Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Automatic Control Laboratory

SA project

Matrix Completion for Computational Education Recommendation System

TANG YU
February 16, 2023

Advisors
Dr. Alberto Padoan
Dr. Mathias Hudoba de Badyn
Matilde Gargiani
Prof. Dr. John Lygeros

Chapter 1

Abstract

This report considers the problem of designing an education recommendation system via matrix completion techniques. With the increasing amount of information, a computational recommendation system is desired in many fields, including education. We model the course structure as a bipartite graph, and the student-question performance as a binary matrix. We then employ the model proposed by [16] to perform content and learning analysis. Based on the partial observations of the binary matrix, we aim to recover a complete matrix in order to give an overall estimate of student performance. We use the nuclear norm convex optimization technique for reconstructing matrices. In the report, we discuss our modification of convex problem formulation and then show the regularization effect of this modification. The impact of varying the observation data set and changing problem dimensions are discussed. An approximate percentage of required observations in the student-question performance matrix is concluded, and the comparison of the matrix completion performance of different problem dimensions is elaborated. A detailed discussion on course structure design is presented, aiming at having a good estimate on student-question performance and maximizing the overall students' knowledge. A future work recommendation is given at the end of this report.

Contents

1 Abstract **2**

List of Figures **4**

2 Introduction **1**

3 Preliminary **3**
3.1 Nuclear norm optimization 3
3.2 Matrix coherence 4
3.3 Graph Theory 5

4 Problem Statement **6**
4.1 Model Structure 6
4.2 Model Size and Interpretation 7
4.3 Matrix Completion for Education Recommendation 8

5 Experiment **9**
5.1 Experimental Setup 9
5.2 Results and Discussion 10
5.2.1 Add regularization constraint into convex optimization 10
5.2.2 Impact of the number of incomplete observations 11
5.2.3 Impact of sparsity level of W 11
5.2.4 Impact of problem size 12
5.2.5 Advantages and Limitations 15

6 Conclusion **17**

Bibliography **17**

List of Figures

4.1 Illustration of the model structure for learning performance analysis. (a) shows the "gradebook" matrix Y of student responses to questions. This "gradebook" is usually only partially observed; (b) is a bipartite graph which depicts the connection between questions and concepts [16]. 7

5.1 Comparison of two optimization formulations: no regularization constraint vs. with regularization constraint. The comparison is in terms of the change of course structure matrix W. (a) plots the E_Z error while (b) plots the E_Y error. 10

5.2 Reconstruction error E_Z and E_Y as a function of the number of randomly observed entries. (a) small-size setup $Q = 7$, $N = 6$, $K = 3$; (b) large-size setup $Q = 10$, $N = 50$, $K = 5$. 11

5.3 The number of required observations for exact matrix recovery as a function of the number of connecting edges in the bipartite graph corresponding to W. (a) small-size setup $Q = 7$, $N = 6$, $K = 3$, with 100 independent simulations; (b) large-size setup $Q = 10$, $N = 50$, $K = 5$, with 25 independent simulations. 13

5.4 The estimated overall students' knowledge capacity $\hat{C}_{\text{sum}}$ as a function of the number of connecting edges in the bipartite graph corresponding to W. (a) small-size setup $Q = 7$, $N = 6$, $K = 3$; (b) large-size setup $Q = 10$, $N = 50$, $K = 5$. . . 14

5.5 Comparison of the estimation performance of different problem sizes in terms of reconstruction errors. The compared four different sizes are: small size $Q = 7$, $N = 6$, $K = 3$; large-size $Q = 10$, $N = 50$, $K = 5$; high-dimensional size $Q = 20$, $N = 100$, $K = 10$; higher-dimensional size $Q = 50$, $N = 200$, $K = 10$. (a) reconstruction error of $\hat{Z}$; (b) reconstruction error of $\hat{Y}$. 16

Chapter 2

Introduction

With the increasing amount of information and the arising complexity of content, the information search and gathering task becomes time-consuming for users. Thus, a recommendation system can be leveraged to provide personalized recommendations to users based on their preferences, behavior, and past interactions with a system or platform. A wide range of its applications include e-commerce, social media, e-business, e-learning, e-resource services, etc [17]. Specifically, in the field of e-learning, an effective recommendation system will provide personalized feedback and improve students' learning process. Various types of recommendation algorithms exist, such as content-based filtering, collaborative filtering, hybrid recommendation systems, matrix factorization and matrix completion, deep learning-based recommendation systems, etc [1]. The success of an education recommendation system depends on the quality of the data it collects and the fidelity of its recommendation algorithm.

Matrix completion is a technique used in recommendation systems to fill in missing values in a user-item interaction matrix. The missing values in this matrix represent instances where the user has not interacted with the item. In the context of education, a matrix completion algorithm can be used to recommend courses to students based on their past course selections. Additionally, a matrix completion algorithm can also help course designer to build an estimate of overall student performance in a course based on the incomplete student-question performance data. By using matrix completion techniques, we can fill in the missing entries in the matrix, thus obtaining a complete picture of the student-question performance data. This can assist in designing a course structure that is tailored to the needs of the students, tractable for the teacher to predict overall student performance, and is likely to improve their knowledge capability.

There are many different matrix completion algorithms that can be used for education recommendation systems. Singular Value Thresholding (SVT) aims at finding a low-rank reconstrcution matrix by solving a non-convex optimization problem. One type of SVT algorithm is called iterative hard thresholding, which iteratively computes the top several singular values of a projection matrix [13]. This algorithm converges quickly when the ground truth matrix is very low-rank and its computation time is much faster than a full Singular Value Decomposition (SVD) [24]. The other type of SVT is soft-thresholding algorithm, which replaced the singular values of the partially observed matrix with their soft-thresholded counterparts. The resulting matrix is then reconstructed using the truncated SVD [18]. Alternating Least Squares (ALS) is another space efficient technique that stores the iterates in a factored form [6]. This method reformulates the low-rank constraint as a least-square problem utilizing a matrix product formulation. Although there are some convergence guarantees established in the literature [14], [15], [3], the optimization problem is non-convex and the final solution depends significantly on the initialization of matrix factored form. On the other hand, Nuclear Norm Minimization (NNM) method refor-

mulates the low-rank constraint into a convex optimization problem. It involves minimizing the sum of singular values of the partially observed matrix over the given affine space, subject to the observed entries [22]. The minimum-rank solution can be recovered by this convex optimization method. Besides the aforementioned fundamental matrix completion techniques, many other algorithms and variants are developed for different scenarios. The work in [20] considers side information and feature selection to perform inductive matrix completion. The implementation in [9] copes with high-rank matrix completion by assuming columns of the high-rank matrix lie within a union of multiple low-rank sub-spaces. [12] and [11] modify NNM method by treating each singular value differently, and then formulate this weighted NNM into a standard quadratic programming problem. The work in [7] proposes a binary matrix completion (BMC) problem, where observations are 1-bit measurements. It shows that maximum likelihood estimate under a suitable constraint gives an accurate reconstruction of the matrix.

In this report, we employ the NNM matrix completion technique for designing education recommendation system by virtue of the efficiency and scalability of convex optimization. We summarize the contributions of this report as follows. We modify the standard NNM problem formulation by adding one extra optimization variable and one more convex constraint. The matrix completion performance with this modification surpasses the standard one in our specific scenario. Meanwhile, we give a result on the required number of observed entries to exactly complete matrix and obtain a prediction of overall student performance. In addition, we discuss how the course structure will influence the student-question response matrix recovery, and draw a suggestion on the course structure design. In the end, we demonstrate that our matrix completion technique can be leveraged for problems of different dimensions, i.e., different class sizes.

This report is structured as follows. A theoretical background is described in Section.3 to provide a collection of the related mathematical preliminaries. Section.4 gives an overview of the employed model in this report. Every required symbols are defined there and the methodologies for the subsequent section are interpreted in this part. Section.5 presents synthetic data simulation for our model. In this section, it discusses the regularization effect of our convex optimization formulation, the impact of the number of observations and the problem size. A discussion regarding the course structure design is also presented in this section. Eventually, the conclusion is given in Section.6 and we also discuss some future work in this section.

Chapter 3

Preliminary

This section serves as a review of mathematical concepts and results to be used throughout this report. The main concepts involved in the experiment are: convex optimization, matrix coherence, graph theory. As mentioned in Section. 2, there are many existing algorithms to recover a matrix. In this report, a specific technique is employed which converts matrix completion problem into a nuclear norm convex optimization problem. Meanwhile, matrix coherence is critical for our low-rank structure of the matrix, which plays an important role in matrix reconstruction. Additionally, our model as will be discussed in Section. 4 is closely related to graph theory. All these concepts will be illustrated in the following sections.

3.1 Nuclear norm optimization

For a given real-valued matrix Z, let Ω denotes the set of indices of elements that have been observed from the matrix Z. In other words, $(i,j) \in \Omega$ if the entry Z_{ij} is observed. In the meantime, we assume the locations of all observed entries are sampled uniformly at random. Then based on Ω, one wants to reconstruct the ground truth matrix Z with sufficiently small recovery error. The reconstructed matrix is denoted as $\hat{Z}$, and we assume a low-rank structure of $\hat{Z}$. One can find matrix $\hat{Z}$ by solving a non-convex optimization problem

$$\begin{array}{ll} \text{minimize} & \text{rank}(\hat{\boldsymbol{Z}}) \\ \text{subject to} & \hat{Z}_{ij} = Z_{ij} \quad (i,j) \in \Omega \end{array} \tag{3.1}$$

where $\text{rank}(\hat{Z})$ is the rank of the matrix $\hat{Z}$. However, the rank of a matrix is a very discrete and combinatorial measure of a matrix. In fact, the optimization Problem (3.1) cannot be efficiently solved by existing algorithms because it is not only NP-hard, but all existing algorithms which provide exact solutions require computation time doubly exponential in the dimension of the matrix [4].

The rank of a matrix is equal to the number of its nonzero singular values. Let $Sp(\hat{Z}) = [\sigma_1, \sigma_2, ..., \sigma_n]^T$ denote the vector of singular values of $\hat{Z}$, where $\sigma_k(\hat{Z})$ denotes the k-th largest singular value of $\hat{Z}$. Then $\text{rank}(\hat{Z}) = \|Sp(\hat{Z})\|_0$, where $\|\cdot\|_0$ simply indicates the number of nonzero elements in a vector. As proposed in [2], we approximate Problem (3.1) by replacing the ℓ_0-norm with the *trace norm*. The trace norm, also known as *nuclear norm*, of $\hat{Z}$ is the sum of its singular values (ℓ_1-norm of $Sp(\hat{Z})$). The nuclear norm is defined as follows

$$\|\hat{\boldsymbol{Z}}\|_* = \sum_{k=1}^{n} \sigma_k(\hat{\boldsymbol{Z}}). \tag{3.2}$$

The resulting nuclear norm optimization problem is

$$\begin{array}{ll} \text{minimize} & \|\hat{\boldsymbol{Z}}\|_* \\ \text{subject to} & \hat{Z}_{ij} = Z_{ij} \quad (i,j) \in \Omega \end{array} \tag{3.3}$$

Minimizing the sum of the magnitude of singular values of $\hat{Z}$ can heuristically minimize the number of non-zero singular values. The convex ℓ_1-norm is analogous to the counting ℓ_0-norm in the area of sparse signal recovery [2]. The resulting optimization Problem (3.3) is convex and can be efficiently solved via existing convex optimization solvers such as CVX, Yalmip, GAMS, etc. If the number of observed entries (cardinality of Ω) is sufficiently large and with every entry being uniformly sampled, one can hope that there is a unique low-rank matrix $\hat{Z}$ which exactly recovers Z. The remaining parts of the report will utilize the formulation (3.3) for matrix completion simulation and analysis.

3.2 Matrix coherence

There are some matrices whose entries encode different amount of information, i.e., have different importance in terms of observation). As a result, it is improbable to recover the original matrix if the important entries are missing. In mathematical terms, the eigenvectors of such matrices are too much aligned with the standard canonical basis of R^d, where d represents the dimension of a square matrix. Such property can be interpreted as matrix coherence [6]. Matrix coherence measures the degree to which the singular vectors of a matrix are correlated with the standard basis. This concept is employed to characterize the capability to obtain overall information from a subset of matrix entries in the context of low-rank approximations and other sampling-based algorithms [19]. We report the formal definition of matrix coherence:

Definition [2]: Let U be a subspace of R^d of dimension r and P_U the orthogonal projection onto U. Then the coherence of U with respect to the standard basis e_i, for $i = 1, 2..., d$, is defined as

$$\mu(U) := \frac{d}{r} \max_{i=1,..,d} \|P_U \cdot e_i\|^2. \tag{3.4}$$

Note that the largest value that $\mu(U)$ can take is $\frac{d}{r}$, which happens when $e_i \in U$ so that $\|P_U \cdot e_i\| = 1$. The smallest possible value for $\mu(U)$ is 1, which occurs when U is spanned by the vector with all element being $1/\sqrt{d}$. Matrices whose column and row spaces have a low coherence (close to 1) are called *incoherent* [6]. In general, matrices with low coherence are easier to recover because most entries of the vector have roughly the same order of magnitude and hence the information are comparably shared among almost all the entries [6]. One thing to notice is that no matter how coherent the matrix is, if there is a column/row that does not have any observed entries, it is generally impossible to recover the matrix exactly. Detailed theory regarding the conditions for exact matrix recovery is summarized below.

Theorem 1 (Exact Recovery Guarantee [2])**.** *Let N be the number of observed entries of a $p_1 \times p_2$ matrix with rank r, and let $p = \max(p_1, p_2)$. And C is a numerical constant depending on the coherence of the matrix. Then, if the location of each observation is sampled uniformly at random and $N \geq Crp^{5/4} \log p$, then there is a high probability (at least $1 - cp^{-3}$) that we recover the matrix exactly.*

Corollary 1.1. *For low matrix coherence, $N \approx rp \log p$ for exact recovery; whereas for matrix high coherence, $N \approx p^2 \log p$ for no recovery error [2].*

Corollary 1.2. *Low-coherent matrix generally possesses low-rank structure. If the matrix rank $r \leq p^{1/5}$, then we need fewer observations ($N \geq Crp^{6/5} \log p$) to exactly recover matrix with probability at least $1 - cp^{-3}$ [2].*

The theorem and its corollaries give a concrete relationship regarding the matrix dimension, matrix rank, matrix coherence, and the required number of observations. The probability bound on the exact recovery is also described. If the matrix has high coherence, it is almost intractable to exactly reconstruct the matrix unless roughly all the entries are observed. In contrast, it is possible to recover a low-coherence matrix with only observing a subset of the total entries. Generally, the lower the matrix coherence, statistically the fewer observations needed to recover the matrix.

3.3 Graph Theory

A graph is composed of vertices (nodes) V and edges E. A general graph G can be represented as $G = (V, E)$. A graph can be categorized into directed graph and undirected graph. For undirected graph, the edges E do not have directions, and the edge indicate a two-way relationship, which means each edge can be traversed in both directions. On the other hand, the edges of directed graph have a direction. The edges encode information of a one-way relationship, namely, each edge can only be traversed in a single direction.

A bipartite graph (or bigraph) is a special type of graph whose vertices can be divided into two disjoint and independent sets U and V, that is every edge connects a vertex U to one in V, often denoted as $G = (U, V, E)$. In our scenario, we only consider undirected bipartite graph. The degree of a vertex $\deg(v)$ of an undirected graph is the number of edges associated to this vertex. One property about bipartite graph is that the total sum degree of one vertices set is equal to that of the other, which also equals the cardinality of the edge set [8].

$$\sum_{v \in V} \deg(v) = \sum_{u \in U} \deg(u) = |E|. \tag{3.5}$$

Chapter 4

Problem Statement

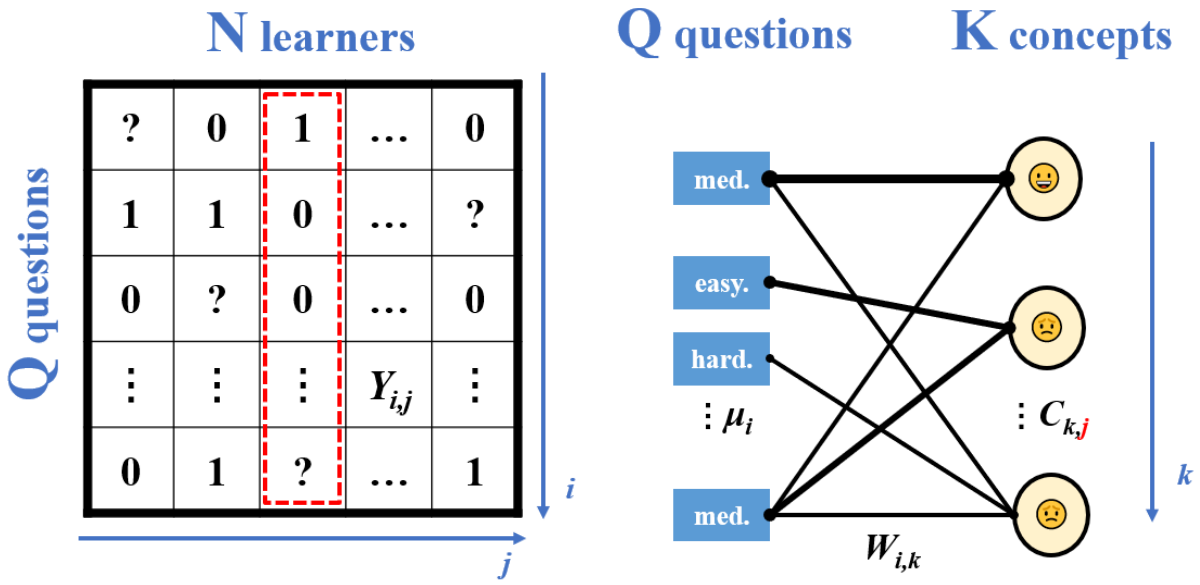
4.1 Model Structure

We employ the model developed in [16] for course structure and learning performance analysis. Specifically, we are interested in the probability that a student answers a question correctly. The correctness of an answer to a given question is influenced by three factors: the relatedness of each question to the underlying concepts, learners' knowledge towards the concepts, and each question's intrinsic difficulty. The correctness of students' responses is collected into a "gradebook" matrix Y, which is a common practice in classical test theory [21]. A graphical demonstration of such model is depicted in Figure.4.1(a), where each row represents a question in a course and each column represents a student taking this course. The Y is a binary matrix with each entry $Y_{i,j} \in \{0, 1\}$, depending on whether student j provides a correct answer to question i. Value 1 indicates a correct response while 0 a false answer. This matrix is only partially observed with question marks representing the missing values of observation data, due to the fact that the corresponding questions are not responded by or assigned to students.

Figure.4.1(b) is essentially a bipartite graph relating the questions set (rectangles) to the concepts set (circles). This undirected bipartite graph models any course structure in which questions and concepts are associated via edges in the graph. The corresponding matrix of this bipartite graph is W, where each entry $W_{i,k}$ represents one edge weight of the bipartite graph, indicating the extent to which question i involves the underlying concept k. We can also design question i with its intrinsic easiness level μ_i. Such information is encoded by matrix M, whose row i contains the intrinsic question easiness μ_i. Meanwhile, the learners' knowledge towards concepts is modeled by matrix C, where j-th learner's knowledge of the k-th concept is denoted by $C_{k,j}$. Thereby, we incorporate the information of the aforementioned three factors into matrix Z by $Z = WC + M$. The dimensions of this model will be discussed in Section.4.2. We then calculate the probabilities that students answer the questions correctly. A large value of a entry in Z should reflect a large probability of success. A common approach to obtaining probabilities is by applying sigmoid function to each entry of Z [23]. The sigmoid function is defined as

$$\Phi(x) = \frac{1}{1 + e^{-x}}. \tag{4.1}$$

Here, $\Phi(x)$ maps a real value $x \in (-\infty, \infty)$ to a success probability $\Phi(x) \in [0, 1]$, such that extremely negative entries are mapped to probability zero of success while extremely positive entries are mapped to probability one of success. Eventually, we are interested in whether learners can correctly respond to questions, and hence the following model for a binary-valued graded



(a) Student responses to questions

(b) Question-concept association graph

Figure 4.1: Illustration of the model structure for learning performance analysis. (a) shows the "gradebook" matrix Y of student responses to questions. This "gradebook" is usually only partially observed; (b) is a bipartite graph which depicts the connection between questions and concepts [16].

performance variable $Y_{i,j} \in \{0, 1\}$ is proposed:

$$\begin{aligned} Z_{i,j} &= \mathbf{W}_i^T \mathbf{C}_j + \mu_i, \quad \forall i, j, \\ Y_{i,j} &\sim \text{Ber}\left(\Phi\left(Z_{i,j}\right)\right), \quad (i,j) \in \Omega. \end{aligned} \tag{4.2}$$

Here, Ber(x) designates a Bernoulli distribution with success probability x. The set of observation $\Omega \subseteq \{1, ..., Q\} \times \{1, ..., N\}$ contains the indices which are the positions of the observed data of students' responses. Based on the incomplete observation data of the students' response matrix Y, our goal is to recover the matrix $\hat{Y}$ utilizing the structure of matrices W, C, and M.

4.2 Model Size and Interpretation

Let N denote the total number of students, Q the total number of questions, and K the total number of underlying concepts. Thereby, C is a $K \times N$ matrix and each column $C_j \in R^K$, $j \in \{1, ..., N\}$ of C can be interpreted as a measure of j-th learner's knowledge regarding all concepts, with larger $C_{k,j}$ values implying more knowledge. And W is a $Q \times K$ matrix and each row $W_i^T \in R^K$, $i \in \{1, ..., Q\}$, of W can be interpreted as a measure of involvement of question i to all concepts, with larger $W_{i,k}$ values indicating strong association between question i and concept k. Lastly, each question i has its intrinsic difficulty $\mu_i \in R$, with larger values indicating easier questions. Then, M becomes a $Q \times N$ matrix with each row $M_i = \mu_i \cdot 1_{1\times N}$, $i \in \{1, ..., Q\}$, representing the difficulty level of question i. Note that $1_{1\times N}$ is a all-ones row vector of size N because each question imposes the same intrinsic difficulty for N students. Therefore, by equality $Z = WC + M$, Z becomes a $Q \times N$ matrix, whose entry $Z_{i,j}$, $i \in \{1, ..., Q\}, j \in \{1, ..., N\}$ associated with the probability of question i is correctly answered by student j.

In our educational domain of interest, typically, there is only a small number of concepts, i.e., $K < Q, N$. As a result, W becomes a tall and narrow $Q \times K$ matrix while C becomes short and wide $K \times N$ matrix. The rank of W and C are determined by value K. Meanwhile, each question is generally only related to a subset of the whole concepts, which enforces sparse structure of W. The number of non-zero entries in W indicates the degree of the corresponding bipartite graph. Furthermore, the numerical values of each entry in W should be non-negative, while each entry of C can be either positive or negative. Large positive values of $C_{k,j}$ implies strong knowledge of learner j towards the concept k, whereas negative values reveal insufficient knowledge. Because we postulate that having strong concept knowledge should never deteriorate a student to give correct response. Thus, imposing non-negative constraint on W ensures that large positive values in C represent strong knowledge and hence make students more probable to answer questions correctly [16]. Additionally, since each row of M is formed by $\mu_i \cdot 1_{1\times N}$, the rank of M is purely decided by the number of different difficulty levels μ, which we assume is much smaller than Q and N, and even smaller than K. Overall, the educational course information is encoded in matrix Z. After applying sigmoid function to each element $Z_{i,j}$, the larger the entry value, the larger the probability of success. Note that by formulation $Z = WC + M$, a low-rank structure is naturally imposed on Z, whose rank is determined by the value of K.

4.3 Matrix Completion for Education Recommendation

Our goal is to give an estimate $\hat{Y}$ on the overall students' performance on the questions given that we have incomplete observation data $Y_{i,j}$, $(i,j) \in \Omega$, of the ground truth matrix Y. To this end, we first recover a matrix $\hat{Z}$ and then utilizing the relation $\hat{Y} \sim \text{Ber}\left(\Phi\left(\hat{Z}\right)\right)$ to obtain $\hat{Y}$. In the following simulation in Section.5.2, we employ the nuclear norm optimization technique as mentioned in Section.3.1 to complete matrix $\hat{Z}$. The way of generating synthetic data of ground truth matrices Y and Z will be elaborated in Section.5.1. Note that during course design process, we have control over the W and M matrices so that in our convex optimization formulation, we would assume W and M are as given.

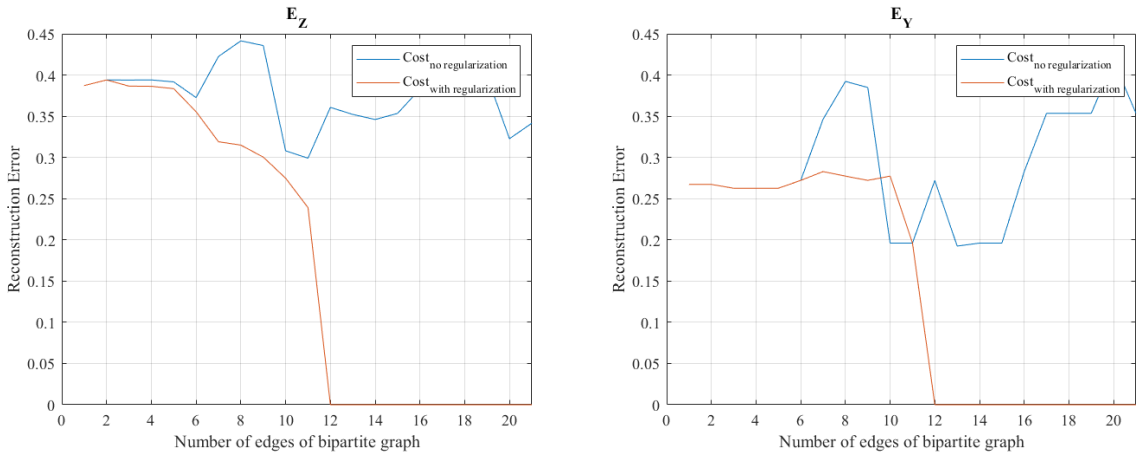
Chapter 5

Experiment

In this section, we simulate the matrix completion results via nuclear norm optimization technique. Using synthetic data, we evaluate the performance of matrix completion by comparing the fidelity of the estimates $\hat{Y}$ and $\hat{Z}$ to the ground truth Y and Z. The performance metrics used in our simulation are the relative reconstruction errors $E_Y = \|\hat{Y} - Y\|_F / \|Y\|_F$ and $E_Z = \|\hat{Z} - Z\|_F / \|Z\|_F$. The discussion on synthetic experiments is organized as follows. In Section.5.1 we specify the synthetic data generation process as well as experimental setup. In Section.5.2.1 we demonstrate the effect of adding an extra regularization constraint to the original convex optimization problem. In Section.5.2.2 we study how the percentage of i.i.d. observations will influence the matrix completion performance, in order to give an estimate on how many observations of students' grades are sufficient to give a good prediction on the overall student performance. Then we discuss the impact of varying W (course structure) on the matrix completion performance in Section.5.2.3. The result will provide a suggestion on designing course structure. Finally, we compare the performance of matrix recovery on different sizes of problem setup in Section.5.2.4.

5.1 Experimental Setup

In the simulation, we generate instances of W, C, and μ under pre-defined distributions and then generate the ground truth matrices Y and Z according to Eq.4.2. Specifically, each element of C is generated according to a Gaussian distribution $C_{k,j} \sim \mathcal{N}(0, 5)$. Meanwhile, we assume that there are only three levels of easiness for each question: easy ($\mu_e = 2$), medium ($\mu_m = 0.5$), hard ($\mu_h = -1$). The percentage of the number of questions in each easiness level is: 25% easy, 50% medium, 25% hard, which matches with pragmatic situation. The W matrix will vary among different simulations because it is related to course structure design. The detailed way of generating W will be discussed in Section.5.2.3. The data is generated independently by random number generator. Each observation entry (i, j) is sampled uniformly from the ground truth matrix. For the pragmatic consideration, we will consider both small size and large size problem settings. Small-size case corresponds to intensive mini-class teaching while large-size scenario represents standard classes. In the small size setting, we set $Q = 7$, $N = 6$, $K = 3$; while in the large size setting, we set $Q = 10$, $N = 50$, $K = 5$. Since we consider practical situation for course structure design, the cardinality of Q is usually not very large whereas the cardinality of N is not limited. To perform nuclear norm optimization, we employ the existing convex optimization solver CVX from MATLAB interface. All simulations are run with Intel(R) Core(TM) i7-10510U CPU @ 1.80GHz 2.30 GHz CPU on LAPTOP-HVTN1FK1.



(a) Reconstruction error of $\hat{Z}$

(b) Reconstruction error of $\hat{Y}$

Figure 5.1: Comparison of two optimization formulations: no regularization constraint vs. with regularization constraint. The comparison is in terms of the change of course structure matrix W. (a) plots the E_Z error while (b) plots the E_Y error.

5.2 Results and Discussion

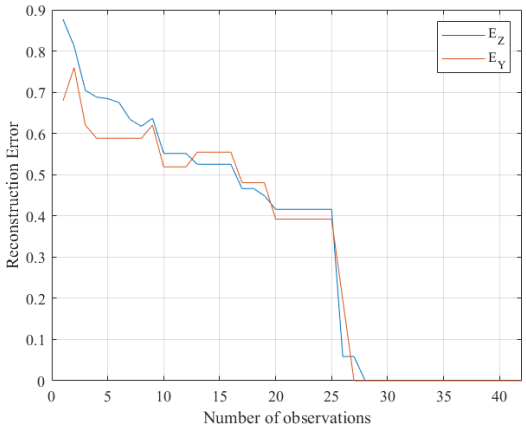
5.2.1 Add regularization constraint into convex optimization

As discussed in Section.4.2, Z is inherently low-rank. Meanwhile, Section.3.2 mentions the co-existence property of low-rank and low-coherence, and gives an approximate number on the required observations for exact recovery. In this experiment, we employ small-size case and we assume 75% of the total entries of Z matrix are observed, which approximately equals the number suggested in Section.3.2. More precise discussion regarding the required number of observations will be presented later.

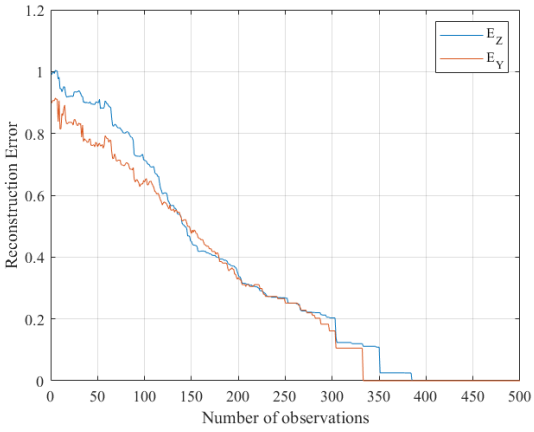
Here, we want to compare whether adding a regularization constraint (i.e., $\hat{Z} = W \cdot \hat{C} + M$) on the convex optimization problem will benefit the reconstruction. Specifically, the optimization problem with regularization constraint becomes:

$$\begin{aligned} \underset{\hat{Z},\hat{C}}{\text{minimize}} \quad & \|\hat{\boldsymbol{Z}}\|_* \\ \text{subject to} \quad & \hat{Z}_{ij} = Z_{ij} \quad (i,j) \in \Omega \\ & \hat{Z} = W \cdot \hat{C} + M \end{aligned} \tag{5.1}$$

The intuition is that we give more information and confidence about the structure of $\hat{Z}$ through this regularization constraint. With this, we have one more optimization variable $\hat{C}$ while the objective remains unchanged. Figure.5.1 presents the matrix completion results of two convex optimization constraints formulation in small size case: no regularization as Eq.3.3, vs. with regularization as Eq.5.1. Because we are eventually interested in the course structure matrix W, we gradually modify W by increasing the number of edges in the corresponding bipartite graph, and see how the reconstruction errors E_Y and E_Z will change in both aforementioned problem formulations. It shows that adding the constraint $\hat{Z} = W\hat{C} + M$ into the convex optimization program can better regulate the performance of matrix completion, and have lower reconstruction error and stable performance for both $\hat{Z}$ and $\hat{Y}$.



(a) Small-size $Q = 7$, $N = 6$, $K = 3$



(b) Large-size $Q = 10$, $N = 50$, $K = 5$

Figure 5.2: Reconstruction error E_Z and E_Y as a function of the number of randomly observed entries. (a) small-size setup $Q = 7$, $N = 6$, $K = 3$; (b) large-size setup $Q = 10$, $N = 50$, $K = 5$.

5.2.2 Impact of the number of incomplete observations

In this simulation, we study the impact of the number of observations in ground truth matrix Y and consequently Z. Both small-size and large-size cases are discussed here. We vary the extent of observation from no observation to full observation. The locations of the observed entries are generated independent and identically distributed (i.i.d.) and uniformly over the entire matrix. Eq.5.1 convex formulation is employed in this simulation. The W matrix in this simulation is generated randomly and its sparsity structure is taken into account.

Figure.5.2 demonstrates that the reconstruction performance gradually elevates as the number of observations increases for both small-size and large-size problems. For both cases, the reconstruction errors E_Z and E_Y become trivial when roughly 67% of the total entries are randomly observed. This percentage number can fluctuate among different simulations yet the value generally agrees with the bound as given in Section.3.2. Note that the recovery performance of E_Y is slightly better than that of E_Z.

5.2.3 Impact of sparsity level of W

In the simulation, we investigate how the change in the sparsity structure of W will influence the matrix completion performance. Essentially the course design is reflected by the structure and properties of W. This analysis is motivated by the intent of re-designing the course structure in order to have a better estimate on students' performance and maximize the overall students' knowledge. The estimate on students' performance is reflected by E_Z and E_Y, or reflected by the number of needed observations when the reconstruction errors are sufficiently small. And the overall students' knowledge is encoded in every entries of C. As mentioned in Section.4.2, large positive values of $C_{k,j}$ implies strong knowledge of learner j towards the concept k, whereas negative values reveal insufficient knowledge. Thus, a metric to evaluate the overall students' knowledge is $C_{\text{sum}} \in R$, which arithmetically sums up all entries of C. A large positive value of C_{sum} indicates a better overall students' knowledge. The usage of C_{sum} as a metric implicitly assumes that every concept carries the same amount of academic importance and hence are equally important for students. In this experiment, therefore, we are interested in empirically studying how the sparsity structure of W impacts on the accuracy of the matrix completion, and on the

overall capacity of students' knowledge. We consider both small-size and large-size problems.

We start by initializing W with all-zero entries and then in each iteration, randomly selecting one of the zero entries in W and replacing it with $W_{i,k} \sim \mathcal{U}(0,1)$. Such operation is equivalent to randomly adding one more edge with a random edge weight $W_{i,k} \sim \mathcal{U}(0,1)$ in the corresponding bipartite graph in each iteration, until the graph becomes a fully connected bipartite graph. Subsequently, for each new W in every iteration, we obtain the number of observations needed when the normalized reconstruction error E_Y is below a certain threshold ϵ, which is set to be 1×10^{-6} in this experiment. The value of ϵ is chosen to represent exact matrix recovery and take numerical stability into account. Note that varying the number of non-zero entries in W is equivalent to varying the number of concepts covered by the questions. We perform 100 Monte-Carlo trials for small-size case and 25 for large-size case given the restrained computation power. The results are summarized into box plots in Figure.5.3.

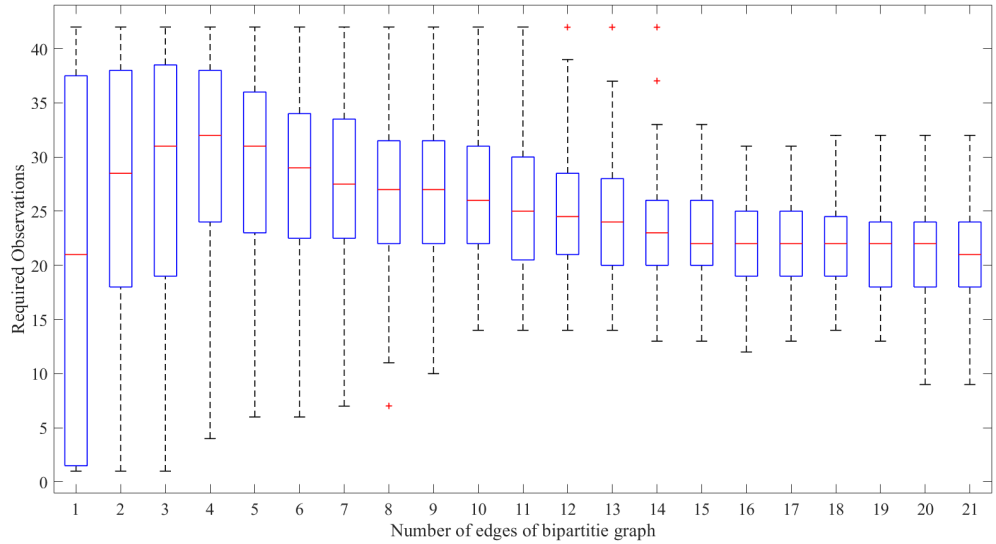
Figure.5.3 (a) and (b) show that generally the required number of observations decreases as W becomes denser, i.e., sparsity level decreases. For the small-size case, the average needed percentage of observations levels off to roughly 60% when the number of edges in the bipartite graph becomes 12. It corresponds to W with slightly more than half of the whole entries being non-trivial, i.e., sparsity level slight less than 50%. For the large-size case, the average required percentage of observations reaches a plateau of roughly 72% when the sparsity level of W is less than 50%. We would appreciate fewer observations to achieve exact recovery because in real setting we always have budget regarding how many observations we can possibly obtain.

Additionally, the optimization Problem (5.1) returns two variables: $\hat{Z}$ and $\hat{C}$. The returned $\hat{C}$ represents our current estimate on the capacity of students' knowledge. Thus, we also observe the trend of the $\hat{C}_{\text{sum}}$ as we increase the number of non-zero entries in W using the aforementioned procedure. The percentage of observations is 67% in this simulation. This simulation intents to discuss how the capacity of students' knowledge, based on our estimation, will change as we vary the course structure. For both small-size and large-size cases, 25 Monte-Carlo trials are simulated and summarized in Figure.5.4.

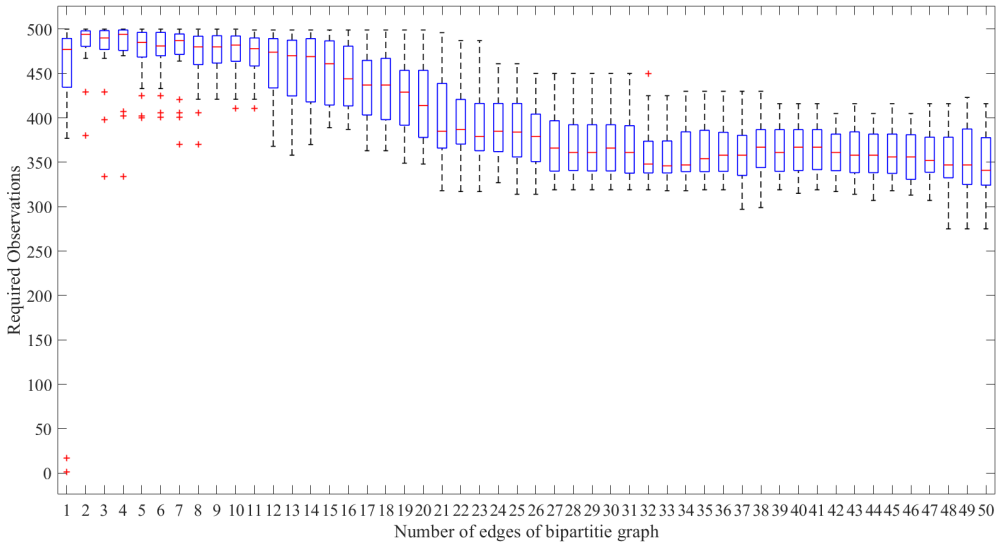
As depicted in Figure.5.4, the $\hat{C}_{\text{sum}}$ increases and roughly reaches a plateau when the sparsity level of W decreases from 100% to 50%. It indicates that the overall capacity of students' knowledge will not significantly change as the course structure becomes more complex, i.e., the number of edges in the corresponding bipartite graph increases. Both results in Figure.5.3 and Figure.5.4 suggest that when designing course structure, a reasonable choice is to construct a corresponding bipartite graph such that it connects slightly more than half of the total edges with respect to the complete graph. The addition of more edges will neither help dramatically reduce the number of required observations nor significantly improve students' knowledge; yet it makes the bipartite graph more complicated in a sense that more concepts are covered by more questions and hence students' workload will exacerbate.

5.2.4 Impact of problem size

In this experiment, we compare the reconstruction error between small-size and large-size problem. It tells us which setting provides a better predication of the students' performance given some practical limits, e.g., number of obtained observations, complexity of the course structure, etc. To this end, besides the above-referenced two different size scenarios, we additionally simulate two bigger sizes scenarios: high dimensional size ($Q = 20$, $N = 100$, $K = 10$) and higher

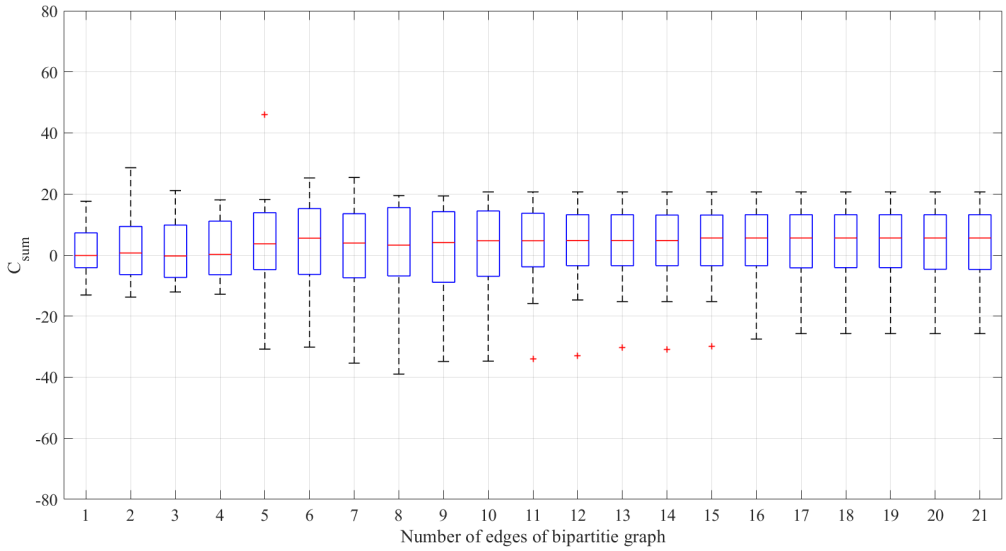


(a) Small-size $Q = 7$, $N = 6$, $K = 3$

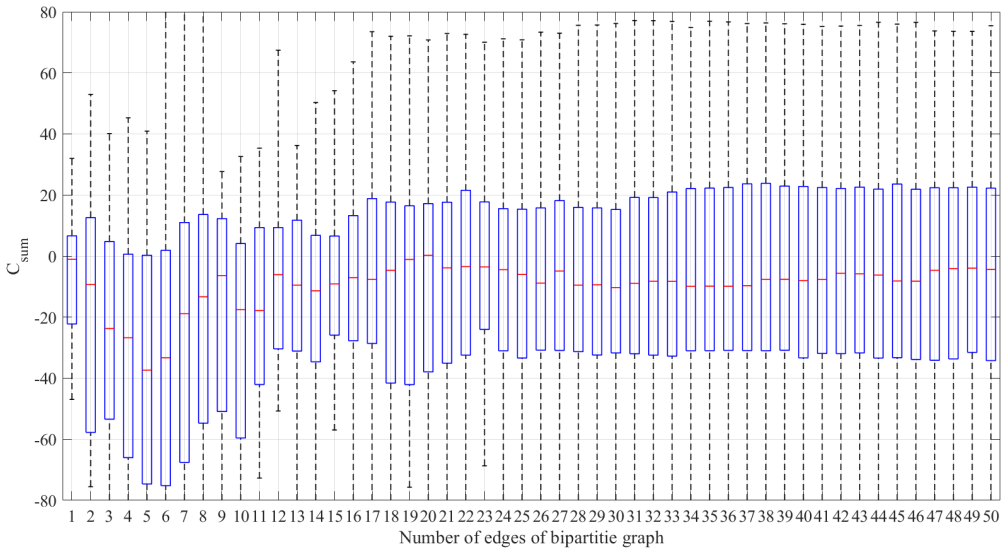


(b) Large-size $Q = 10$, $N = 50$, $K = 5$

Figure 5.3: The number of required observations for exact matrix recovery as a function of the number of connecting edges in the bipartite graph corresponding to W. (a) small-size setup $Q = 7$, $N = 6$, $K = 3$, with 100 independent simulations; (b) large-size setup $Q = 10$, $N = 50$, $K = 5$, with 25 independent simulations.



(a) Small-size $Q = 7$, $N = 6$, $K = 3$



(b) Large-size $Q = 10$, $N = 50$, $K = 5$

Figure 5.4: The estimated overall students' knowledge capacity $\hat{C}_{\text{sum}}$ as a function of the number of connecting edges in the bipartite graph corresponding to W. (a) small-size setup $Q = 7$, $N = 6$, $K = 3$; (b) large-size setup $Q = 10$, $N = 50$, $K = 5$.

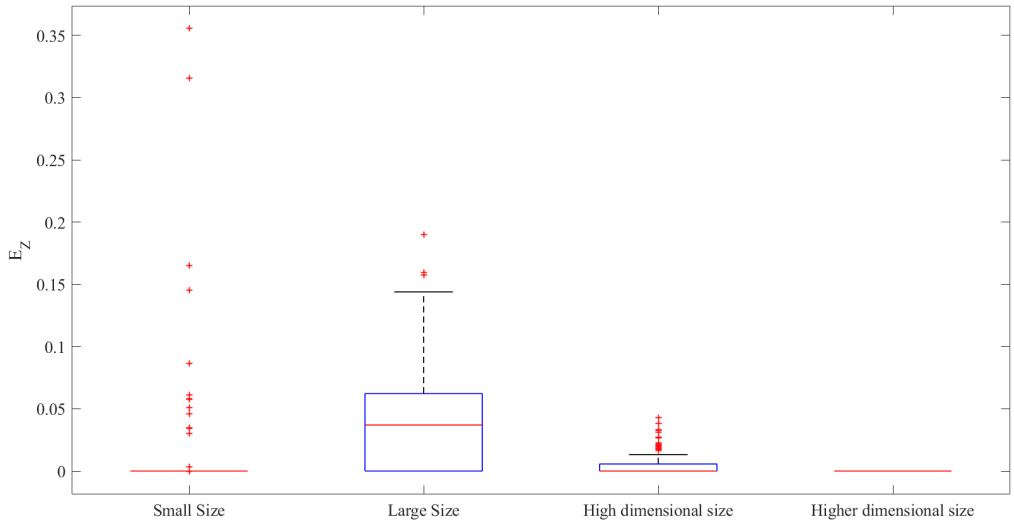
dimensional size ($Q = 50$, $N = 200$, $K = 10$). For all four size cases, we assume observing 75% of the total entries and 40% of the entries of W being zero. Such experimental setup is practical in real setting in that the course content structure should neither be too straightforward nor be too involved, and hence a sparsity level of 40% represents a moderate course complication level. We perform 100 Monte-Carlo trials and summarize the result in Figure.5.5.

As shown in Figure.5.5, with the aforementioned number of observations and level of sparsity in W, both small-size and higher dimensional size cases achieve zero reconstruction error in terms of their median values. However, all simulations in higher dimensional case yield zero error; whereas for small-size case, the result possesses large variance in that many trials return non-zero reconstruction error with totally different values. We can conclude that the dimension of small size problem ($Q = 7$, $N = 6$, $K = 3$) is so small that the estimation performance cannot always be guaranteed. This result suggests that in mini-class teaching scenario, sometimes each individual might possess distinct characteristics, i.e., ground truth matrix has high matrix coherence. As a result, estimating the overall students' performance is hard with incomplete observations. On the other hand, when comparing the estimation performance of the other three cases with high dimension, we observe that the reconstruction performance improves as we further enlarge the problem size. This observation agrees with the work by [16].

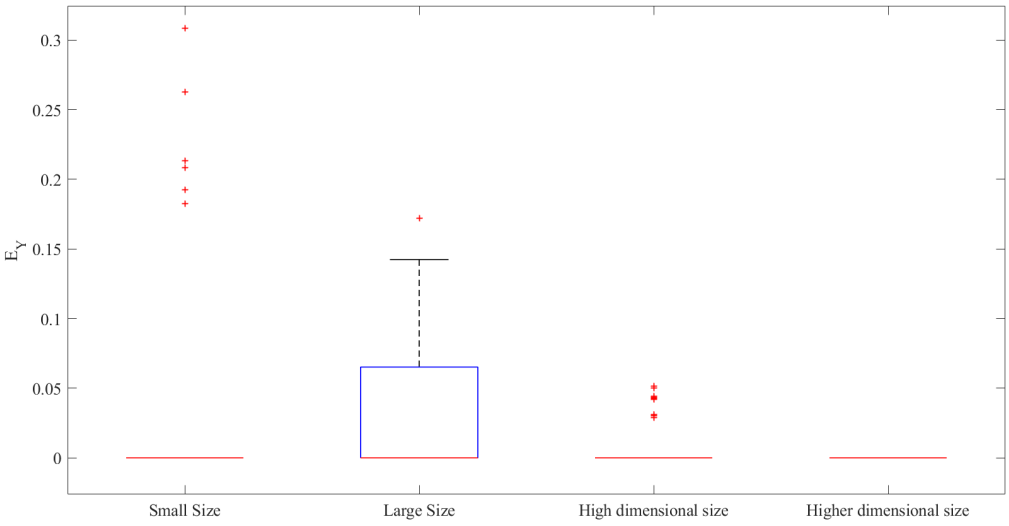
5.2.5 Advantages and Limitations

In this section, we discuss the merits and drawbacks using our method. First of all, we complete the estimate $\hat{Y}$ via reconstructing $\hat{Z}$ and then applying Eq.4.2. The benefit of this procedure is that we don't have to deal with binary matrix completion, especially when this matrix is potentially sparse (i.e., many false answers in the students' response binary matrix). When creating synthetic data, we notice that the ground truth Y is full-rank with large probability. [10] gives a large probability bound for binary matrix being full-rank. Additionally, the sparsity structure may enhance its probability of being full rank as elaborated in [5]. The rank structure is closely related to the coherence and convex optimization, and a low-rank structure is needed for a decent nuclear-norm-based matrix completion. As a result, by virtue of the model $Z = WC + M$, we impose inherent low-rank structure on Z and circumvent the issues related to completing a low-rank binary matrix.

Nevertheless, one drawback of using model $Z = WC + M$ is that we cannot say about the correctness of this model and its variables' distributions. For example, we apply sigmoid function on Z to obtain probability of success. While it is a common practice for obtaining probability, one cannot guarantee that the real relationship between Y and Z complies with this function. In addition, we generate data of W, C, and M based on the distribution described in Section.5.1. There is also another proposed variable distribution mentioned in work [16]. Yet, since the authentic underlying distributions of these variables are not obtainable, one cannot testify the correctness of these proposed distributions. Furthermore, one cannot prove the representativeness of the model $Z = WC + M$ unless substantial real data experiments are performed, which are lacked in this report.



(a) Reconstruction error of $\hat{Z}$



(b) Reconstruction error of $\hat{Y}$

Figure 5.5: Comparison of the estimation performance of different problem sizes in terms of reconstruction errors. The compared four different sizes are: small size $Q = 7$, $N = 6$, $K = 3$; large-size $Q = 10$, $N = 50$, $K = 5$; high-dimensional size $Q = 20$, $N = 100$, $K = 10$; higher-dimensional size $Q = 50$, $N = 200$, $K = 10$. (a) reconstruction error of $\hat{Z}$; (b) reconstruction error of $\hat{Y}$.

Chapter 6

Conclusion

In this report, we have discussed the problem of designing education recommendation system using matrix completion techniques. We have assumed that the student performance is impacted by three factors: the association of each question to the underlying concepts, students' knowledge of the concepts, and each question's intrinsic difficulty. We model the course structure as a bipartite graph, and the student-question performance as a binary matrix. Consequently, we can encode these three aforementioned factors into different matrices, which will mutually influence the student-question performance matrices. We have employed nuclear norm convex optimization technique for reconstructing concerned matrices. In the synthetic data simulation, we have shown the regularization effect of our modification of convex problem formulation. In the meantime, we have discussed the impact of varying the number of observations in "gradebook" matrix, as well as the influence of changing problem sizes on the matrix reconstruction performance. We have obtained an approximate percentage of required observations. We have also found that the matrix reconstruction performance improves as the problem size enlarges, with an exception at the case when the problem size is sufficiently small. Additionally, we have concluded that a reasonable choice on designing the course structure bipartite graph is to connect slightly more than half of the total edges with respect to the complete bipartite graph. This design choice balances the prediction result on overall student-question performance, the capacity the overall students' knowledge, and the students' workload.

A future work of this report is to incorporate dynamic update to maximize the student performance given some constraints on the course structure and question easiness level. Specifically, after designing a course structure and question easiness level in year k, we can perform matrix completion to obtain our estimated student performance and student knowledge $\hat{Y}_k$ and $\hat{C}_k$ in year k, respectively. Subsequently, we would like to re-design course structure and easiness level (W_k and M_k) with the objective of maximizing the student knowledge $\hat{Y}_{k+1}$ in the following year $k+1$. The observation data set Ω_{k+1} may change in the next year, and the overall student knowledge capacity C_{k+1} might fluctuate. We can model this by adding a white noise to each entry of $\hat{C}_k$ to make it a ground truth student knowledge C_{k+1} in year $k+1$. Certain bounds on the $\|M\|_F$ and $\|W\|_F$ shall be applied. Moreover, each question has a limited number of concepts it can cover, which enforces a ℓ_0-norm constraint on each row of W. Some constraints of this problem formulation can be non-convex, and hence require a convex relaxation or non-convex optimization solver.

Bibliography

[1] Nana Yaw Asabere. "Review of Recommender Systems for Learners in Mobile Social/Collaborative Learning". In: *International Journal of Information* 2.5 (2012).

[2] Emmanuel Candes and Benjamin Recht. "Exact matrix completion via convex optimization". In: *Communications of the ACM* 55.6 (2012), pp. 111–119.

[3] Steve Chien et al. "Private alternating least squares: Practical private matrix completion with tighter rates". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 1877–1887.

[4] Alexander L Chistov and D Yu Grigor'Ev. "Complexity of quantifier elimination in the theory of algebraically closed fields". In: *International Symposium on Mathematical Foundations of Computer Science*. Springer. 1984, pp. 17–31.

[5] Kevin P Costello and Van Vu. "On the rank of random sparse matrices". In: *Combinatorics, Probability and Computing* 19.3 (2010), pp. 321–342.

[6] Mark A Davenport and Justin Romberg. "An overview of low-rank matrix recovery from incomplete observations". In: *IEEE Journal of Selected Topics in Signal Processing* 10.4 (2016), pp. 608–622.

[7] Mark A Davenport et al. "1-bit matrix completion". In: *Information and Inference: A Journal of the IMA* 3.3 (2014), pp. 189–223.

[8] Reinhard Diestel. "Graph theory 3rd ed". In: *Graduate texts in mathematics* 173.33 (2005), p. 12.

[9] Brian Eriksson, Laura Balzano, and Robert Nowak. "High-rank matrix completion". In: *Artificial Intelligence and Statistics*. PMLR. 2012, pp. 373–381.

[10] Paulo JSG Ferreira et al. "The rank of random binary matrices and distributed storage applications". In: *IEEE communications letters* 17.1 (2012), pp. 151–154.

[11] Shuhang Gu et al. "Weighted nuclear norm minimization and its applications to low level vision". In: *International journal of computer vision* 121 (2017), pp. 183–208.

[12] Shuhang Gu et al. "Weighted nuclear norm minimization with application to image denoising". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 2862–2869.

[13] Prateek Jain, Raghu Meka, and Inderjit Dhillon. "Guaranteed rank minimization via singular value projection". In: *Advances in Neural Information Processing Systems* 23 (2010).

[14] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. "Low-rank matrix completion using alternating minimization". In: *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. 2013, pp. 665–674.

[15] Raghunandan Hulikal Keshavan. *Efficient algorithms for collaborative filtering*. Stanford University, 2012.

[16] Andrew S Lan et al. "Sparse factor analysis for learning and content analytics". In: *Journal of Machine Learning Research (JMLR)* 15.57 (2014), pp. 1959–2008.

[17] Jie Lu et al. "Recommender system application developments: a survey". In: *Decision support systems* 74 (2015), pp. 12–32.

[18] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. "Spectral regularization algorithms for learning large incomplete matrices". In: *The Journal of Machine Learning Research* 11 (2010), pp. 2287–2322.

[19] Mehryar Mohri and Ameet Talwalkar. "Can matrix coherence be efficiently and accurately estimated?" In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2011, pp. 534–542.

[20] Ivan Nazarov et al. "Sparse group inductive matrix completion". In: *arXiv preprint arXiv:1804.10653* (2018).

[21] Melvin R Novick. "The axioms and principal results of classical test theory". In: *Journal of mathematical psychology* 3.1 (1966), pp. 1–18.

[22] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization". In: *SIAM review* 52.3 (2010), pp. 471–501.

[23] Sagar Sharma, Simone Sharma, and Anidhya Athaiya. "Activation functions in neural networks". In: *Towards Data Sci* 6.12 (2017), pp. 310–316.

[24] Horst D Simon and Hongyuan Zha. "Low-rank matrix approximation using the Lanczos bidiagonalization process with applications". In: *SIAM Journal on Scientific Computing* 21.6 (2000), pp. 2257–2274.