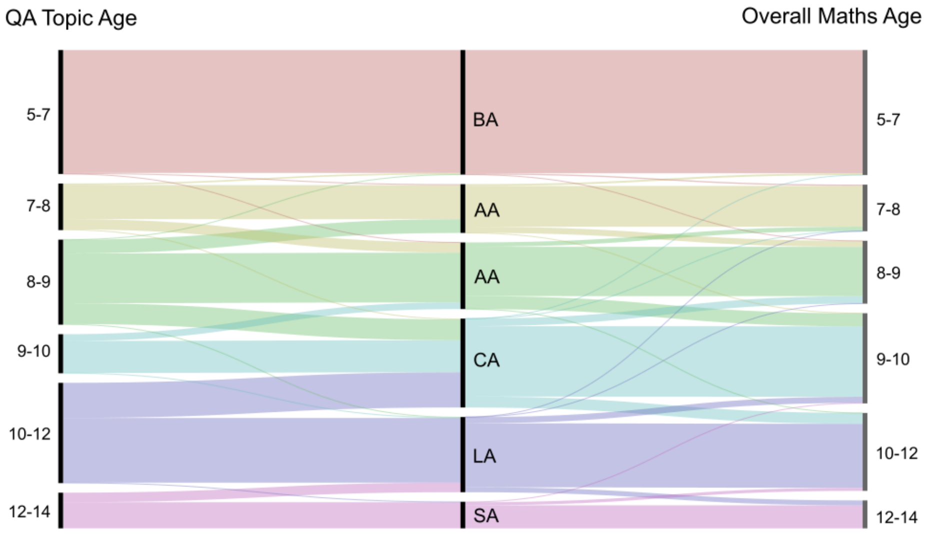
EPSRC
Engineering and Physical Sciences
Research Council

InFoMM
Industrially Focused
Mathematical Modelling

EPSRC Centre for Doctoral Training in Industrially Focused Mathematical Modelling



Designing Subsampling Strategies

Rodrigo Leal Cervantes

UNIVERSITY OF
OXFORD

whizz
EDUCATION

Contents

1. Introduction

Background

Maths-Whizz, the flagship product of Whizz Education, is a virtual tutor that teaches Mathematics to children aged 5 to 13. The tutor strives to replicate a human tutor as closely as possible, selecting appropriate subjects to cover based on the strengths and weaknesses of each child.

The Maths-Age is the metric used to measure the level of competence in a given mathematical topic.

When students sign up for the service, they undergo an assessment that measures their skill in different mathematical topics. For each topic assessed, a student gets a Maths-Age—a metric developed by Whizz similar to the Reading Age. The resolution of the Maths-Age is one quarter of a year.

In the current assessment, students are given between 3 and 10 topics out of a total of 13, with the more advanced students needing to take the most topics. The initial assessment was designed to take between 20 and 50 minutes but in reality 50 minutes is the average length. The assessment is too long for children, and it impacts the user experience negatively. This is why Whizz has asked us to look at subsampling strategies that could make the assessment shorter.

Subsampling strategies

A subsampling strategy involves assessing the student in fewer topics and then using the information that was acquired to estimate a likely age for the topics left out. Whizz is already using such a strategy: indeed, at the end of the assessment, topics that are in range in the curriculum but that were not assessed are assigned the overall Maths-Age (rounded down to the nearest quarter).

In order to clearly define the subsampling problem, we categorise the students into six Maths-Age ranges which comprise different topics. For example, a student that lies within the 5-7 range needs to be assessed in four topics and a student in the 10-12 range needs eight topics.

However, it is important to mention that Whizz is not classifying the students in any way, nor is it distinguishing them in any way. What we call the Maths-Age ranges come from the overall Maths-Age of the students, which also controls the pattern of topics that they encountered in the assessment. We insist on having these ranges because they are crucial for the clear definition of what is the subsampling problem; in particular, they define what are the topics left unassessed that still need to be predicted.

In order to successfully use subsampling in the assessment, we address two different problems. Firstly, assuming that the Maths-Age ranges of the students are known, we need to find a suitable way of predicting the information left unassessed. Secondly, we have to determine how to classify the students quickly into those ranges.

In the next section we discuss reconstruction methods and in Section 3 we give a solution to the classification problem. The solution to the two problems together forms a subsampling strategy. In Section 4 we discuss the newly proposed model and we give suggestions as to how it could be implemented. Finally, in Section 5, we address the potential impact of our research.

Key assumption in our approach

- We assume that all the topics are independent of each other and the result in one does not affect the result in another.
- We also suppose the data is the ground truth for the purpose of measuring the accuracy of our strategies. This in turn means that that the current assessment gives an accurate reflection of the skills of any child.

Glossary of terms

We define some of the terms that we will encounter several times in the report.

- **Maths-Age:** The metric used to measure the skill of a student in different mathematical topics.
- **Age matrix:** A grid; the mathematical object that we use to store the information of the assessment. The rows are topics and the columns are students.
- **Matrix completion:** The problem of predicting missing values inside a matrix.
- **Budget:** The number k of topics that we are willing to assess per student.
- **Nonnegative matrix:** A matrix that has all of its entries greater than or equal to zero, such as the Age matrices.
- **Partial Least Squares regression:** A linear regression model that predicts the remaining rows of a matrix using a model built from k rows (where k is the budget).

2. Prediction of unassessed data

We were given the assessment scores and the age of 32,771 UK students. From this data, we fill age matrices that have topics along the rows and students along the columns. We can see the age matrices of the six different ranges in Figure 1. The age scores are indicated by the colour. The codes AA, BA, QA, etc., are used internally by Whizz to refer to the topics (for example Place Value or Measures).

In a real assessment, we need to determine the Maths-Age range of each student as we decide the sequence of topics that will be most predictive: classification and prediction must be performed together. However, to simplify matters, we first assume that the ranges are known and we consider how to recover information that we purposefully leave aside.

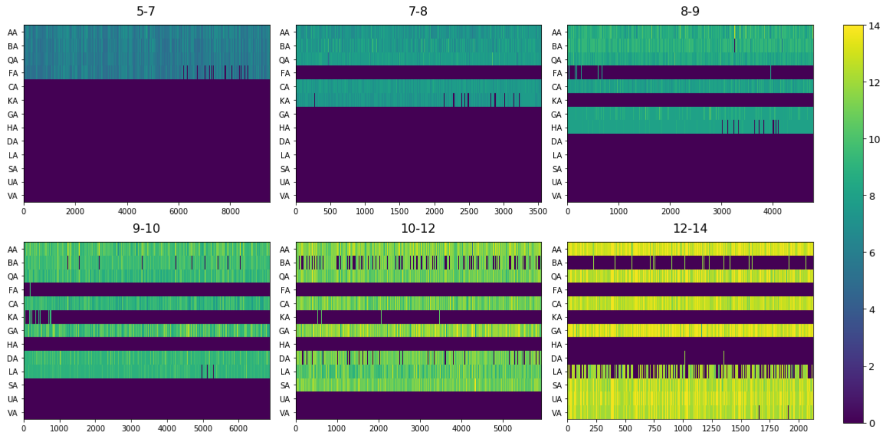


Figure 1 – The age matrices for the six Maths-Age ranges. The rows are the topics and the columns are students. The score is indicated by the colour: light blue is an age of 5, yellow is an age of 14 and dark blue is missing information.

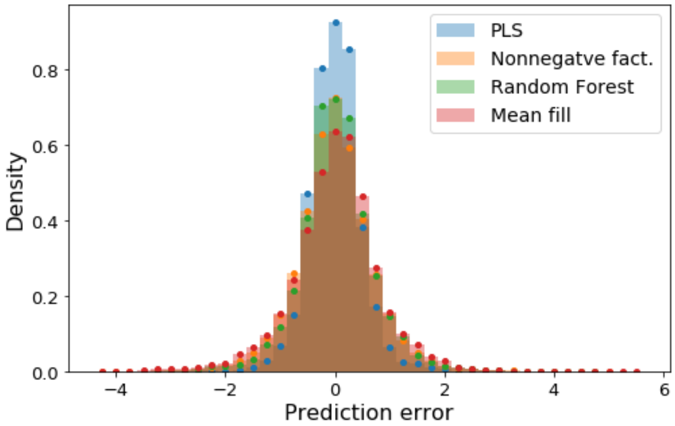


Figure 2 – Histogram of the prediction error made using different algorithms. The error is calculated as the predicted age minus the actual age as found in Whizz' data, therefore values to the right are overestimations and to the left are underestimations. The taller and thinner the distribution, the more accurate the method.

We perform an experiment for the students in the Maths-Age range 10-12. There are 8 topics and we leave 4 of them unassessed.

Let us suppose that we are interested in the range where the overall Maths age lies between the ages 10 and 12: there are 8 topics in this range. In principle we could make the assessment half as long if we were to leave 4 topics unassessed. We explore whether it is better to leave out information selected randomly or in a structured way.

If we use random subsampling, we select four entries in every column of the matrix and we remove the information. This is similar to the Netflix recommender problem where any given user has only seen and ranked a small fraction of all the movies available, and Netflix uses those scores and those of other users to predict the likely movie ratings for unseen movies.

The Netflix data and the age matrices are example of nonnegative matrices that have all of their entries greater or equal than zero. For random subsampling, we tested two algorithms to recover the missing information: Nonnegative Matrix Factorisation and Random Forest regression.

If we do structured subsampling instead of random subsampling, we leave whole topics unassessed. We then need to recover those topics in the form of whole rows of the age matrices. In our example, we could use the first 4 rows to predict the Maths-Ages in the last 4 rows, or we could use every odd row to predict all of the even rows. A method that solves the problem in this setting is called Partial Least Squares (PLS) regression.

The methods are compared in Figure 2, where we see a histogram of the prediction errors when there are 8 topics in total and we leave 4 of them unassessed. The bar above zero gives the frequency (essentially the number) of perfectly recovered entries, thus the taller and thinner a distribution is, the more accurate the method is. From this comparison we find visually that structured subsampling, with PLS regression as the method for predicting the missing data, gives the most accurate results.

Partial Least Squares (PLS) gives the best results when predicting the unassessed data.

We summarise the prediction results averaged over several experiments in Table 1. Using this table, we reach the same conclusion found visually: PLS outperforms the other two methods. Moreover, we find that a large majority of the predictions fall within half a year of the original measure.

It is also relevant to compare the results obtained using PLS with the baseline of simply averaging the four known topics to fill the four missing entries (included in last row of the table under 'Mean Fill'). This approach is currently used by Whizz for the unassessed topics that are in range in the curriculum, and we find that the PLS completion greatly outperforms this baseline.

Algorithm	Exact (%)	Within .25 (%)	Within .50 (%)	Underest. (%)	Overest. (%)	RMSE
Nonnegative Fact.	18.4	49.6	70.4	42.1	39.6	0.69
Random Forest	17.0	51.4	73.6	41.6	41.4	0.66
PLS	23.9	64.6	86.0	37.6	38.5	0.47
Mean Fill	16.2	45.2	66.0	40.2	43.6	0.79

Table 1 – Summary of the error prediction results for the main recovery methods. The first three columns give the percentages of predicted ages that fall within a range of the actual data, the next columns give the percentage of underestimated and overestimated entries and the last column shows the Root Mean Square Error, the average overall error made with the predictions.

PLS emerges from this analysis as the best choice of method to use for predicting the missing information. Furthermore, because it uses whole rows to predict other whole rows, we can rank the topics in order, from the most informative to the least informative. We do it in the following way: we make a prediction using one topic only and we find the minimum error, then we use two topics and we find the pair that gives the minimum error, and we keep repeating until we have no more topics to predict. This order will appear again in the next section.

3. Using PLS as part of a subsampling strategy

Having found the most useful way of recovering the missing information, we have to figure out a practical way in which this would be applied. The main problem that we face is to determine as quickly as possible the Maths-Age range within which a student belongs.

Classification of students into the correct Maths-Age range

Ideally we would like to present the topics to the students in the order suggested by the PLS results but this order is different for different Maths-Age ranges. The actual age of the students, which is available before starting the assessment, is of little use: indeed, the correlation between the actual age and the overall Maths-Age which measures the mathematical skills of the children is (perhaps unsurprisingly) low. Therefore, it is important that the first topic assessed is highly discriminatory between Maths-Age ranges.

Only two topics are given to every student, topics AA and QA. If we compare their scores with the overall Maths-Age ranges, they give respectively 66.2 and 73.8% correct classifications.

We determine the Maths-Age of the students using two topics, chosen to maximise the classification rate.

This indicates that QA is more discriminatory and we use it as the first topic in the assessment. From then on, we have a provisional classification and we can assess topics that are present for different ranges. Next, we choose the second topic amongst the 12 topics that are left to maximise the number of correct classifications. Using two topics we can drive the number up to 86.1% correct classifications. The alluvial diagram that serves as a cover for this report summarises this classification, with the students classified on the left according to their age in QA, in the middle using an average that includes the second topic, and on the right according to the overall Math-Age. After the two topics used for classification, we use the PLS ordering of the topics so as to optimise the prediction accuracy.

The subsampling strategy

We now have all the pieces in place for subsampling strategy: we use the first two topics to determine the Maths-Age range and once we have this information, we know what is the best topic ordering to use in order to get the minimal error for the predictions. This strategy can be can be most easily understood in the form of a decision tree, which we find in Figure 3.

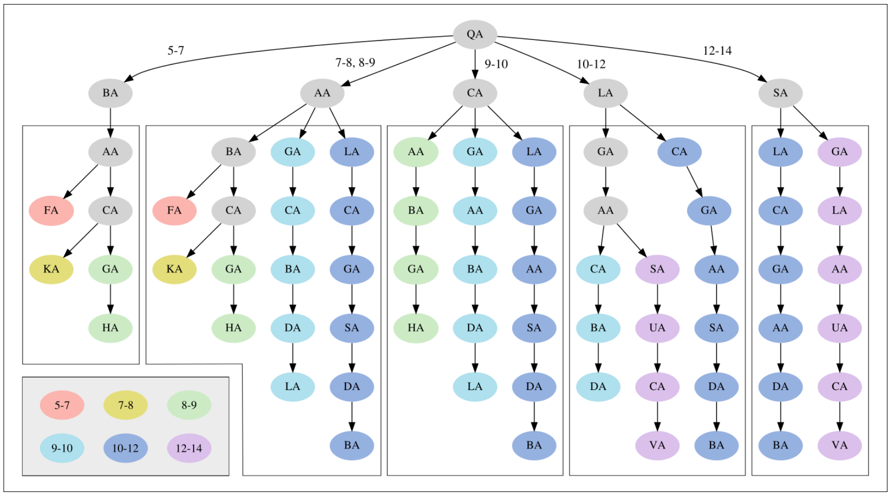

QA
5-7
7-8, 8-9
9-10
10-12
12-14
BA AA CA LA SA
AA BA GA LA AA GA LA GA CA LA GA
FA CA FA CA CA CA BA AA GA AA GA CA LA
KA GA KA GA BA GA GA BA AA CA SA AA GA AA
HA HA DA SA HA DA SA BA UA SA AA UA
LA DA LA DA DA CA DA DA CA
BA BA VA BA BA VA
5-7 7-8 8-9
9-10 10-12 12-14

Figure 3 – The subsampling strategy as a decision tree. The first topic to be assessed is QA at the top and to decide the path that we follow at a given point we use the average of all the assessment undertaken up to that point. Grey denotes a Maths-Age range that still needs to be determined, and the other colours indicate the final classification.

In order to use subsampling and reduce the time of the assessment, we assess the topics in the order given in the tree until we reach our budget k, and at this point we use a PLS model (previously trained on the data) to predict the topics left.

We report the accuracy of our subsampling strategy as a function of the budget k within any of the Maths-Age ranges. For the range of 10-12, which we have previously used in our examples, the percentages of perfectly predicted entries and within half a year are, respectively: 14.7% and 70.1% for k=1; 17.4% and 74.6% for k=2; 19.7% and 77.0% for k=3; 22.6% and 82.2% for k=4; 26.6% and 86.8% for k=5; 28.5% and 88.0% for k=6; and 27.7% and 88.1% using k=7. There are 8 topics in this range and $k = 7$ is the maximum budget.

The Maths-Age revisited

The average Maths-Age is used by Whizz after every new topic is calculated to determine the topics that should be assessed next. In Figure 4, we compare how this average converges to the final overall score when we use the current ordering of topics and when we use our newly proposed ordering. For the subset of students used to generate the image, we observe a faster convergence to the final average score, and that the standard deviation averaged over all of the students is decreased from 0.48 to 0.28.

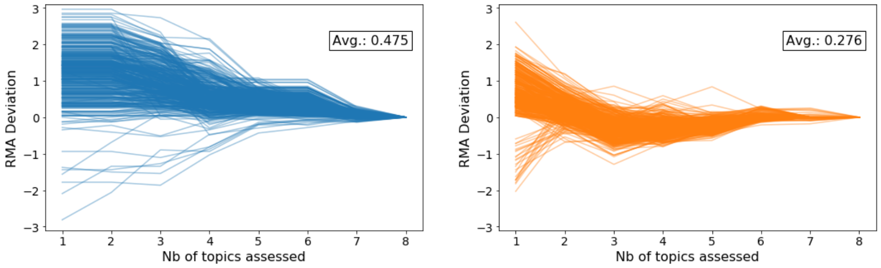


Figure 4 – Convergence of the Maths-Age to the its final value throughout the assessment. We are comparing the same data with the topics being ordered on the left according to the current ordering and on the right according to our proposed strategy. The data relates to the students whose final Maths-Age is within the 10-12 range but who have an age for topic QA outside this range (i.e. the age for that particular topic is lower than 10 years or higher than 12 years).

4. Discussion, conclusions and recommendations

Accuracy of the recovery

We have shown that prediction using PLS outperforms other matrix completion techniques, and it improved significantly upon the baseline of a simple average. We believe that this conclusion is particularly important for Whizz, because there is no reason why our proposed prediction strategy could not be used to predict topics that are currently estimated with the mean.

Another advantage of this method is that it leads naturally to a ranking of the topics based on how informative they are. During this analysis we observed that the least informative topics usually have a majority of students scoring the lowest possible age. This is probably an indication that they are given those topics for an incorrect reason.

Classification of the students

We mentioned above that we can correctly classify 86.1% of the students using only two topics. It is important to ask what happens to those students that we misclassify, and our answer to this question is that most of the misclassified students seem to have a wildly varying age profile, and they often exhibit the behaviour of scoring the lowest possible age in the most advanced topics. When this is the case, it is an indication that the students should not be assessed in the advanced topics and that they were misclassified by Whizz. Nevertheless, there is almost surely some misclassifications using our method too.

How to make the most out of our strategy?

> Improvements to the assessment will have to be done in an iterative way since we expect the data to change.

The results that we provide with our assessment methodology give the average accuracy that can be achieved for a given budget. It is for Whizz to determine the appropriate compromise between the time that is saved and the accuracy that is lost.

We are aware that our methodology might have flaws that cannot be seen in the data because it was gathered using a different assessment procedure to the one we propose. We expect future changes to the assessment to also change the data, and therefore we think that a substantial improvement can only be achieved iteratively. Metrics such as the convergence of the average Maths-Age can be used to benchmark the new changes.

5. Potential impact

Even modest improvements to the assessment can have a big effect on the business of Whizz. A shorter assessment will improve the user experience of the Maths-Whizz tutor and it could increase the conversion of trial users to a paid home plan, resulting in higher revenue.

Perhaps more importantly, we believe that our work has the potential to make teaching mathematics to young children a nicer experience.

Junaid Mubeen, Director of Education at Whizz commented: *"We are very pleased with the outcome of this work. The prospect of addressing the longstanding issue of lengthy assessment bodes well both for the user experience within Maths-Whizz, and for our home market revenue. Rodrigo's methods are clearly laid out and we are particularly encouraged by the relative ease with which his final model can be implemented. We expect to see some form of this deliverable realised in the product over the next year."*