\cdot j} = (1111000000)$.

Experiments

We conduct experiments to evaluate the effectiveness of our algorithms on real MOOCs on Coursera. We also demonstrate the robustness of our approach in terms of parameter sensitivity. In our experiments, we use the first offerings of three Coursera MOOCs from Education, Economics and Computer Science offered by The University of Melbourne. They are *Assessment and Teaching of 21st Century Skills*, *Principles of Macroeconomics*, *Discrete Optimisation* and are named EDU, ECON and OPT for short respectively.

Dataset Preparation

We focus on the students who contributed posts or comments in forums. For each student, we aggregate all posts/comments that s/he contributed. After stemming, removing stop words and html tags, a word-student matrix with normalised tf-idf in [0,1] is produced. The statistics of words and students for the MOOCs are displayed in Table 3.

Table 3: Statistics of Datasets

MOOC	#Words	#Students
EDU	20,126	1,749
ECON	22,707	1,551
OPT	17,059	1,092

Baseline Approach and Evaluation Metrics

Since there has been no prior method to automatically generate topics forming a Guttman scale, we compare our algorithm with standard NMF (with no regularisation on $\mathbf{H}_{ideal}$).

We adopt the Coefficient of Reproducibility (CR) as it is commonly used to evaluate Guttman scale quality:

$$CR = 1 - \frac{\text{No. of errors}}{\text{No. of possible errors(Total responses)}} \quad .$$

CR measures how well a student's responses can be predicted given his/her position on the scale, *i.e.*, total score. By convention, a scale is accepted with items scaled unidimensionally, if its CR is at least 0.90 (Guttman 1950).

To guarantee binary $\mathbf{H}$, we first scale to $\frac{h_{ij} - \min(\mathbf{H})}{\max(\mathbf{H}) - \min(\mathbf{H})} \in [0, 1]$, then threshold against a value in $[0.1, 0.2, \cdots, 0.9]$ maximising CR, so that we *conservatively* report CR.

Experimental Setup

Evaluation Setting We split data into a training set (70% students) and a test set (30% students). The topics are generated by optimising the objective function (2) on the training set, and evaluated using CR and the quality of approximation $\|\mathbf{V} - \mathbf{WH}\|_F^2$. To simulate the inferring responses for new students, which has not been explored previously, the trained model is evaluated on the test set using Precision-Recall and ROC curves. Note that in the psychometric literature, validation typically ends with an accepted (> 0.9) CR on the training set.

After learning on the training set word-student matrix $\mathbf{V}^{(train)}$, two matrices are produced: a word-topic matrix $\mathbf{W}^{(train)}$ and topic-student matrix $\mathbf{H}^{(train)}$. To evaluate the trained model on the test set $\mathbf{V}^{(test)}$, we apply the trained word-topic matrix $\mathbf{W}^{(train)}$. Together, we have the relations

$$\begin{aligned} \mathbf{V}^{(train)} &= \mathbf{W}^{*(train)}\mathbf{H}^{(train)} \\ \mathbf{V}^{(test)} &= \mathbf{W}^{*(train)}\mathbf{H}^{(test)} \quad . \end{aligned}$$

Solving for $\mathbf{H}^{(test)}$ yields

$$\mathbf{H}^{(test)} = \mathbf{H}^{(train)}(\mathbf{V}^{(train)})^{\dagger}\mathbf{V}^{(test)} \quad .$$

where $(\mathbf{V}^{(train)})^{\dagger}$ denotes the pseudoinverse of $\mathbf{V}^{(train)}$.

Hyperparameter Settings Table 4 shows the parameter values used for parameter sensitivity experiments, where the default values in boldface are used in other experiments.

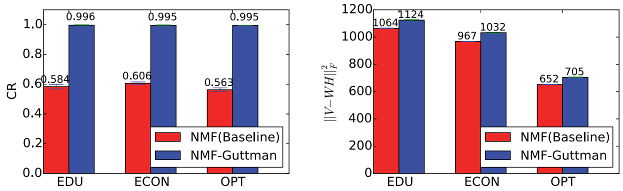
Table 4: Hyperparameter Settings

Parameter	Values Explored (**Default Value**)
λ_0	$[10^{-4}, 10^{-3}, 10^{-2}, \mathbf{10^{-1}}, 10^{0}, 10^{1}, 10^{2}]$
λ_1	$[10^{-4}, 10^{-3}, 10^{-2}, \mathbf{10^{-1}}, 10^{0}, 10^{1}, 10^{2}]$
λ_2	$[10^{-4}, 10^{-3}, \mathbf{10^{-2}}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}]$
k	$[5, \mathbf{10}, 15, 20, 25, 30]$

Results

In this group of experiments, we examine how well the generated topics conform to a Guttman scale, and the quality of approximation $\mathbf{WH}$ to $\mathbf{V}$. The reported results are the results averaged over 10 runs. The parameters are set using the values in boldface in Table 4. Figure 1 displays the comparison between our algorithm NMF-Guttman and the baseline NMF in terms of CR, and the quality of approximation $\mathbf{WH}$ to $\mathbf{V}$ on the training set.

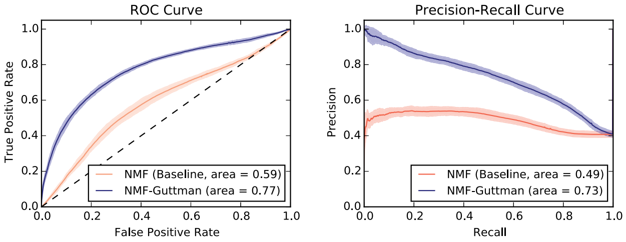
It is clear that our algorithm NMF-Guttman can provide excellent performance in terms of CR with nearly a perfect 1.0, well above the 0.9 cutoff for acceptance. This significantly outperforms baseline which has 0.60 CR across the MOOCs, below Guttman scale acceptance. Meanwhile, NMF-Guttman maintains good quality of approximation, with only slightly inferior $\|\mathbf{V} - \mathbf{WH}\|_F^2$ comparing to NMF (5%, 6%, 8% worse on EDU, ECON, OPT). This is reasonable, as NMF-Guttman has more constraints hence the model itself is less likely to approximate $\mathbf{V}$ as well as the less constrained standard NMF.



(a) CR (b) $\|\mathbf{V} - \mathbf{WH}\|_F^2$

Figure 1: Comparison of NMF and NMF-Guttman in terms of CR and $\|\mathbf{V} - \mathbf{WH}\|_F^2$.

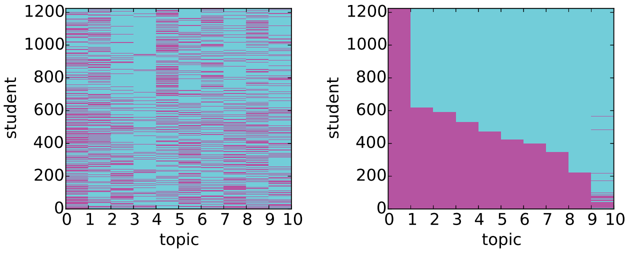
The ROC and Precision-Recall curves (averaged curves with standard deviation over 10 runs) on test set for the ECON MOOC are shown in Figure 2. It can be seen that NMF-Guttman significantly dominates NMF, with around 20%-30% better performance, demonstrating the possibility of using the topics for inferring the response of unseen students. Similar results can be found on the remaining MOOCs in He et al. (2015a).



(a) ECON (b) ECON

Figure 2: Comparison of NMF and NMF-Guttman in terms of ROC curve and Precision-Recall curve.

We next visualise the student-topic matrix $\mathbf{H}^T$ produced by NMF and NMF-Guttman respectively. Figure 3 is a clear demonstration that NMF-Guttman can produce excellent Guttman scales, while NMF may not. Around half of the cohort (having grade=0) only contribute to topic 1—the easiest—while only a few students contribute to topic 10—the most difficult.



(a) NMF (b) NMF-Guttman

Figure 3: Student-topic matrix generated by NMF and NMF-Guttman for MOOC EDU; fuchsia for 1, cyan for 0.

NMF-Guttman can discover items (topics) with responses conforming to a Guttman scale while maintaining the quality of factorisation approximation. It also effectively infers new students' responses.

Validity

The results above establish that our algorithm generates items (topics) with responses conforming to the Guttman scale. Next we test validity—whether topics are meaningful in two aspects: **a) Interpretability**: Are the topics interpretable? **b) Difficulty level**: Do topics exhibit different difficulty levels as inferred by our algorithm and implied by the Guttman scale?

Qualitative Survey To answer the above questions, we interviewed experts with relevant course background. We showed them the topics (each topic is represented by top 10 words) discovered by our algorithm NMF-Guttman and those generated from the baseline NMF, while blinding interviewees to the topic set's source. We randomised topic orders, and algorithm order, since our algorithm naturally suggests topic order. We posed the following questions for the topics from NMF-Guttman and NMF respectively:

Q1. *Interpretation:* interpret the topic's meaning based on its top 10 word description.

Q2. *Interpretability:* how easy is interpretation? 1=Very difficult; 2=Difficult; 3=Neutral; 4=Easy; 5=Very easy.

Q3. *Difficulty level:* how difficult is the topic to learn? 1=Very easy; 2=Easy; 3=Neutral; 4=Difficult; 5=Very difficult.

Q4. *Ranking:* rank the topics according to their difficulty levels. From 1=easiest; to 10=most difficult.

a) OPT MOOC We interviewed 5 experts with PhDs in discrete optimisation for the OPT MOOC. To validate the topics' difficulty levels, we compute the Spearman's rank correlation coefficient between the ranking from our algorithm and the one from each interviewee, which is shown in Table 5. There is high correlation between the NMF-Guttman ranking and those of the Interviewees, *suggesting the topics' Guttman scale relates to difficulty.*

Table 5: Survey for OPT MOOC.

Interviewee	Background	Spearman's rank correlation coefficient
1	Works in optimisation and took OPT MOOC	0.71
2	Tutor for OPT MOOC	0.37
3	Professor who teaches optimisation courses	0.90
4	Works in optimisation	0.67
5	Works in optimisation	0.41

Table 6 depicts the interpretation on a selection of four topics by Interviewee 1, who took the OPT MOOC previously. The compelete interpretation for the topics from NMF and NMF-Guttman can be found in He et al. (2015a). It can be seen that the topics from NMF-Guttman are interpretable and exhibit different difficulty levels, qualitatively

Table 6: Interviewee 1's interpretation on OPT MOOC topics generated from NMF-Guttman with inferred difficulty ranking.

No.	Topics	Interpretation	Inferred Ranking
1	python problem file solver assign pi class video course use	How to use platform/python	1 (Easiest)
2	submit thank please pyc grade feedback run solution check object	Platform/submission issues	2
5	color opt random search local greedy swap node good get	Understand and implement local search	5
8	time temperature sa move opt would like well start ls	How to design and tune simulated annealing and local search	8

validating the topics can be used to measure students' latent skill. Note that the topics produced by NMF do not conform to a Guttman scale and are not designed for measurement. Indeed we observed informally that NMF-Guttman's topics were more diverse than those of NMF. For OPT MOOC, half of the topics are not relevant to the course content directly, *i.e.*, feedback about the course and platform/submission issues. While most of the topics from NMF-Guttman are closely relevant to the course content, which are more useful to measure students' skill or conduct curriculum refinement.

b) EDU MOOC The course coordinator who has detailed understanding of the course, its curriculum and its forums, was interviewed to answer our survey questions. A 0.8 Spearman's rank correlation coefficient is found between the NMF-Guttman ranking and that of the course coordinator, supporting that the inferred difficulty levels are meaningful. Furthermore, most of the NMF-Guttman's topics are interpretable, less fuzzy, and less overlapping than those of NMF, ad judged by the course coordinator. The topic interpretations can be found in He et al. (2015a).

Parameter Sensitivity

To validate the robustness of parameters and analyse the effect of the parameters, a group of experiments were conducted. The parameter settings are shown in Table 4. Due to space limitation, we only report the results for λ_1 on OPT MOOC. Results for parameter λ_0,λ_1,λ_2 and k on all three MOOCs can be found in He et al. (2015a).

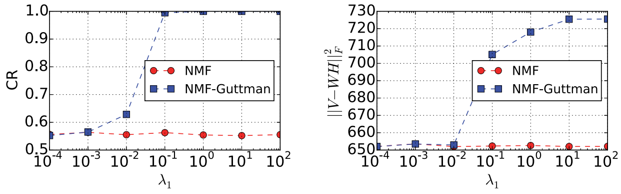
Regularisation Parameter λ_1 The performance of CR and $\|\mathbf{V} - \mathbf{WH}\|_F^2$ with varying λ_1 is shown in Figure 4. It can be seen that NMF-Guttman's high performance is stable for λ_1 varying over a wide range 10^{-1} to 10^2.

Similar results are found for λ_0, λ_2, and k. NMF-Guttman is not sensitive to λ_0 and k. For λ_2, NMF-Guttman stably performs well when λ_2 varies from 10^{-4} to 10^{-2}.

Overall, our algorithm NMF-Guttman is robust, consistently achieves much higher CR than NMF with varying λ_0, λ_1, λ_2 and k, while maintaining the quality of approximation $\|\mathbf{V} - \mathbf{WH}\|_F^2$.

Conclusion

This is the first study that combines a machine learning technique (topic modelling) with measurement theory (psychometrics) as used in education. Our focus is measurement for curriculum design and assessment in MOOCs. Motivated by findings that participation level in online forums is predictive of student performance (Beaudoin 2002), we aim to automatically discover forum post topics on which student engagement forms a so-called Guttman scale: such scales are evidence of measuring educationally-meaningful skill attainment (Guttman 1950). We achieve this goal by a novel regularisation of non-negative matrix factorisation.



(a) CR on OPT

(b) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on OPT

Figure 4: Comparison of NMF and NMF-Guttman in terms of CR and $\|\mathbf{V} - \mathbf{WH}\|_F^2$ with varying λ_1.

Our empirical results are compelling and extensive. We contribute a quantitative validation on three Coursera MOOCs, demonstrating our algorithm conforms to Guttman scaling (shown with high coefficients of reproducibility), strong quality of factorisation approximation, and predictive power on unseen students (via ROC and PR curve analysis). We also contribute a qualitative study of domain expert interpretations on two MOOCs, showing that most of the topics with difficulty levels inferred, are interpretable and meaningful.

This paper opens a number of exciting directions for further research. Broadly speaking, the consequences of content-based measurement on educational theories and practice requires further understanding, while the study of statistical models for psychometrics by computer science will stimulate interesting new machine learning.

Our approach could be extended to incorporate partial prior knowledge. For example, an education researcher or instructor might already possess certain items for student engagement in MOOCs (*e.g.*, watching videos, clickstream observations, completing assignments, etc.) to measure some latent attribute. We are interested in exploring how to discover topics that measure the same attribute as measured by existing items.

References

Abdi, H. 2010. Guttman scaling. In Salkind, N. J., ed., *Encyclopedia of Research Design*. SAGE Publications.

Anderson, A.; Huttenlocher, D.; Kleinberg, J.; and Leskovec, J. 2014. Engaging with massive online courses.

In *Proceedings of the 23rd International Conference on World Wide Web*, 687–698.

Apaza, R. G.; Cervantes, E. V.; Quispe, L. C.; and Luna, J. O. 2014. Online courses recommendation based on LDA. In *1st Symposium on Information Management and Big Data*, 42.

Beaudoin, M. F. 2002. Learning or lurking?: Tracking the "invisible" online student. *The Internet and Higher Education* 5(2):147–155.

De Ayala, R. J. 2013. *Theory and practice of item response theory*. Guilford Publications.

Dıez, J.; Luaces, O.; Alonso-Betanzos, A.; Troncoso, A.; and Bahamonde, A. 2013. Peer assessment in MOOCs using preference learning via matrix factorization. In *NIPS Workshop on Data Driven Education*.

Gillani, N.; Eynon, R.; Osborne, M.; Hjorth, I.; and Roberts, S. 2014. Communication communities in MOOCs. *arXiv preprint arXiv:1403.4640*.

Guttman, L. 1950. The basis for scalogram analysis. In Stouffer, S., ed., *Measurement and Prediction: The American Soldier*. Wiley, New York.

Halawa, S.; Greene, D.; and Mitchell, J. 2014. Dropout prediction in MOOCs using learner activity features. *Experiences and best practices in and around MOOCs* 7.

He, J.; Bailey, J.; Rubinstein, B. I.; Zhang, R.; Milligan, S.; and Chan, J. 2015a. MOOCs meet measurement theory: A topic-modelling approach. Technical Report arXiv:1511.07961 [cs.LG], arXiv.

He, J.; Bailey, J.; Rubinstein, B. I.; and Zhang, R. 2015b. Identifying at-risk students in massive open online courses. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Kizilcec, R. F.; Piech, C.; and Schneider, E. 2013. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, 170–179. ACM.

Kloft, M.; Stiehler, F.; Zheng, Z.; and Pinkwart, N. 2014. Predicting MOOC dropout over weeks using machine learning methods. *EMNLP 2014* 60.

Lee, D. D., and Seung, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788–791.

Lee, D. D., and Seung, H. S. 2001. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, 556–562.

Mi, F., and Yeung, D.-Y. 2015. Probabilistic graphical models for boosting cardinal and ordinal peer grading in MOOCs. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Milligan, S. 2015. Crowd-sourced learning in MOOCs: learning analytics meets measurement theory. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, 151–155. ACM.

Pedhazur, E. J., and Schmelkin, L. P. 1991. *Measurement, design, and analysis: An integrated approach (student ed.)*. Lawrence Erlbaum Associates, Inc.

Piech, C.; Huang, J.; Chen, Z.; Do, C.; Ng, A.; and Koller, D. 2013. Tuned models of peer assessment in MOOCs. *arXiv preprint arXiv:1307.2579*.

Ramesh, A.; Goldwasser, D.; Huang, B.; Daume III, H.; and Getoor, L. 2014a. Understanding MOOC discussion forums using seeded lda. In *Proc. of 9th Workshop on Innovative Use of NLP for Building Educational Applications*, 28–33.

Ramesh, A.; Goldwasser, D.; Huang, B.; Daume III, H.; and Getoor, L. 2014b. Learning latent engagement patterns of students in online courses. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Ramesh, A.; Goldwasser, D.; Huang, B.; Daume III, H.; and Getoor, L. 2014c. Uncovering hidden engagement patterns for predicting learner performance in MOOCs. In *Proceedings of the first ACM conference on Learning@ scale conference*, 157–158. ACM.

Ramesh, A.; Kumar, S. H.; Foulds, J.; and Getoor, L. 2015. Weakly supervised models of aspect-sentiment for online course discussion forums. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Robinson, A. C. 2015. Exploring class discussions from a massive open online course (MOOC) on cartography. In *Modern Trends in Cartography*. Springer. 173–182.

Wen, M.; Yang, D.; and Rose, C. 2014. Sentiment analysis in MOOC discussion forums: What does it tell us? In *Educational Data Mining 2014*.

Xu, B., and Yang, D. 2015. Study partners recommendation for xMOOCs learners. *Computational Intelligence and Neuroscience* 2015.

Yang, D.; Adamson, D.; and Rosé, C. P. 2014. Question recommendation with constraints for massive open online courses. In *Proceedings of the 8th ACM Conference on Recommender systems*, 49–56. ACM.

Yang, D.; Sinha, T.; Adamson, D.; and Rose, C. P. 2013. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-Driven Education Workshop*, volume 11, 14.

Yang, D.; Wen, M.; Howley, I.; Kraut, R.; and Rose, C. 2015. Exploring the effect of confusion in discussion forums of massive open online courses. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, 121–130. ACM.

Zhang, Z.; Ding, C.; Li, T.; and Zhang, X. 2007. Binary matrix factorization with applications. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, 391–400. IEEE.

Zhang, Z.-Y.; Li, T.; Ding, C.; Ren, X.-W.; and Zhang, X.-S. 2010. Binary matrix factorization for analyzing gene expression data. *Data Mining and Knowledge Discovery* 20(1):28–52.