RICE UNIVERSITY

Sparse Factor Analysis for Learning and Content Analytics

by

Andrew S. Lan

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree Master of Science

APPROVED, THESIS COMMITTEE:

Richard G. Baraniuk, Chair
Victor E. Cameron Professor of Electrical and Computer Engineering and Computer Science

Asif V. Veeraraghavan
Assistant Professor of Electrical and Computer Engineering

Genevera I. Allen
Assistant Professor of Statistics and Electrical and Computer Engineering

Houston, Texas
April, 2014
ABSTRACT

Sparse Factor Analysis for Learning and Content Analytics

by

Andrew S. Lan

We develop a new model and algorithms for machine learning-based learning analytics, which estimate a learner’s knowledge of the concepts underlying a domain, and content analytics, which estimate the relationships among a collection of questions and those concepts. Our model represents the probability that a learner provides the correct response to a question in terms of three factors: their understanding of a set of underlying concepts, the concepts involved in each question, and each question’s intrinsic difficulty. We estimate these factors given the graded responses to a collection of questions. The underlying estimation problem is ill-posed in general, especially when only a subset of the questions are answered. The key observation that enables a well-posed solution is the fact that typical educational domains of interest involve only a small number of key concepts. Leveraging this observation, we develop a bi-convex maximum-likelihood solution to the resulting SPARse Factor Analysis (SPARFA) problem. We also incorporate instructor-defined tags on questions and question text to facilitate the interpretability of the estimated factors. Experiments with synthetic and real-world data demonstrate the efficacy of our approach.
Acknowledgement

First of all, I would like to thank my advisor, Prof. Richard Baraniuk for his inspiring advices and motivating guidance throughout this period, without which this work could not be done. His insights on personalized learning is of paramount importance to my research.

Then I would also like to thank Dr. Christoph Studer and Dr. Andrew Waters, for countless insightful comments, helpful discussions and alcohol. Also thanks to my officemate Amirali Aghazadeh for a fantastical working environment!

I would also like to thank Prof. Wotao Yin for helpful discussions on the convergence proof of SPARFA-M, Prof. Genevera Allen for insights into probit regression, and also Prof. Ashok Veeraraghavan for serving as my thesis committee. Your comments are very helpful.

I would also like to thank Daniel Calderón for organizing and administering the Amazon Mechanical Turk experiments in Section 6.3.2, and Carlos Monroy and Reid Whitaker for providing the STEMscopes data. Also thanks to every DSP group member for a cool research atmosphere.

Most importantly, I want to thank my parents and all my non-Rice friends for their support. Every one of you reminds me of the joyful times we had prior to my arrival at Rice, supporting me through my ups and downs.

Please see our website www.sparfa.com, where you can learn more about the project and purchase SPARFA t-shirts and other merchandise.
Contents

Abstract ii
List of Illustrations vii
List of Tables x

1 Introduction 1
1.1 The Promise of Personalized Learning 1
1.2 Sparse Factor Analysis (SPARFA) 3

2 Statistical Model for Learning and Content Analytics 9
2.1 Model for Graded Learner Response Data 9
2.2 Joint Estimation of Concept Knowledge and Question–Concept Association 11

3 SPARFA-M: Maximum Likelihood Sparse Factor Analysis 14
3.1 Problem Formulation 14
3.2 The SPARFA-M Algorithm 16
3.3 Accelerated First-Order Methods for Regularized Probit/Logit Regression 18
3.4 Convergence Analysis of SPARFA-M 20
  3.4.1 Convergence Analysis of Regularized Regression using FISTA 21
  3.4.2 Convergence Analysis of SPARFA-M 22
3.5 Algorithmic Details and Improvements for SPARFA-M 23
  3.5.1 Reducing Computational Complexity in Practice 23
3.5.2 Reducing the Chance of Getting Stuck in Local Minima . . . . 24
3.5.3 Parameter Selection . . . . . . . . . . . . . . . . . . . . . . . . 25

4 Tag Analysis: Post-Processing to Interpret the Estimated Concepts

4.1 Incorporating Question–Tag Information . . . . . . . . . . . . . . . 27
4.2 Estimating the Concept–Tag Associations and Learner–Tag Knowledge 28
  4.2.1 Associating Tags to Each Concept . . . . . . . . . . . . . . . . 28
  4.2.2 Learner Tag Knowledge Profiles . . . . . . . . . . . . . . . . . 29

5 Experiments

5.1 Synthetic Data Experiments . . . . . . . . . . . . . . . . . . . . . . 30
  5.1.1 Baseline Algorithm: K-SVD+ . . . . . . . . . . . . . . . . . . . 31
  5.1.2 Performance Measures . . . . . . . . . . . . . . . . . . . . . . 32
  5.1.3 Impact of Problem Size and Number of Concepts . . . . . . . 33
  5.1.4 Impact of the Number of Incomplete Observations . . . . . . 34
  5.1.5 Impact of Sparsity Level . . . . . . . . . . . . . . . . . . . . . 36
  5.1.6 Impact of Model Mismatch . . . . . . . . . . . . . . . . . . . . 37
5.2 Real Data Experiments . . . . . . . . . . . . . . . . . . . . . . . . . . 38
  5.2.1 Undergraduate DSP course . . . . . . . . . . . . . . . . . . . . 39
  5.2.2 Grade 8 science course . . . . . . . . . . . . . . . . . . . . . . 42
  5.2.3 Algebra Test Administered on Amazon Mechanical Turk . . . 42

6 Extension to SPARFA: Ordinal SPARFA-Tag

6.1 Ordinal SPARFA-Tag Statistical Model . . . . . . . . . . . . . . . . 47
  6.1.1 Model for ordinal learner response data . . . . . . . . . . . . . 47
  6.1.2 Additional assumptions . . . . . . . . . . . . . . . . . . . . . . 50
6.2 Algorithm . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 50
  6.2.1 Ordinal SPARFA-M . . . . . . . . . . . . . . . . . . . . . . . . 51
Illustrations

1.1 (a) The SPARFA framework processes a (potentially incomplete) binary-valued dataset of graded learner–question responses to (b) estimate the underlying questions-concept association graph and the abstract conceptual knowledge of each learner (illustrated here by smiley faces for learner $j = 3$, the column in (a) selected by the red dashed box).

1.2 (a) Sparse question–concept association graph and (b) most important tags associated with each concept for Grade 8 Earth science with $N = 135$ learners answering $Q = 80$ questions. Only 13.5% of all graded learner–question responses were observed.

5.1 Performance comparison of SPARFA-M, SPARFA-B, and K-SVD$_+$ for different problem sizes $Q \times N$ and number of concepts $K$. The performance naturally improves as the problem size increases, while both SPARFA algorithms outperform K-SVD$_+$. $M$ denotes SPARFA-M, $B$ denotes SPARFA-B, and $K$ denotes K-SVD$_+$. 

5.2 Performance comparison of SPARFA-M, SPARFA-B, and K-SVD$_+$ for different percentages of observed entries in $Y$. The performance degrades gracefully as the number of observations decreases, while the SPARFA algorithms outperform K-SVD$_+$. 
5.3 Performance comparison of SPARFA-M, SPARFA-B, and K-SVD+ for different sparsity levels in the rows in $W$. The performance degrades gracefully as the sparsity level increases, while the SPARFA algorithms outperform K-SVD+. 

5.4 Performance comparison of SPARFA-M, SPARFA-B, and K-SVD+ with probit/logit model mismatch; $M_P$ and $M_L$ indicate probit and logit SPARFA-M, respectively. In the left/right halves of each box plot, we generate $Y$ according to the inverse probit/logit link functions. The performance degrades only slightly with mismatch, while both SPARFA algorithms outperform K-SVD+. 

5.5 (a) Question–concept association graph and (b) most important tags associated with each concept for an undergraduate DSP course with $N = 15$ learners answering $Q = 44$ questions. 

5.6 (a) Question–concept association graph and (b) most important tags associated with each concept for a high-school algebra test carried out on Amazon Mechanical Turk with $N = 99$ users answering $Q = 34$ questions. 

6.1 Performance comparison of Ordinal SPARFA-M vs. K-SVD+. “SP” denotes Ordinal SPARFA-M without given support $\Gamma$ of $W$, “SPP” denotes the variant with estimated precision $\tau$, and “SPT” denotes Ordinal SPARFA-Tag. “KS” stands for K-SVD+, and “KST” denotes its variant with given support $\Gamma$. 

6.2 Performance comparison of Ordinal SPARFA-M vs. K-SVD+ by varying the number of quantization bins. “SP” denotes Ordinal SPARFA-M, “KSY” denotes K-SVD+ operating on $Y$, and “KSZ” denotes K-SVD+ operating on $Z$ in (2.2) (the unquantized data).
6.3 Question–concept association graph for a high-school algebra test
with $N = 99$ users answering $Q = 34$ questions. Boxes represent
questions; circles represent concepts. We furthermore show the
unique tag associated with each concept. . . . . . . . . . . . . . . . . 58

6.4 Question–concept association graph for a grade 8 Earth Science
course with $N = 145$ learners answering $Q = 80$ questions ($Y$ is
highly incomplete with only 13.5% entries observed). We furthermore
show the unique tag associated with each concept. . . . . . . . . . . . 59

6.5 Prediction performance on the Mechanical Turk algebra test dataset.
We compare the collaborative filtering methods SVD++ and OrdRec
to various Ordinal SPARFA-M based methods: “Nuc” uses the
nuclear norm constraint, “Fro” uses the Frobenius norm constraint,
“Bin” and “BinInd” learn the bin boundaries, whereas “Bin” learns
one set of bin boundaries for the entire dataset and “BinInd” learns
individual bin boundaries for each question. . . . . . . . . . . . . . . 63

7.1 Question–concept association graph and most important keywords
recovered by SPARFA-Top for the STEMscopes dataset; boxes
represent questions, circles represent concepts, and thick lines
represent strong question–concept associations. . . . . . . . . . . . . 72

7.2 Question–concept association graph and 3 most important keyword
recovered by SPARFA-Top for the algebra test dataset; boxes
represent questions, circles represent concepts, and thick lines
represent strong question–concept associations. . . . . . . . . . . . . 73

7.3 Average predicted likelihood on 20% hold-out data in $Y$ using
SPARFA-Top with different precision parameters $\tau$. For $\tau \to \infty$
SPARFA-Top corresponds to SPARFA. . . . . . . . . . . . . . . . . . 74
Tables

5.1 Selected tag knowledge of Learner 1. .......................... 41
5.2 Average tag knowledge of all learners. .......................... 41
5.3 Graded responses and their underlying concepts for Learner 1 (1
designates a correct response and 0 an incorrect response). ........... 45
5.4 Estimated concept knowledge for Learner 1. ...................... 45
Chapter 1

Introduction

Textbooks, lectures, and homework assignments were the answer to the main educational challenges of the 19th century, but they are the main bottleneck of the 21st century. Today’s textbooks are static, linearly organized, time-consuming to develop, soon out-of-date, and expensive. Lectures remain a primarily passive experience of copying down what an instructor says and writes on a board (or projects on a screen). Homework assignments that are not graded for weeks provide poor feedback to learners (e.g., students) on their learning progress. Even more importantly, today’s courses provide only a “one-size-fits-all” learning experience that does not cater to the background, interests, and goals of individual learners.

1.1 The Promise of Personalized Learning

We envision a world where access to high-quality, personally tailored educational experiences is affordable to all of the world’s learners. The key is to integrate textbooks, lectures, and homework assignments into a personalized learning system (PLS) that closes the learning feedback loop by (i) continuously monitoring and analyzing learner interactions with learning resources in order to assess their learning progress and (ii) providing timely remediation, enrichment, or practice based on that analysis.
See [2], [3], [4], [5], [6], and [7] for various visions and examples.

Some progress has been made over the past few decades on personalized learning; see, for example, the sizable literature on intelligent tutoring systems discussed in [8].

To date, the lionshare of fielded, intelligent tutors have been rule-based systems that are hard-coded by domain experts to give learners feedback for pre-defined scenarios (e.g., [9], [10], [11], [12], and [13]). The specificity of such systems is counterbalanced by their high development cost in terms of both time and money, which has limited their scalability and impact in practice.

In a fresh direction, recent progress has been made on applying machine learning algorithms to mine learner interaction data and educational content (see the overview articles by [14] and [15]). In contrast to rule-based approaches, machine learning-based PLSs promise to be rapid and inexpensive to deploy, which will enhance their scalability and impact. Indeed, the dawning age of “big data” provides new opportunities to build PLSs based on data rather than rules. We conceptualize the architecture of a generic machine learning-based PLS to have three interlocking components:

- **Learning analytics**: Algorithms that estimate what each learner does and does not understand based on data obtained from tracking their interactions with learning content.

- **Content analytics**: Algorithms that organize learning content such as text, video, simulations, questions, and feedback hints.
• **Scheduling**: Algorithms that use the results of learning and content analytics to suggest to each learner at each moment what they should be doing in order to maximize their learning outcomes, in effect closing the learning feedback loop.

### 1.2 Sparse Factor Analysis (SPARFA)

In this work, we develop a new model and a suite of algorithms for joint machine learning-based *learning analytics* and *content analytics*. Our model (developed in Section 2) represents the probability that a learner provides the correct response to a given question in terms of three factors: their knowledge of the underlying concepts, the concepts involved in each question, and each question’s intrinsic difficulty.

Figure 1.1 provides a graphical depiction of our approach. As shown in Figure 1.1(a), we are provided with data relating to the correctness of the learners’ responses to a collection of questions. We encode these graded responses in a “gradebook,” a source of information commonly used in the context of classical test theory ([16]). Specifically, the “gradebook” is a matrix with entry $Y_{i,j} = 1$ or 0 depending on whether learner $j$ answers question $i$ correctly or incorrectly, respectively. Question marks correspond to incomplete data due to unanswered questions. Working left-to-right in Figure 1.1(b), we assume that the collection of questions (rectangles) is related to a small number of abstract concepts (circles) by a graph, where the edge weight $W_{i,k}$ indicates the degree to which question $i$ involves concept $k$. We also assume that question $i$ has intrinsic difficulty $\mu_i$. Denoting learner $j$’s knowledge of concept $k$ by
(a) Graded learner–question responses. (b) Inferred question–concept association graph.

Figure 1.1: (a) The SPARFA framework processes a (potentially incomplete) binary-valued dataset of graded learner–question responses to (b) estimate the underlying questions-concept association graph and the abstract conceptual knowledge of each learner (illustrated here by smiley faces for learner $j = 3$, the column in (a) selected by the red dashed box).

For $C_{k,j}$, we calculate the probabilities that the learners answer the questions correctly as $WC + M$, where $W$ and $C$ are matrix versions of $W_{i,k}$ and $C_{k,j}$, respectively, and $M$ is a matrix containing the intrinsic question difficulty $\mu_i$ on row $i$. We transform the probability of a correct answer to an actual 1/0 correctness via a standard probit or logit link function (see [17]).

Armed with this model and given incomplete observations of the graded learner–question responses $Y_{i,j}$, our goal is to estimate the factors $W$, $C$, and $M$. Such a factor-analysis problem is ill-posed in general, especially when each learner answers only a small subset of the collection of questions (see [18] for a factor analysis overview). Our first key observation that enables a well-posed solution is the fact that
typical educational domains of interest involve only a small number of key concepts. Consequently, \( W \) becomes a tall, narrow matrix that relates the questions to a small set of abstract concepts, while \( C \) becomes a short, wide matrix that relates learner knowledge to that same small set of abstract concepts. Note that the concepts are “abstract” in that they will be estimated from the data rather than dictated by a subject matter expert. Our second key observation is that each question involves only a small subset of the abstract concepts. Consequently, the matrix \( W \) is sparsely populated. Our third observation is that the entries of \( W \) should be non-negative to ensure that large positive values in \( C \) represent strong knowledge of the associated abstract concepts.

Leveraging these observations, we propose below a suite of new algorithms for solving the \textit{SPARse Factor Analysis} (SPARFA) problem. Section 3 develops SPARFA-M, which uses an efficient bi-convex optimization approach to produce maximum likelihood point estimates of the factors. Section 4 develops a novel method for incorporating user-defined tags that label the questions, in order to facilitate interpretation of the abstract concepts estimated by the SPARFA algorithms.

In Section 5, we report on a range of experiments with a variety of synthetic and real-world data that demonstrate the wealth of information provided by the estimates of \( W, C \) and \( M \). As an example, Figure 1.2 provides the results for a dataset collected from learners using [19], a science curriculum platform. The dataset consists of 145 Grade 8 learners from a single school district answering a tagged set of 80 questions.
on Earth science; only 13.5% of all graded learner–question responses were observed. We applied the SPARFA-M algorithm to retrieve the factors $\mathbf{W}$, $\mathbf{C}$, and $\mathbf{M}$ using 5 latent concepts. The resulting sparse matrix $\mathbf{W}$ is displayed as a bipartite graph in Figure 1.2(a); circles denote the abstract concepts and boxes denote questions. Each question box is labeled with its estimated intrinsic difficulty $\mu_i$, with large positive values denoting easy questions. Links between the concept and question nodes represent the active (non-zero) entries of $\mathbf{W}$, with thicker links denoting larger values $W_{i,k}$. Unconnected questions are those for which no abstract concept explained the learners’ answer pattern; these questions typically have either very low or very high intrinsic difficulty, resulting in nearly all learners answering them correctly or incorrectly. The tags in Figure 1.2(b) provide interpretation to the estimated abstract concepts.

We envision a range of potential learning and content analytics applications for the SPARFA framework that go far beyond the standard practice of merely forming column sums of the “gradebook” matrix (with entries $Y_{i,j}$) to arrive at a final scalar numerical score for each learner (which is then often further quantized to a letter grade on a 5-point scale). Each column of the estimated $\mathbf{C}$ matrix can be interpreted as a measure of the corresponding learner’s knowledge about the abstract concepts. Low values indicate concepts ripe for remediation, while high values indicate concepts ripe for enrichment. The sparse graph stemming from the estimated $\mathbf{W}$ matrix automatically groups questions into similar types based on their concept association. The
(a) Inferred question–concept association graph.

<table>
<thead>
<tr>
<th>Concept 1</th>
<th>Concept 2</th>
<th>Concept 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Changes to land (45%)</td>
<td>Evidence of the past (74%)</td>
<td>Alternative energy (76%)</td>
</tr>
<tr>
<td>Properties of soil (28%)</td>
<td>Mixtures and solutions (14%)</td>
<td>Environmental changes (19%)</td>
</tr>
<tr>
<td>Uses of energy (27%)</td>
<td>Environmental changes (12%)</td>
<td>Changes from heat (5%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Concept 4</th>
<th>Concept 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Properties of soil (77%)</td>
<td>Formulation of fossil fuels (54%)</td>
</tr>
<tr>
<td>Environmental changes (17%)</td>
<td>Mixtures and solutions (28%)</td>
</tr>
<tr>
<td>Classifying matter (6%)</td>
<td>Uses of energy (18%)</td>
</tr>
</tbody>
</table>

(b) Most important tags and relative weights for the estimated concepts.

Figure 1.2: (a) Sparse question–concept association graph and (b) most important tags associated with each concept for Grade 8 Earth science with \( N = 135 \) learners answering \( Q = 80 \) questions. Only 13.5% of all graded learner–question responses were observed.
matrix makes it straightforward to find a set of questions similar to a given target question. Finally, the estimated $M$ matrix (with entries $\mu_i$ on each row) provides an estimate of each question’s intrinsic difficulty. This enables an instructor to assign questions in an orderly fashion as well as to prune out potentially problematic questions that are either too hard, too easy, too confusing, or unrelated to the concepts underlying the collection of questions.

In Section 6 and Section 7, we extend the basic SPARFA framework from analyzing binary graded learner responses to ordinal responses, while also utilizing tags and question text to provide more interpretability of the estimated concepts. An overview of related work on machine learning-based personalized learning is provided in Section 8, and we conclude in Section 9. All proofs are relegated to two appendices.
Chapter 2

Statistical Model for Learning and Content Analytics

Our approach to learning and content analytics is based on a new statistical model that encodes the probability that a learner will answer a given question correctly in terms of three factors: (i) the learner’s knowledge of a set of latent, abstract concepts, (ii) how the question is related to each concept, and (iii) the intrinsic difficulty of the question.

2.1 Model for Graded Learner Response Data

Let $N$ denote the total number of learners, $Q$ the total number of questions, and $K$ the number of latent abstract concepts. We define $C_{k,j}$ as the concept knowledge of learner $j$ on concept $k$, with large positive values of $C_{k,j}$ corresponding to a better chance of success on questions related to concept $k$. Stack these values into the column vector $c_j \in \mathbb{R}^K$, $j \in \{1, \ldots, N\}$ and the $K \times N$ matrix $C = [c_1, \ldots, c_N]$.

We further define $W_{i,k}$ as the question–concept association of question $i$ with respect to concept $k$, with larger values denoting stronger involvement of the concept. Stack these values into the column vector $\bar{w}_i \in \mathbb{R}^K$, $i \in \{1, \ldots, Q\}$ and the $Q \times K$ matrix $W = [\bar{w}_1, \ldots, \bar{w}_Q]^T$. Finally, we define the scalar $\mu_i \in \mathbb{R}$ as the intrinsic difficulty of question $i$, with larger values representing easier questions. Stack these values into
the column vector $\mu$ and form the $Q \times N$ matrix $M = \mu 1_{1 \times N}$ as the product of $\mu = [\mu_1, \ldots, \mu_Q]^T$ with the $N$-dimensional all-ones row vector $1_{1 \times N}$.

Given these definitions, we propose the following model for the binary-valued graded response variable $Y_{i,j} \in \{0, 1\}$ for learner $j$ on question $i$, with 1 representing a correct response and 0 an incorrect response:

$$Z_{i,j} = \mathbf{w}_i^T c_j + \mu_i, \quad \forall i, j,$$

$$Y_{i,j} \sim Ber(\Phi(Z_{i,j})), \quad (i, j) \in \Omega_{\text{obs}}.$$  \hfill (2.1)

Here, $Ber(z)$ designates a Bernoulli distribution with success probability $z$, and $\Phi(z)$ denotes an inverse link function that maps a real value $z$ to the success probability of a binary random variable. Thus, the slack variable $\Phi(Z_{i,j}) \in [0, 1]$ governs the probability of learner $j$ answering question $i$ correctly. The set $\Omega_{\text{obs}} \subseteq \{1, \ldots, Q\} \times \{1, \ldots, N\}$ in (2.1) contains the indices associated with the observed graded learner response data. Hence, our framework is able to handle the case of incomplete or missing data (e.g., when the learners do not answer all of the questions). Stack the values $Y_{i,j}$ and $Z_{i,j}$ into the $Q \times N$ matrices $Y$ and $Z$, respectively. We can conveniently rewrite (2.1) in matrix form as

$$Y_{i,j} \sim Ber(\Phi(Z_{i,j})), \quad (i, j) \in \Omega_{\text{obs}} \quad \text{with} \quad Z = WC + M.$$  \hfill (2.2)

In this paper, we focus on the two most commonly used link functions in the machine learning literature. The inverse probit function is defined as

$$\Phi_{\text{pro}}(x) = \int_{-\infty}^x N(t) \, dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} \, dt.$$  \hfill (2.3)
where \( \mathcal{N}(t) = \frac{1}{\sqrt{2 \pi}} e^{-t^2/2} \) is the probability density function (PDF) of the standard normal distribution (with mean zero and variance one). The inverse logit link function is defined as

\[ \Phi_\text{log}(x) = \frac{1}{1 + e^{-x}}. \]  

As we noted in the Introduction, \( C, W \), and \( \mu \) (or equivalently, \( M \)) have natural interpretations in real education settings. Column \( j \) of \( C \) can be interpreted as a measure of learner \( j \)'s knowledge about the abstract concepts, with larger \( C_{k,j} \) values implying more knowledge. The non-zero entries in \( W \) can be used to visualize the connectivity between concepts and questions (see Figure 1.1(b) for an example), with larger \( W_{i,k} \) values implying stronger ties between question \( i \) and concept \( k \). The values of \( \mu \) contains estimates of each question's intrinsic difficulty.

### 2.2 Joint Estimation of Concept Knowledge and Question–Concept Association

Given a (possibly partially observed) matrix of graded learner response data \( Y \), we aim to estimate the learner concept knowledge matrix \( C \), the question–concept association matrix \( W \), and the question intrinsic difficulty vector \( \mu \). In practice, the latent factors \( W \) and \( C \), and the vector \( \mu \) will contain many more unknowns than we have observations in \( Y \); hence, estimating \( W \), \( C \), and \( \mu \) is, in general, an ill-posed inverse problem. The situation is further exacerbated if many entries in \( Y \) are unobserved.
To regularize this inverse problem, prevent over-fitting, improve identifiability,* and enhance interpretability of the entries in \( C \) and \( W \), we appeal to the following three observations regarding education that are reasonable for typical exam, homework, and practice questions at all levels. We will exploit these observations extensively in the sequel as fundamental assumptions:

(A1) *Low-dimensionality:* The number of latent, abstract concepts \( K \) is small relative to both the number of learners \( N \) and the number of questions \( Q \). This implies that the questions are redundant and that the learners’ graded responses live in a low-dimensional space. We further assume that the number of latent concepts \( K \) is known a priori.†

(A2) *Sparsity:* A learner’s ability to answer a given question correctly (or incorrectly) depends only on a small subset of the concepts. In other words, we assume that the matrix \( W \) is sparsely populated, i.e., contains mostly zero entries.

(A3) *Non-negativity:* A learner’s knowledge of a given concept does not negatively affect their probability of correctly answering a given question, i.e., prior knowledge of a concept is not “harmful.” In other words, the entries of \( W \) are non-

---

*If \( Z = WC \), then for any orthonormal matrix \( H \) with \( H^T H = I \), we have \( Z = WH^T HC = \tilde{W} \tilde{C} \). Hence, the estimation of \( W \) and \( C \) is, in general, non-unique up to a unitary matrix rotation.

†Small \( K \) extracts just a few concepts, whereas large \( K \) extracts more concepts but can result in over-fitting. If the goal is to predict missing entries in \( Y \), then standard techniques like cross-validation ([20]) can be used to select \( K \).
negative, which provides a natural interpretation for the entries in $C$: Large
values $C_{k,j}$ indicate strong knowledge of the corresponding concept, whereas
negative values indicate weak knowledge.

In practice, $N$ can be larger than $Q$ and vice versa, and hence, we do not impose any
additional assumptions on their values.

We will refer to the problem of estimating $W$, $C$, and $\mu$, given the observations $Y$,
under the assumptions (A1)–(A3) as the SPARse Factor Analysis (SPARFA) problem.
We now develop two complementary algorithms to solve the SPARFA problem. In
Section 3, we introduce SPARFA-M, a computationally efficient matrix-factorization
approach that produces maximum-likelihood point estimates of the quantities of in-
terest.
Chapter 3

SPARFA-M: Maximum Likelihood Sparse Factor Analysis

Our first algorithm, SPARFA-M, solves the SPARFA problem using maximum-likelihood-based probit or logit regression.

3.1 Problem Formulation

To estimate $W$, $C$, and $\mu$, we maximize the likelihood of the observed data $Y_{i,j}$, $(i,j) \in \Omega_{\text{obs}}$

$$p(Y_{i,j}|\bar{w}_i, c_j) = \Phi(\bar{w}_i^T c_j)^{Y_{i,j}} (1 - \Phi(\bar{w}_i^T c_j))^{1-Y_{i,j}}$$

given $W$, $C$, and $\mu$ and subject to the assumptions (A1), (A2), and (A3) from Section 2.2. This yields the following optimization problem:

$$(P^*) \left\{ \begin{array}{l} \text{maximize} \sum_{(i,j) \in \Omega_{\text{obs}}} \log p(Y_{i,j}|\bar{w}_i, c_j) \\ \text{subject to} \|\bar{w}_i\|_0 \leq s \forall i, \|\bar{w}_i\|_2 \leq \kappa \forall i, \ W_{i,k} \geq 0 \forall i, k, \|C\|_F = \xi. \end{array} \right.$$ 

Let us take a quick tour of the problem $(P^*)$ and its constraints. The intrinsic difficulty vector $\mu$ is incorporated as an additional column of $W$, and $C$ is augmented with an all-ones row accordingly. We impose sparsity on each vector $\bar{w}_i$ to comply with (A2) by limiting its maximum number of nonzero coefficients using the constraint
\[ \| \bar{w}_i \|_0 \leq s; \text{ here } \|a\|_0 \text{ denotes the number of non-zero entries in the vector } a. \] The \( \ell_2 \)-norm constraint on each vector \( \bar{w}_i \) with \( \kappa > 0 \) is required for our convergence proof below. We enforce non-negativity on each entry \( W_{i,k} \) to comply with (A3). Finally, we normalize the Frobenius norm of the concept knowledge matrix \( C \) to a given \( \xi > 0 \) to suppress arbitrary scalings between the entries in both matrices \( W \) and \( C \).

Unfortunately, optimizing over the sparsity constraints \( \| \bar{w}_i \|_0 \leq s \) requires a combinatorial search over all \( K \)-dimensional support sets having \( s \) non-zero entries. Hence, \( (P^*) \) cannot be solved efficiently in practice for the typically large problem sizes of interest. In order to arrive at an optimization problem that can be solved with a reasonable computational complexity, we relax the sparsity constraints \( \| \bar{w}_i \|_0 \leq s \) in \( (P^*) \) to \( \ell_1 \)-norm constraints as in [21] and move them, the \( \ell_2 \)-norm constraints, and the Frobenius norm constraint into the objective function via Lagrange multipliers:

\[
(P) \quad \min_{w_i, c; W_{i,k} \geq 0, \bar{w}_i, k} - \log p(Y_{i,j} | \bar{w}_i, c_j) + \lambda \sum_{i} \| \bar{w}_i \|_1 + \frac{\mu}{2} \sum_{i} \| \bar{w}_i \|_2^2 + \frac{\gamma}{2} \| C \|_F^2.
\]

The first regularization term \( \lambda \sum_{i} \| \bar{w}_i \|_1 \) induces sparsity on each vector \( \bar{w}_i \), with the single parameter \( \lambda > 0 \) controlling the sparsity level. Since one can arbitrarily increase the scale of the vectors \( \bar{w}_i \) while decreasing the scale of the vectors \( c_j \) accordingly (and vice versa) without changing the likelihood, we gauge these vectors using the second and third regularization terms \( \frac{\mu}{2} \sum_i \| \bar{w}_i \|_2^2 \) and \( \frac{\gamma}{2} \| C \|_F^2 \) with the regularization parameters \( \mu > 0 \) and \( \gamma > 0 \), respectively.* We emphasize that since \( \| C \|_F^2 = \)

*The first, \( \ell_1 \)-norm, regularization term already gauges the norm of the \( \bar{w}_i \). Hence, the additional
\[ \sum_j \| c_j \|_2^2, \] we can impose a regularizer on each column rather than the entire matrix \( C \), which facilitates the development of the efficient algorithm detailed below. We finally note that (P) exhibits similarities to the optimization problem for missing data imputation outlined in [22, Eq. 7]; the problem (P) studied here, however, includes an additional non-negativity constraint on \( W \) and the regularization term \( \mu \frac{1}{2} \sum_i \| \bar{w}_i \|_2^2 \).

3.2 The SPARFA-M Algorithm

Since the first negative log-likelihood term in the objective function of (P) is convex in the product \( WC \) for both the probit and the logit functions (see, e.g., [20]), and since the rest of the regularization terms are convex in either \( W \) or \( C \) while the non-negativity constraints on \( W_{i,k} \) are with respect to a convex set, the problem (P) is biconvex in the individual factors \( W \) and \( C \). More importantly, with respect to blocks of variables \( \bar{w}_i, c_j \), the problem (P) is block multi-convex in the sense of [1].

SPARFA-M is an alternating optimization approach to (approximately) solving (P) that proceeds as follows. We initialize \( W \) and \( C \) with random entries and then iteratively optimize the objective function of (P) for both factors in an alternating fashion. Each outer iteration involves solving two kinds of inner subproblems. In the first subproblem, we hold \( W \) constant and separately optimize each block of variables in \( c_j \); in the second subproblem, we hold \( C \) constant and separately optimize each \( \ell_2 \)-norm regularizer \( \frac{\mu}{2} \sum_i \| \bar{w}_i \|_2^2 \) in (P) can be omitted in practice. However, this regularizer is employed in our convergence analysis of SPARFA-M as detailed in Section 3.4.
block of variables $\mathbf{w}_i$. Each subproblem is solved using an iterative method; see Section 3.3 for the respective algorithms. The outer loop is terminated whenever a maximum number of outer iterations $I_{\text{max}}$ is reached, or if the decrease in the objective function of (P) is smaller than a certain threshold.

The two subproblems constituting the inner iterations of SPARFA-M correspond to the following convex $\ell_1/\ell_2$-norm and $\ell_2$-norm regularized regression (RR) problems:

\[(\text{RR}_1^+) \quad \text{minimize} \sum_{j:(i,j) \in \Omega_{\text{obs}}} - \log p(Y_{i,j} | \mathbf{w}_i, \mathbf{c}_j) + \lambda \|\mathbf{w}_i\|_1 + \frac{\mu}{2} \|\mathbf{w}_i\|_2^2, \]

\[(\text{RR}_2) \quad \text{minimize} \sum_{i:(i,j) \in \Omega_{\text{obs}}} - \log p(Y_{i,j} | \mathbf{w}_i, \mathbf{c}_j) + \frac{\gamma}{2} \|\mathbf{c}_j\|_2^2. \]

For the logit link function, one can solve (RR$_1^+$) and (RR$_2$) using an iteratively reweighed second-order algorithm as in [20], [23], [24], [25], or an interior-point method as in [26]. However, none of these techniques extend to the case of the probit link function, which is essential for some applications, e.g., in noisy compressive sensing recovery from 1-bit measurements (e.g., [27, 28]). Moreover, second-order techniques typically do not scale well to high-dimensional problems due to the calculation of the Hessian. Consequently, we develop two novel first-order methods that efficiently solve (RR$_1^+$) and (RR$_2$) for both probit and logit regression while scaling well to high-dimensional problems. We build our algorithm on the fast iterative soft-thresholding algorithm (FISTA) framework developed in [29], which provides accelerated convergence compared to the probit regression methods used in [30] and [22].
3.3 Accelerated First-Order Methods for Regularized Probit/Logit Regression

The FISTA framework ([29]) iteratively solves optimization problems whose objective function is given by $f(\cdot) + g(\cdot)$, where $f(\cdot)$ is a continuously differentiable convex function and $g(\cdot)$ is convex but potentially non-smooth. This approach is particularly well-suited to the inner subproblem (RR$_1^+$) due to the presence of the non-smooth $\ell_1$-norm regularizer and the non-negativity constraint. Concretely, we associate the log-likelihood function plus the $\ell_2$-norm regularizer $\frac{\mu_2}{2}\|\bar{w}_i\|^2_2$ with $f(\cdot)$ and the $\ell_1$-norm regularization term with $g(\cdot)$. For the inner subproblem (RR$_2$), we associate the log-likelihood function with $f(\cdot)$ and the $\ell_2$-norm regularization term with $g(\cdot)$.

Each FISTA iteration consists of two steps: (i) a gradient-descent step in $f(\cdot)$ and (ii) a shrinkage step determined by $g(\cdot)$. For simplicity of exposition, we consider the case where all entries in $Y$ are observed, i.e., $\Omega_{\text{obs}} = \{1, \ldots, Q\} \times \{1, \ldots, N\}$; the extension to the case with missing entries in $Y$ is straightforward. We will derive the algorithm for the case of probit regression first and then point out the departures for logit regression.

For (RR$_1^+$), the gradients of $f(\bar{w}_i)$ with respect to the $i^\text{th}$ block of regression

\[ \frac{\partial f}{\partial \bar{w}_i} \]

\[ \frac{\partial g}{\partial \bar{w}_i} \]

\[ \frac{\partial f}{\partial \bar{w}_i} \] and \[ \frac{\partial f}{\partial \bar{w}_i} \] are smooth in this case. Hence, we could also apply an accelerated gradient-descent approach instead, e.g., as described in [31].
coefficients \( \hat{w}_i \) are given by

\[
\nabla f^i_{\text{pro}} = \nabla_{\hat{w}_i} \left( - \sum_j \log p_{\text{pro}}(Y_{i,j} | \hat{w}_i, c_j) + \frac{\mu}{2} \| \hat{w}_i \|_2^2 \right) = -C \bar{D}^i (\bar{y}^i - \hat{p}^i_{\text{pro}}) + \mu \bar{w}_i, \tag{3.1}
\]

where \( \bar{y}^i \) is an \( N \times 1 \) column vector corresponding to the transpose of the \( i \)th row of \( Y \). \( \hat{p}^i_{\text{pro}} \) is an \( N \times 1 \) vector whose \( j \)th element equals the probability of \( Y_{i,j} \) being 1; that is, \( p_{\text{pro}}(Y_{i,j} = 1 | \hat{w}_i, c_j) = \Phi_{\text{pro}}(\hat{w}_i^T c_j) \). The entries of the \( N \times N \) diagonal matrix are given by

\[
D^i_{j,j} = N(\hat{w}_i^T c_j) \Phi_{\text{pro}}(\hat{w}_i^T c_j)(1 - \Phi_{\text{pro}}(\hat{w}_i^T c_j)).
\]

The gradient step in each FISTA iteration \( \ell = 1, 2, \ldots \) corresponds to

\[
\hat{w}^{\ell+1}_i \leftarrow \hat{w}_i^\ell - t_\ell \nabla f^i_{\text{pro}}, \tag{3.2}
\]

where \( t_\ell \) is a suitable step-size. To comply with (A3), the shrinkage step in (RR\(_1^+\)) corresponds to a non-negative soft-thresholding operation

\[
\bar{w}^{\ell+1}_i \leftarrow \max\{\hat{w}^{\ell+1}_i - \lambda t_\ell, 0\}. \tag{3.3}
\]

For (RR\(_2\)), the gradient step becomes

\[
\hat{c}^{\ell+1}_j \leftarrow c^\ell_j - t_\ell \nabla f^i_{\text{pro}},
\]

which is the same as (3.1) and (3.2) after replacing \( C \) with \( W^T \) and \( \mu \) with \( \gamma \). The shrinkage step for (RR\(_2\)) is the simple re-scaling

\[
c^{\ell+1}_j \leftarrow \frac{1}{1 + \gamma t_\ell} \hat{c}^{\ell+1}_j. \tag{3.4}
\]
In the logit regression case, the steps (3.2), (3.3), and (3.4) remain the same but the gradient changes to

$$\nabla f_{\log}^i = \nabla_{\bar{w}_i}^{\log} \left( - \sum_j \log p_{\log}(Y_{i,j} \mid \bar{w}_i, c_j) + \frac{\mu}{2} \| \bar{w}_i \|_2^2 \right) = -C(y^i - p_{\log}^i) + \mu \bar{w}_i,$$

where the $N \times 1$ vector $p_{\log}^i$ has elements $p_{\log}(Y_{i,j} = 1 \mid \bar{w}_i, c_j) = \Phi^{\log}(\bar{w}_i^T c_j)$.

The above steps require a suitable step-size $t_{\ell}$ to ensure convergence to the optimal solution. A common approach that guarantees convergence is to set $t_{\ell} = 1/L$, where $L$ is the Lipschitz constant of $f(\cdot)$ (see [29] for the details). The Lipschitz constants for both the probit and logit cases are analyzed in Theorem 1 below. Alternatively, one can also perform backtracking, which—under certain circumstances—can be more efficient; see [29, p. 194] for more details.

### 3.4 Convergence Analysis of SPARFA-M

While the SPARFA-M objective function is guaranteed to be non-increasing over the outer iterations ([32]), the factors $W$ and $C$ do not necessarily converge to a global or local optimum due to its biconvex (or more generally, block multi-convex) nature. It is difficult, in general, to develop rigorous statements for the convergence behavior of block multi-convex problems. Nevertheless, we can establish the global convergence of SPARFA-M from any starting point to a critical point of the objective function using recent results developed in [1]. The convergence results below appear to be novel for both sparse matrix factorization as well as dictionary learning.
3.4.1 Convergence Analysis of Regularized Regression using FISTA

In order to establish the SPARFA-M convergence result, we first adapt the convergence results for FISTA in [29] to prove convergence on the two subproblems \((RR^+_1)\) and \((RR_2)\). The following theorem is a consequence of [29, Thm. 4.4] combined with Lemmata 3 and 4 in Appendix A. If back-tracking is used to select step-size \(t_\ell\) [29, p. 194], then let \(\alpha\) correspond to the back-tracking parameter. Otherwise set \(\alpha = 1\) and for \((RR^+_1)\) let \(t_\ell = 1/L_1\) and for \((RR_2)\) let \(t_\ell = 1/L_2\). In Lemma 4, we compute that \(L_1 = \sigma^2_{\max}(C) + \mu\) and \(L_2 = \sigma^2_{\max}(W)\) for the probit case, and \(L_1 = \frac{1}{4}\sigma^2_{\max}(C) + \mu\) and \(L_2 = \frac{1}{4}\sigma^2_{\max}(W)\) for the logit case.

**Theorem 1 (Linear convergence of RR using FISTA)** Given \(i\) and \(j\), let

\[
F_1(\bar{w}_i) = \sum_{j: (i,j) \in \Omega_{obs}} -\log p(Y_{i,j}|\bar{w}_i, c_j) + \lambda \|\bar{w}_i\|_1 + \frac{\mu}{2} \|\bar{w}_i\|_2^2, \quad W_{i,k} \geq 0 \forall k,
\]

\[
F_2(c_j) = \sum_{i: (i,j) \in \Omega_{obs}} -\log p(Y_{i,j}|\bar{w}_i, c_j) + \frac{\gamma}{2} \|c_j\|_2^2
\]

be the cost functions of \((RR^+_1)\) and \((RR_2)\), respectively. Then, we have

\[
F_1(\bar{w}_i^\ell) - F_1(\bar{w}_i^*) \leq \frac{2\alpha L_1 \|\bar{w}_i^0 - \bar{w}_i^*\|^2}{(\ell + 1)^2},
\]

\[
F_2(c_j^\ell) - F_2(c_j^*) \leq \frac{2\alpha L_2 \|c_j^0 - c_j^*\|^2}{(\ell + 1)^2},
\]

where \(\bar{w}_i^0\) and \(c_j^0\) are the initialization points of \((RR^+_1)\) and \((RR_2)\), \(\bar{w}_i^\ell\) and \(c_j^\ell\) designate the solution estimates at the \(\ell\)th inner iteration, and \(\bar{w}_i^*\) and \(c_j^*\) denote the optimal solutions.

In addition to establishing convergence, Theorem 1 reveals that the difference between the cost functions at the current estimates and the optimal solution points,
$F_1(\bar{w}_i) - F_1(\bar{w}_i^*)$ and $F_2(c_j) - F_2(c_j^*)$, decrease as $O(\ell^{-2})$.

### 3.4.2 Convergence Analysis of SPARFA-M

We are now ready to establish global convergence of SPARFA-M to a critical point. To this end, we first define $x = [\bar{w}_1^T, \ldots, \bar{w}_Q^T, c_1^T, \ldots, c_N^T]^T \in \mathbb{R}^{(N+Q)K}$ and rewrite the objective function $(P)$ of SPARFA-M as follows:

$$F(x) = \sum_{(i,j) \in \Omega_{\text{obs}}} - \log p(Y_{i,j}|\bar{w}_i, c_j)$$

$$+ \frac{\mu}{2} \sum_i \|\bar{w}_i\|_2^2 + \lambda \sum_i \|\bar{w}_i\|_1 + \sum_{i,k} \delta(W_{i,k} < 0) + \frac{\gamma}{2} \sum_j \|c_j\|_2^2,$$

where $\delta(z < 0) = \infty$ if $z < 0$ and 0 otherwise. Note that we have re-formulated the non-negativity constraint as a set indicator function and added it to the objective function of $(P)$. Since minimizing $F(x)$ is equivalent to solving $(P)$, we can now use the results developed in [1] to establish the following convergence result for the SPARFA-M algorithm. The proof can be found in Appendix B.

**Theorem 2 (Global convergence of SPARFA-M)** From any starting point $x^0$, let $\{x^t\}$ be the sequence of estimates generated by the SPARFA-M algorithm with $t = 1, 2, \ldots$ as the outer iteration number. Then, the sequence $\{x^t\}$ converges to the finite limit point $\hat{x}$, which is a critical point of $(P)$. Moreover, if the starting point $x^0$ is within a close neighborhood of a global optimum of $(P)$, then SPARFA-M converges to this global optimum.
Since the problem (P) is bi-convex in nature, we cannot guarantee that SPARFA-M always converges to a *global optimum* from an *arbitrary* starting point. Nevertheless, the use of multiple randomized initialization points can be used to increase the chance of being in the close vicinity of a global optimum, which improves the (empirical) performance of SPARFA-M (see Section 3.5 for details). Note that we do not provide the convergence rate of SPARFA-M, since the associated parameters in [1, Thm. 2.9] are difficult to determine for the model at hand; a detailed analysis of the convergence rate for SPARFA-M is part of ongoing work.

### 3.5 Algorithmic Details and Improvements for SPARFA-M

In this section, we outline a toolbox of techniques that improve the empirical performance of SPARFA-M and provide guidelines for choosing the key algorithm parameters.

#### 3.5.1 Reducing Computational Complexity in Practice

To reduce the computational complexity of SPARFA-M in practice, we can improve the convergence rates of (RR$^+_1$) and (RR$^+_2$). In particular, the regularizer $\frac{\mu}{2} \| \overline{\mathbf{w}}_i \|_2^2$ in (RR$^+_1$) has been added to (P) to facilitate the proof for Theorem 2. This term, however, typically slows down the (empirical) convergence of FISTA, especially for large values of $\mu$. We therefore set $\mu$ to a small positive value (e.g., $\mu = 10^{-4}$), which leads to fast convergence of (RR$^+_1$) while still guaranteeing convergence of SPARFA-
Selecting the appropriate (i.e., preferably large) step-sizes $t_\ell$ in (3.2), (3.3), and (3.4) is also crucial for fast convergence. In Lemmata 3 and 4, we derive the Lipschitz constants $L$ for $(RR_1^+)$ and $(RR_2)$, which enables us to set the step-sizes $t_\ell$ to the constant value $t = 1/L$. In all of our experiments below, we exclusively use constant step-sizes, since we observed that backtracking provided no advantage in terms of computational complexity for SPARFA-M.

To further reduce the computational complexity of SPARFA-M without degrading its performance, we have found that instead of running the large number of inner iterations it typically takes to converge, we can run just a few (e.g., 10) inner iterations per outer iteration.

3.5.2 Reducing the Chance of Getting Stuck in Local Minima

The performance of SPARFA-M strongly depends on the initialization of $W$ and $C$, due to the bi-convex nature of $(P)$. We have found that running SPARFA-M multiple times with different starting points and picking the solution with the smallest overall objective function delivers excellent performance. In addition, we can deploy the standard heuristics used in the dictionary-learning literature [33, Section IV-E] to further improve the convergence towards a global optimum. For example, every few outer iterations, we can evaluate the current $W$ and $C$. If two rows of $C$ are similar (as measured by the absolute value of the inner product between them), then we
re-initialize one of them as an i.i.d. Gaussian vector. Moreover, if some columns in $\mathbf{W}$ contain only zero entries, then we re-initialize them with i.i.d. Gaussian vectors. Note that the convergence proof in Section 3.4 does not apply to implementations employing such trickery.

### 3.5.3 Parameter Selection

The input parameters to SPARFA-M include the number of concepts $K$ and the regularization parameters $\gamma$ and $\lambda$. The number of concepts $K$ is a user-specified value. In practice, cross-validation could be used to select $K$ if the task is to predict missing entries of $\mathbf{Y}$ (e.g., [20]). Selecting $\gamma$ is not critical in most situations, since it only gauges the scaling of $\mathbf{W}$ and $\mathbf{C}$, and hence, only needs to be set to an arbitrary (bounded) positive value. The sparsity parameter $\lambda$ strongly affects the output of SPARFA-M; it can be selected using any of a number of criteria, including the Bayesian information criterion (BIC), detailed in [20].
Chapter 4

Tag Analysis: Post-Processing to Interpret the Estimated Concepts

So far we have developed SPARFA-M to estimate $W$, $C$, and $\mu$ (or equivalently, $M$) in (2.2) given the partial binary observations in $Y$. Both $W$ and $C$ encode a small number of latent abstract concepts. As we initially noted, the concepts are "abstract" in that they are estimated from the data rather than dictated by a subject matter expert. In this section we develop a principled way to interpret the true meaning of the abstract concepts, which is important if our results are to be usable for learning analytics and content analytics in practice. Our approach applies when the questions come with a set of user-generated "tags" or "labels" that describe in a free-form manner what ideas underlie each question.

We develop a post-processing algorithm for the estimated matrices $W$ and $C$ that estimates the association between the latent concepts and the user-generated tags. Additionally, we show how to extract a personalized tag knowledge profile for each learner. The efficacy of our tag-analysis framework will be demonstrated in the real-world experiments in Section 5.2.
4.1 Incorporating Question–Tag Information

Suppose that a set of tags has been generated for each question that represent the topic(s) or theme(s) of each question. The tags could be generated by the course instructors, subject matter experts, learners, or, more broadly, by crowd-sourcing. In general, the tags provide a redundant representation of the true knowledge components, i.e., concepts can be associated to a “bag of tags.”

Assume that there is a total number of $M$ tags associated with the $Q$ questions. We form a $Q \times M$ matrix $T$, where each column of $T$ is associated to one of the $M$ pre-defined tags. We set $T_{i,m} = 1$ if tag $m \in \{1, \ldots, M\}$ is present in question $i$ and 0 otherwise. Now, we postulate that the question association matrix $W$ extracted by SPARFA can be further factorized as $W = TA$, where $A$ is an $M \times K$ matrix representing the tags-to-concept mapping. This leads to the following additional assumptions:

- **Non-negativity**: The matrix $A$ is non-negative. This increases the interpretability of the result, since concepts should not be negatively correlated with any tags, in general.

- **Sparsity**: Each column of $A$ is sparse. This ensures that the estimated concepts relate to only a few tags.
4.2 Estimating the Concept–Tag Associations and Learner–Tag Knowledge

The additional assumptions above enable us to extract $\mathbf{A}$ using $\ell_1$-norm regularized non-negative least-squares as described in [20] and [21]. Specifically, to obtain each column $\mathbf{a}_k$ of $\mathbf{A}$, $k = 1, \ldots, K$, we solve the following convex optimization problem, a non-negative variant of basis pursuit denoising:

$$\text{(BPDN)} \quad \min_{\mathbf{a}_k : A_{m,k} \geq 0 \forall m} \frac{1}{2} \| \mathbf{w}_k - \mathbf{T} \mathbf{a}_k \|_2^2 + \eta \| \mathbf{a}_k \|_1.$$  

Here, $\mathbf{w}_k$ represents the $k$th column of $\mathbf{W}$, and the parameter $\eta$ controls the sparsity level of the solution $\mathbf{a}_k$.

We propose a first-order method derived from the FISTA framework in [29] to solve (BPDN$_+\!$). The algorithm consists of two steps: A gradient step with respect to the $\ell_2$-norm penalty function, and a projection step with respect to the $\ell_1$-norm regularizer subject to the non-negative constraints on $\mathbf{a}_k$. By solving (BPDN$_+$) for $k = 1, \ldots, K$, and building $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_K]$, we can (i) assign tags to each concept based on the non-zero entries in $\mathbf{A}$ and (ii) estimate a tag-knowledge profile for each learner.

4.2.1 Associating Tags to Each Concept

Using the concept–tag association matrix $\mathbf{A}$ we can directly associate tags to each concept estimated by the SPARFA algorithms. We first normalize the entries in $\mathbf{a}_k$ such that they sum to one. With this normalization, we can then calculate percentages...
that show the proportion of each tag that contributes to concept $k$ corresponding to the non-zero entries of $a_k$. This concept tagging method typically will assign multiple tags to each concept, thus, enabling one to identify the coarse meaning of each concept (see Section 5.2 for examples using real-world data).

4.2.2 Learner Tag Knowledge Profiles

Using the concept–tag association matrix $A$, we can assess each learner’s knowledge of each tag. To this end, we form an $M \times N$ matrix $U = AC$, where the $U_{m,j}$ characterizes the knowledge of learner $j$ of tag $m$. This information could be used, for example, by a PLS to automatically inform each learner which tags they have strong knowledge of and which tags they do not. Course instructors can use the information contained in $U$ to extract measures representing the knowledge of all learners on a given tag, e.g., to identify the tags for which the entire class lacks strong knowledge. This information would enable the course instructor to select future learning content that deals with those specific tags. A real-world example demonstrating the efficacy of this framework is shown below in Section 5.2.1.
Chapter 5

Experiments

In this section, we validate SPARFA-M on both synthetic and real-world educational data sets. First, using synthetic data, we validate that SPARFA-M can accurately estimate the underlying factors from binary-valued observations and characterize their performance under different circumstances. Specifically, we benchmark the factor estimation performance of SPARFA-M against a variant of the well-established K-SVD algorithm ([33]) used in dictionary-learning applications and also SPARFA-B, a solution to SPARFA based on Bayesian statistics, developed in [34]. Second, using real-world graded learner-response data we demonstrate the efficacy SPARFA-M (both probit and logit variants) for learning and content analytics. Specifically, we showcase how the estimated learner concept knowledge, question–concept association, and intrinsic question difficulty can support machine learning-based personalized learning.

5.1 Synthetic Data Experiments

We first characterize the estimation performance of SPARFA-M using synthetic test data generated from a known ground truth model. We generate instances of $\mathbf{W}$, $\mathbf{C}$, and $\mu$ under pre-defined distributions and then generate the binary-valued observations $\mathbf{Y}$ according to (2.2).
Our report on the synthetic experiments is organized as follows. In Section 5.1.1, we outline K-SVD+, a variant of the well-established K-SVD dictionary-learning (DL) algorithm originally proposed in [33]; we use it as a baseline method for comparison SPARFA-M. In Section 5.1.2 we detail the performance metrics. We compare SPARFA-M to K-SVD+ and SPARFA-B as we vary the problem size and number of concepts (Section 5.1.3), observation incompleteness (Section 5.1.4), and the sparsity of \( W \) (Section 5.1.5). In the above-referenced experiments, we simulate the observation matrix \( Y \) via the inverse probit link function and use only the probit variant of SPARFA-M in order to make a fair comparison with SPARFA-B. In a real-world situation, however, the link function is generally unknown. In Section 5.1.6 we conduct model-mismatch experiments, where we generate data from one link function but analyze assuming the other.

In all synthetic experiments, we average the results of all performance measures over 25 Monte-Carlo trials for each instance of the model parameters we control.

5.1.1 Baseline Algorithm: K-SVD+

Since we are not aware of any existing algorithms to solve (2.2) subject to the assumptions (A1)–(A3), we deploy a novel baseline algorithm based on the well-known K-SVD algorithm of [33], which is widely used in various dictionary learning settings but ignores the inverse probit or logit link functions. Since the standard K-SVD algorithm also ignores the non-negativity constraint used in the SPARFA model, we
develop a variant of the non-negative K-SVD algorithm proposed in [35] that we refer to as K-SVD$_+$. In the sparse coding stage of K-SVD$_+$, we use the non-negative variant of orthogonal matching pursuit (OMP) outlined in [36]; that is, we enforce the non-negativity constraint by iteratively picking the entry corresponding to the maximum inner product without taking its absolute value. We also solve a non-negative least-squares problem to determine the residual error for the next iteration. In the dictionary update stage of K-SVD$_+$, we use a variant of the rank-one approximation algorithm detailed in [35, Figure 4], where we impose non-negativity on the elements in $\mathbf{W}$ but not on the elements of $\mathbf{C}$.

K-SVD$_+$ has as input parameters the sparsity level of each row of $\mathbf{W}$. In what follows, we provide K-SVD$_+$ with the known ground truth for the number of non-zero components in order to obtain its best-possible performance. This will favor K-SVD$_+$ over SPARFA-M, since, in practice, such oracle information is not available.

5.1.2 Performance Measures

In each simulation, we evaluate the performance of all algorithms by comparing the fidelity of the estimates $\hat{\mathbf{W}}, \hat{\mathbf{C}},$ and $\hat{\boldsymbol{\mu}}$ to the ground truth $\mathbf{W}, \mathbf{C},$ and $\boldsymbol{\mu}$. Performance evaluation is complicated by the facts that factor-analysis methods are generally susceptible to permutation of the latent factors. We address this concern by normalizing the columns of $\mathbf{W}, \hat{\mathbf{W}}$ and the rows of $\mathbf{C}, \hat{\mathbf{C}}$ to unit $\ell_2$-norm, permuting the columns of $\hat{\mathbf{W}}$ and $\hat{\mathbf{C}}$ to best match the ground truth, and then compare $\mathbf{W}$ and $\mathbf{C}$ with
the estimates $\hat{W}$ and $\hat{C}$. We also compute the Hamming distance between the support set of $W$ and that of the (column-permuted) estimate $\hat{W}$. To summarize, the performance measures used in the sequel are

\[
E_W = \frac{\|W - \hat{W}\|_F^2}{\|W\|_F^2}, \quad E_C = \frac{\|C - \hat{C}\|_F^2}{\|C\|_F^2},
\]

\[
E_\mu = \frac{\|\mu - \hat{\mu}\|_2^2}{\|\mu\|_2^2}, \quad E_H = \frac{\|H - \hat{H}\|_F^2}{\|H\|_F^2},
\]

(5.1)

where $H \in \{0, 1\}^{Q \times K}$ with $H_{i,k} = 1$ if $W_{i,k} > 0$ and $H_{i,k} = 0$ otherwise. The $Q \times K$ matrix $\hat{H}$ is defined analogously using $\hat{W}$.

5.1.3 Impact of Problem Size and Number of Concepts

In this experiment, we study the performance of SPARFA vs. KSVD$^+$ as we vary the number of learners $N$, the number of questions $Q$, and the number of concepts $K$.

Experimental setup We vary the number of learners $N$ and the number of questions $Q \in \{50, 100, 200\}$, and the number of concepts $K \in \{5, 10\}$. For each combination of $(N, Q, K)$, we generate $W, C, \mu,$ and $Y$ according to [34, Eq. 11] with $v_\mu = 1$, $\lambda_k = 2/3 \forall k$, and $V_0 = I_K$. For each instance, we choose the number of non-zero entries in each row of $W$ as $DU(1,3)$ where $DU(a,b)$ denotes the discrete uniform distribution in the range $a$ to $b$. For each trial, we run the probit version of SPARFA-M, SPARFA-B, and K-SVD$^+$ to obtain the estimates $\hat{W}, \hat{C}, \hat{\mu}$, and calculate $\hat{H}$. For all of the synthetic experiments with SPARFA-M, we set the regularization parameters $\gamma = 0.1$ and select $\lambda$ using the BIC ([20]). Hyperparameters of SPARFA-B are set
according to [34].

**Results and discussion**  Figure 5.1 shows box-and-whisker plots for the three algorithms and the four performance measures. We observe that the performance of all of the algorithms generally improves as the problem size increases. We observe that SPARFA-M outperforms K-SVD+ on $E_W$, $E_C$, and especially $E_\mu$. K-SVD+ performs very well in terms of $E_H$ (slightly better than SPARFA-M) due to the fact that we provide it with the oracle sparsity level, which is, of course, not available in practice. SPARFA-B’s slight improvement in estimation accuracy over SPARFA-M comes at the price of significantly higher computational complexity. For example, for $N = Q = 200$ and $K = 5$, SPARFA-B requires roughly 10 minutes on a 3.2 GHz quad-core desktop PC, while SPARFA-M and K-SVD+ require only 6 s.

In summary, SPARFA-M is destined for analyzing large-scale problems where low computational complexity (e.g., to generate immediate learner feedback) is important. The performance of K-SVD+ is rather poor in comparison and requires knowledge of the true sparsity level.

**5.1.4 Impact of the Number of Incomplete Observations**

In this experiment, we study the impact of the number of observations in $\mathbf{Y}$ on the performance of the probit version of SPARFA-M, SPARFA-B, and K-SVD+. 
Figure 5.1: Performance comparison of SPARFA-M, SPARFA-B, and K-SVD for different problem sizes $Q \times N$ and number of concepts $K$. The performance naturally improves as the problem size increases, while both SPARFA algorithms outperform K-SVD. M denotes SPARFA-M, B denotes SPARFA-B, and K denotes KSVD.
Experimental setup  We set $N = Q = 100$, $K = 5$, and all other parameters as in Section 5.1.3. We then vary the percentage $P_{\text{obs}}$ of entries in $Y$ that are observed as 100%, 80%, 60%, 40%, and 20%. The locations of missing entries are generated i.i.d. and uniformly over the entire matrix.

Results and discussion  Figure 5.2 shows that the estimation performance of all methods degrades gracefully as the percentage of missing observations increases. We conclude that SPARFA-M can reliably estimate the underlying factors, even in cases of highly incomplete data.

5.1.5 Impact of Sparsity Level

In this experiment, we study the impact of the sparsity level in $W$ on the performance of the probit version of SPARFA-M, SPARFA-B, and K-SVD$^+$. 

Experimental setup  We choose the active entries of $W$ i.i.d. $\text{Ber}(q)$ and vary $q \in \{0.23, 0.4, 0.6, 0.8\}$ to control the number of non-zero entries in each row of $W$. 
Figure 5.3: Performance comparison of SPARFA-M, SPARFA-B, and K-SVD+ for different sparsity levels in the rows in $W$. The performance degrades gracefully as the sparsity level increases, while the SPARFA algorithms outperform K-SVD+.

All other parameters are set as in Section 5.1.3. This data-generation method allows for scenarios in which some rows of $W$ contain no active entries as well as all active entries.

**Results and discussion** Figure 5.3 shows that sparser $W$ lead to lower estimation errors. This demonstrates that the SPARFA algorithms are well-suited to applications where the underlying factors have a high level of sparsity. The performance of K-SVD+ is worse than SPARFA-M except on the support estimation error $E_H$, which is due to the fact that K-SVD+ is aware of the oracle sparsity level.

**5.1.6 Impact of Model Mismatch**

In this experiment, we examine the impact of model mismatch by using a link function for estimation that does not match the true link function from which the data is generated.

**Experimental setup** We fix $N = Q = 100$ and $K = 5$, and set all other parameters as in Section 5.1.3. Then, for each generated instance of $W$, $C$ and $\mu$, we generate
Figure 5.4: Performance comparison of SPARFA-M, SPARFA-B, and K-SVD+ with probit/logit model mismatch; $M_P$ and $M_L$ indicate probit and logit SPARFA-M, respectively. In the left/right halves of each box plot, we generate $Y$ according to the inverse probit/logit link functions. The performance degrades only slightly with mismatch, while both SPARFA algorithms outperform K-SVD+.

$Y_{pro}$ and $Y_{log}$ according to both the inverse probit link and the inverse logit link, respectively. We then run SPARFA-M (both the probit and logit variants), SPARFA-B (which uses only the probit link function), and K-SVD+ on both $Y_{pro}$ and $Y_{log}$.

**Results and discussion** Figure 5.4 shows that model mismatch does not severely affect $E_W$, $E_C$, and $E_B$ for SPARFA-M. We also see that K-SVD+ performs worse than both SPARFA methods, since it completely ignores the link function.

### 5.2 Real Data Experiments

We next test the SPARFA algorithms on three real-world educational datasets. Since variants of SPARFA-M with different link functions obtained similar results in the synthetic data experiments in Section 5.1, for the sake of brevity, we will often show the results for only one variant of SPARFA-M for each dataset. In what follows, we select the sparsity penalty parameter $\lambda$ in SPARFA-M using the BIC as described in [20].
5.2.1 Undergraduate DSP course

Dataset  We analyze a very small dataset consisting of \( N = 15 \) learners answering \( Q = 44 \) questions taken from the final exam of an introductory course on digital signal processing (DSP) taught at Rice University in Fall 2011 ([37]). There is no missing data in the matrix \( Y \).

Analysis  We estimate \( W \), \( C \), and \( \mu \) from \( Y \) using the logit version of SPARFA-M assuming \( K = 5 \) concepts to prevent over-fitting and achieve a sufficient concept resolution. Since the questions had been manually tagged by the course instructor, we deploy the tag-analysis approach proposed in Section 4. Specifically, we form a \( 44 \times 12 \) matrix \( T \) using the \( M = 12 \) available tags and estimate the \( 12 \times 5 \) concept–tag association matrix \( A \) in order to interpret the meaning of each retrieved concept. For each concept, we only show the top 3 tags and their relative contributions. We also compute the \( 12 \times 15 \) learner tag knowledge profile matrix \( U \).

Results and discussion  Figure 5.5(a) visualizes the estimated question–concept association matrix \( \hat{W} \) as a bipartite graph consisting of question and concept nodes. In the graph, circles represent the estimated concepts and squares represent question, with thicker edges indicating stronger question–concept associations (i.e., larger entries \( \hat{W}_{i,k} \)). Questions are also labeled with their estimated intrinsic difficulty \( \mu_i \), with larger positive values of \( \mu_i \) indicating easier questions. Note that ten questions are not linked to any concept. All \( Q = 15 \) learners answered these questions correctly;
(a) Question–concept association graph. Circles correspond to concepts and rectangles to questions; the values in each rectangle corresponds to that question’s intrinsic difficulty.

<table>
<thead>
<tr>
<th>Concept 1</th>
<th>Concept 2</th>
<th>Concept 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency response (46%)</td>
<td>Fourier transform (40%)</td>
<td>z-transform (66%)</td>
</tr>
<tr>
<td>Sampling rate (23%)</td>
<td>Laplace transform (36%)</td>
<td>Pole/zero plot (22%)</td>
</tr>
<tr>
<td>Aliasing (21%)</td>
<td>z-transform (24%)</td>
<td>Laplace transform (12%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Concept 4</th>
<th>Concept 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fourier transform (43%)</td>
<td>Impulse response (74%)</td>
</tr>
<tr>
<td>Systems/circuits (31%)</td>
<td>Transfer function (15%)</td>
</tr>
<tr>
<td>Transfer function (26%)</td>
<td>Fourier transform (11%)</td>
</tr>
</tbody>
</table>

(b) Most important tags and relative weights for the estimated concepts.

Figure 5.5: (a) Question–concept association graph and (b) most important tags associated with each concept for an undergraduate DSP course with $N = 15$ learners answering $Q = 44$ questions.
Table 5.1: Selected tag knowledge of Learner 1.

<table>
<thead>
<tr>
<th>z-transform</th>
<th>Impulse response</th>
<th>Transfer function</th>
<th>Fourier transform</th>
<th>Laplace transform</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.09</td>
<td>-1.80</td>
<td>-0.50</td>
<td>0.99</td>
<td>-0.77</td>
</tr>
</tbody>
</table>

Table 5.2: Average tag knowledge of all learners.

<table>
<thead>
<tr>
<th>z-transform</th>
<th>Impulse response</th>
<th>Transfer function</th>
<th>Fourier transform</th>
<th>Laplace transform</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.04</td>
<td>-0.03</td>
<td>-0.10</td>
<td>0.11</td>
<td>0.03</td>
</tr>
</tbody>
</table>

as a result nothing can be estimated about their underlying concept structure. Figure 5.5(b) provides the concept–tag association (top 3 tags) for each of the 5 estimated concepts.

Table 5.1 provides Learner 1’s knowledge of the various tags relative to other learners. Large positive values mean that Learner 1 has strong knowledge of the tag, while large negative values indicate a deficiency in knowledge of the tag. Table 5.2 shows the average tag knowledge of the entire class, computed by averaging the entries of each row in the learner tag knowledge matrix $U$ as described in Section 4.2.2. Table 5.1 indicates that Learner 1 has particularly weak knowledges of the tag “Impulse response.” Armed with this information, a PLS could automatically suggest remediation about this concept to Learner 1. Table 5.2 indicates that the entire class has (on average) weak knowledge of the tag “Transfer function.” With this information, a PLS could suggest to the class instructor that they provide remediation about this concept to the entire class.
5.2.2 Grade 8 science course

**Dataset** The STEMscopes dataset was introduced in Section 1.2. There is substantial missing data in the matrix $Y$, with only 13.5% of its entries observed.

**Results and discussion** Again we see a sparse relationship between questions and concepts in Figure 1.2. The few outlier questions that are not associated with any concept are generally those questions with very low intrinsic difficulty or those questions with very few responses.

5.2.3 Algebra Test Administered on Amazon Mechanical Turk

For a final demonstration of SPARFA-M, we analyze a dataset from a high school algebra test carried out by Daniel Calderón of Rice University on Amazon Mechanical Turk, a crowd-sourcing marketplace ([38]).

**Dataset** The dataset consists of $N = 99$ learners answering $Q = 34$ questions covering topics such as geometry, equation solving, and visualizing function graphs. Calderón manually labeled the questions from a set of $M = 10$ tags. The dataset is fully populated, with no missing entries.

**Analysis** We estimate $W$, $C$, and $\mu$ from the fully populated $34 \times 99$ binary-valued matrix $Y$ using the logit version of SPARFA-M assuming $K = 5$ concepts. We deploy the tag-analysis approach proposed in Section 4 to interpret each concept. Additionally, we calculate the likelihoods of the responses using (2.1) and the estimates $\hat{W}$,
\( \hat{C} \) and \( \hat{\mu} \). The results from SPARFA-M are summarized in Figure 5.6. We detail the results of our analysis for Questions 19–26 in Table 5.3 and for Learner 1 in Table 5.4.

**Results and discussion**  With the aid of SPARFA, we can analyze the strengths and weaknesses of each learner’s concept knowledge both individually and relative to other users. We can also detect outlier responses that are due to guessing, cheating, or carelessness. The values in the estimated concept knowledge matrix \( \hat{C} \) measure each learner’s concept knowledge relative to all other learners. The estimated intrinsic difficulties of the questions \( \hat{\mu} \) provide a relative measure that summarizes how all users perform on each question.

Let us now consider an example in detail; see Table 5.3 and Table 5.4. Learner 1 incorrectly answered Questions 21 and 26 (see Table 5.3), which involve Concepts 1 and 2. Their knowledge of these concepts is not heavily penalized, however (see Table 5.4), due to the high intrinsic difficulty of these two questions, which means that most other users also incorrectly answered them. User 1 also incorrectly answered Questions 24 and 25, which involve Concepts 2 and 4. Their knowledge of these concepts is penalized, due to the low intrinsic difficulty of these two questions, which means that most other users correctly answered them. Finally, Learner 1 correctly answered Questions 19 and 20, which involve Concepts 1 and 5. Their knowledge of these concepts is boosted, due to the high intrinsic difficulty of these two questions.

SPARFA can also be used to identify each user’s individual strengths and weaknesses. Continuing the example, Learner 1 needs to improve their knowledge of
(a) Question–concept association graph.

<table>
<thead>
<tr>
<th>Concept 1</th>
<th>Concept 2</th>
<th>Concept 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fractions (57%)</td>
<td>Plotting functions (64%)</td>
<td>Geometry (63%)</td>
</tr>
<tr>
<td>Solving equations (42%)</td>
<td>System of equations (27%)</td>
<td>Simplifying expressions (27%)</td>
</tr>
<tr>
<td>Arithmetic (1%)</td>
<td>Simplifying expressions (9%)</td>
<td>Trigonometry (10%)</td>
</tr>
</tbody>
</table>

(b) Most important tags and relative weights for the estimated concepts.

Figure 5.6: (a) Question–concept association graph and (b) most important tags associated with each concept for a high-school algebra test carried out on Amazon Mechanical Turk with $N = 99$ users answering $Q = 34$ questions.
<table>
<thead>
<tr>
<th>Question number</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>26</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learner’s graded response</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Correct answer likelihood</td>
<td>0.79</td>
<td>0.71</td>
<td>0.11</td>
<td>0.21</td>
<td>0.93</td>
<td>0.23</td>
<td>0.43</td>
<td>0.00</td>
</tr>
<tr>
<td>Underlying concepts</td>
<td>1</td>
<td>1.5</td>
<td>1</td>
<td>2,3,4</td>
<td>3.5</td>
<td>2,4</td>
<td>1,4</td>
<td>2,4</td>
</tr>
<tr>
<td>Intrinsic difficulty</td>
<td>$-1.42$</td>
<td>$-0.46$</td>
<td>$-0.67$</td>
<td>$0.27$</td>
<td>$0.79$</td>
<td>$0.56$</td>
<td>$1.40$</td>
<td>$-0.81$</td>
</tr>
</tbody>
</table>

Table 5.3: Graded responses and their underlying concepts for Learner 1 (1 designates a correct response and 0 an incorrect response).

<table>
<thead>
<tr>
<th>Concept number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept knowledge</td>
<td>0.46</td>
<td>$-0.35$</td>
<td>0.72</td>
<td>$-1.67$</td>
<td>0.61</td>
</tr>
</tbody>
</table>

Table 5.4: Estimated concept knowledge for Learner 1.

Concept 4 (associated with the tags “Simplifying expressions,” “Trigonometry,” and “Plotting functions”) significantly, while their deficiencies on Concepts 2 and 3 are relatively minor.

Finally, by investigating the likelihoods of the graded responses, we can detect outlier responses, which would enable a PLS to detect guessing and cheating. By inspecting the concept knowledge of Learner 1 in Table 5.4, we can identify insufficient knowledge of Concept 4. Hence, Learner 1’s correct answer to Question 22 is likely due to a random guess, since the predicted likelihood of providing the correct answer is estimated at only 0.21.
Chapter 6

Extension to SPARFA: Ordinal SPARFA-Tag

We have seen from above that, leveraging the latent concepts and based only on the graded binary-valued responses (i.e., correct/incorrect) to a set of questions, SPARFA jointly estimates (i) the associations among the questions and the concepts (via a “concept graph”), (ii) learner concept knowledge profiles, and (iii) the underlying question difficulties. As powerful as SPARFA can be, it also suffers from a couple of drawbacks: First of all, it is only capable of analyzing binary graded learner responses, which is quite restrictive for its application in practice. Secondly, it utilizes a post-process step that associates tags with concepts, neglecting the fact that these instructor-generated tags often reveal rich insights toward the underlying content organization of a course.

In this section, we develop Ordinal SPARFA-Tag, a significant extension to the SPARFA framework that enables the exploitation of the additional information that is often available in educational settings. First, Ordinal SPARFA-Tag exploits the fact that responses are often graded on an ordinal scale (partial credit), rather than on a binary scale (correct/incorrect). Second, Ordinal SPARFA-Tag exploits tags/labels (i.e., keywords characterizing the underlying knowledge component related to a question) that can be attached by instructors and other users to questions.

Exploiting pre-specified tags within the estimation procedure provides significantly
more interpretable question–concept associations. Furthermore, this extended statistical framework can discover new concept–question relationships that would not be in the pre-specified tag information but, nonetheless, explain the graded learner–response data.

We showcase the superiority of Ordinal SPARFA-Tag compared to SPARFA via a set of synthetic “ground truth” simulations and on a variety of experiments with real-world educational datasets. We also demonstrate that Ordinal SPARFA-Tag outperforms existing state-of-the-art collaborative filtering techniques in terms of predicting missing ordinal learner responses.

6.1 Ordinal SPARFA-Tag Statistical Model

In this section, we will first extend the SPARFA framework to characterize ordinal (rather than binary-valued) responses, and then impose additional structure in order to model real-world educational behavior more accurately.

6.1.1 Model for ordinal learner response data

Suppose that we have $N$ learners, $Q$ questions, and $K$ underlying concepts. Following the SPARFA model described in Section 2, For the $i^{th}$ question, with $i \in \{1, \ldots, Q\}$, we propose the following model for the learner–response relationships:

$$ Z_{i,j} = \bar{w}_i^T c_j + \mu_i, \quad \forall (i,j), $$

$$ Y_{i,j} = Q(Z_{i,j} + \epsilon_{i,j}), \quad \epsilon_{i,j} \sim \mathcal{N}(0, 1/\tau_{i,j}), \quad (i,j) \in \Omega_{obs}, \quad (6.1) $$
where the quantity $\epsilon_{i,j}$ models the uncertainty of learner $j$ answering question $i$ and $\mathcal{N}(0, 1/\tau_{i,j})$ denotes a zero-mean Gaussian distribution with precision parameter $\tau_{i,j}$, which models the reliability of the observation of learner $j$ answering question $i$. We will further assume $\tau_{i,j} = \tau$, meaning that all the observations have the same reliability.\(^*\)

In (2.1), $Q(\cdot) : \mathbb{R} \to \mathcal{O}$ is a scalar quantizer that maps a real number into $P$ ordered labels according to

$$Q(x) = p \text{ if } \omega_{p-1} < x \leq \omega_p, \ p \in \mathcal{O},$$

where $\{\omega_0, \ldots, \omega_P\}$ is the set of quantization bin boundaries satisfying $\omega_0 < \omega_1 < \cdots < \omega_{p-1} < \omega_p$, with $\omega_0$ and $\omega_p$ denoting the lower and upper bound of the domain of the quantizer $Q(\cdot)$.\(^\dagger\) This quantization model leads to the equivalent input–output relation

$$Z_{i,j} = \mathbf{w}_i^T \mathbf{c}_j + \mu_i, \ \forall (i,j), \text{ and}$$

$$p(Y_{i,j} = p \mid Z_{i,j}) = \int_{\omega_{p-1}}^{\omega_p} \mathcal{N}(s | Z_{i,j}, 1/\tau_{i,j}) \, ds$$

$$= \Phi(\tau(\omega_p - Z_{i,j})) - \Phi(\tau(\omega_{p-1} - Z_{i,j})), (i,j) \in \Omega_{\text{obs}},$$

where $\Phi(x) = \int_{-\infty}^{x} \mathcal{N}(s | 0, 1) \, ds$ denotes the inverse probit function, with $\mathcal{N}(s | 0, 1)$ representing the value of a standard normal evaluated at $s$.\(^\ddagger\)

\(^*\)Accounting for learner/question-varying reliabilities is straightforward and omitted for the sake of brevity.

\(^\dagger\)In most situations, we have $\omega_0 = -\infty$ and $\omega_p = \infty$.

\(^\ddagger\)Space limitations preclude us from discussing a corresponding logistic-based model; the extension...
We can conveniently rewrite (6.1) and (6.2) in matrix form as

\[
\mathbf{Z} = \mathbf{W}\mathbf{C}, \quad \forall (i,j), \quad \text{and}
\]

\[
p(Y_{i,j} | Z_{i,j}) = \Phi(\tau(U_{i,j} - Z_{i,j}))
\]

\[
- \Phi(\tau(L_{i,j} - Z_{i,j})), \; (i,j) \in \Omega_{\text{obs}},
\]  

where \(\mathbf{Y}\) and \(\mathbf{Z}\) are \(Q \times N\) matrices. The \(Q \times (K + 1)\) matrix \(\mathbf{W}\) is formed by concatenating \([\bar{\mathbf{w}}_1, \ldots, \bar{\mathbf{w}}_Q]^T\) with the intrinsic difficulty vector \(\mathbf{\mu}\) and \(\mathbf{C}\) is a \((K + 1) \times N\) matrix formed by concatenating the \(K \times N\) matrix \([\mathbf{c}_1, \ldots, \mathbf{c}_N]\) with an all-ones row vector \(\mathbf{1}_{1 \times N}\). We furthermore define the \(Q \times N\) matrices \(\mathbf{U}\) and \(\mathbf{L}\) to contain the upper and lower bin boundaries corresponding to the observations in \(\mathbf{Y}\), i.e., we have \(U_{i,j} = \omega_{Y_{i,j}}\) and \(L_{i,j} = \omega_{Y_{i,j}} - 1\), \(\forall (i,j) \in \Omega_{\text{obs}}\).

We emphasize that the statistical model proposed above is significantly more general than the original SPARFA model proposed in Section 2, which is a special case of (6.1) with \(P = 2\) and \(\tau = 1\). The precision parameter \(\tau\) does not play a central role in Section 2 (it has been set to \(\tau = 1\)), since the observations are binary-valued with bin boundaries \(\{-\infty, 0, \infty\}\). For ordinal responses (with \(P > 2\)), however, the precision parameter \(\tau\) significantly affects the behavior of the statistical model and, hence, we estimate the precision parameter \(\tau\) directly from the observed data.

is straightforward.
6.1.2  Additional assumptions

Although the assumptions we made in Section 2.2 are reasonable for a wide range of educational contexts, they are hardly complete. In particular, additional information is often available regarding the questions and the learners in some situations. Hence, we impose one additional assumption:

(A4) Oracle support: Instructor-provided tags on questions provide prior information on some question–concept associations. In particular, associating each tag with a single concept will partially (or fully) determine the locations of the non-zero entries in $W$.

As we will see, assumption (A4) significantly improves the limited interpretability of the estimated factors $W$ and $C$ over the conventional SPARFA framework. By utilizing the tags as “oracle” support information on $W$ within the model, enabling each concept to be associated with a predefined tag directly. Note that user-specified tags might not be precise or complete. Hence, the proposed estimation algorithm must be capable of discovering new question–concept associations and removing predefined associations that cannot be explained from the observed data.

6.2  Algorithm

We start by developing Ordinal SPARFA-M, a generalization of SPARFA-M in Section 3 to ordinal response data. Then, we detail Ordinal SPARFA-Tag, which considers prespecified question tags as oracle support information of $W$, to estimate
\( \mathbf{W}, \mathbf{C}, \) and \( \tau, \) from the ordinal response matrix \( \mathbf{Y} \) while enforcing the assumptions (A1)–(A4).

### 6.2.1 Ordinal SPARFA-M

To estimate \( \mathbf{W}, \mathbf{C}, \) and \( \tau \) in (6.3) given \( \mathbf{Y} \), we maximize the log-likelihood of \( \mathbf{Y} \) subject to (A1)–(A4) by solving (P). Here, the likelihood of each response \( p(Y_{i,j} | \tau \mathbf{w}_i^T \mathbf{c}_j) \) is given by (6.2). In contrast to the Frobenius norm regularizer on \( \mathbf{C} \) imposed in the cost function of (P), here we gauge the norm of the matrix \( \mathbf{C} \) by applying a matrix norm constraint \( \| \mathbf{C} \| \leq \eta \). For example, the Frobenius norm constraint \( \| \mathbf{C} \|_F \leq \eta \) can be used. Alternatively, the nuclear norm constraint \( \| \mathbf{C} \|_* \leq \eta \) can also be used, promoting low-rankness of \( \mathbf{C} \) [39], motivated by the facts that (i) reducing the number of degrees-of-freedom in \( \mathbf{C} \) helps to prevent overfitting to the observed data and (ii) learners can often be clustered into a few groups due to their different demographic backgrounds and learning preferences.

The log-likelihood of the observations in (P) is concave in the product \( \tau \mathbf{w}_i^T \mathbf{c}_j \) [40]. Consequently, the problem (P) is tri-convex, in the sense that the problem obtained by holding two of the three factors \( \mathbf{W}, \mathbf{C}, \) and \( \tau \) constant and optimizing the third one is convex. Therefore, to arrive at a practicable way of solving (P), we propose the following computationally efficient block coordinate descent approach, with \( \mathbf{W}, \mathbf{C}, \) and \( \tau \) as the different blocks of variables.

The matrices \( \mathbf{W} \) and \( \mathbf{C} \) are initialized as i.i.d. standard normal random variables,
and we set \( \tau = 1 \). We then iteratively optimize the objective of (P) for all three factors in round-robin fashion. Each (outer) iteration consists of three phases: first, we hold \( \mathbf{W} \) and \( \tau \) constant and optimize \( \mathbf{C} \); second, we hold \( \mathbf{C} \) and \( \tau \) constant and separately optimize each row vector \( \tilde{\mathbf{w}}_i \); third, we hold \( \mathbf{W} \) and \( \mathbf{C} \) fixed and optimize over the precision parameter \( \tau \). These three phases form the outer loop of Ordinal SPARFA-M.

The sub-problems for estimating \( \mathbf{W} \) and \( \mathbf{C} \) correspond to ordinal regression problems [41], and can be solved by deploying similar iterative first-order methods detailed back in Section 3.3, by replacing the gradient step (3.1) with

\[
\nabla f = \nabla_{\mathbf{w}_i} (- \sum_j \log p(Y_{i,j} | \tau \tilde{\mathbf{w}}_i^T \mathbf{c}_j)) = - \mathbf{Cp}.
\]

(6.4)

Here, \( \mathbf{p} \) is a \( N \times 1 \) vector, with the \( j^{th} \) element equal to

\[
\frac{\mathcal{N}(\tau(U_{i,j}^T \mathbf{c}_j)) - \mathcal{N}(\tau(L_{i,j}^T \mathbf{c}_j))}{\Phi(\tau(U_{i,j}^T \mathbf{c}_j) - \Phi(\tau(L_{i,j}^T \mathbf{c}_j))}
\]

where \( \Phi(\cdot) \) is the inverse probit function. The gradient step for \( \mathbf{C} \) is defined analogously.

The projection step (3.4) for \( \mathbf{C} \) where \( \mathbf{C} \) has norm constraint is given by [32] as

\[
\mathbf{C}^{\ell+1} \left\{ \begin{array}{ll}
\hat{\mathbf{C}}^{\ell+1} & \text{if } \|\hat{\mathbf{C}}^{\ell+1}\|_F \leq \eta \\
\eta \frac{\hat{\mathbf{C}}^{\ell+1}}{\|\hat{\mathbf{C}}^{\ell+1}\|_F} & \text{otherwise}.
\end{array} \right.
\]

(6.5)

For the nuclear-norm constraint \( \|\mathbf{C}\|_* \leq \eta \), the projection step is given by

\[
\mathbf{C}^{\ell+1} \left\{ \begin{array}{ll}
\hat{\mathbf{C}}^{\ell+1} & \text{if } \|\hat{\mathbf{C}}^{\ell+1}\|_* \leq \eta \\
\eta \frac{\hat{\mathbf{C}}^{\ell+1}}{\|\hat{\mathbf{C}}^{\ell+1}\|_*} & \text{otherwise}.
\end{array} \right.
\]

(6.6)
where \( \hat{C}^{t+1} = USV^T \) denotes the singular value decomposition, and \( \text{Proj}_\eta(\cdot) \) is the projection onto the \( \ell_1 \)-ball with radius \( \eta \) (see, e.g., [42] for the details).

The Lipschitz constants for both ordinal regression problems can be computed as \( \tau^2 \sigma_{\text{max}}(C) \) and \( \tau^2 \sigma_{\text{max}}(W) \), with \( \sigma_{\text{max}}(X) \) representing the maximum singular value of \( X \).

To optimize the precision parameter \( \tau \), we compute the solution to

\[
\min_{\tau > 0} -\sum_{i,j:(i,j)\in\Omega_{\text{obs}}} \log(\Phi(\tau(U_{i,j} - \bar{w}_i^T c_j)) - \Phi(\tau(L_{i,j} - \bar{w}_i^T c_j)))
\]

via the secant method [43].

Instead of fixing the quantization bin boundaries \( \{\omega_0, \ldots, \omega_p\} \) introduced in Section 6.1.1 and optimizing the precision and intrinsic difficulty parameters, one can fix \( \tau = 1 \) and optimize the bin boundaries instead, an approach used in, e.g., [44]. We emphasize that optimization of the bin boundaries can also be performed straightforwardly via the secant method, iteratively optimizing each bin boundary while keeping the others fixed. We omit the details for the sake of brevity. Note that we have also implemented variants of Ordinal SPARFA-M that directly optimize the bin boundaries, while keeping \( \tau \) constant; the associated prediction performance is shown in Section 6.3.3.
6.2.2 Ordinal SPARFA-Tag

We now develop the Ordinal SPARFA-Tag algorithm that incorporates (A4). Assume that the total number of tags associated with the \( Q \) questions equal \( K \) (each of the \( K \) concepts correspond to a tag), and define the set \( \Gamma = \{(i, k) : \text{question } i \text{ has tag } k\} \) as the set of indices of entries in \( W \) identified by pre-defined tags, and \( \bar{\Gamma} \) as the set of indices not in \( \Gamma \), we can re-write the optimization problem (P) as:

\[
(P_{\Gamma}) \quad \text{minimize} \quad \sum_{i,j:(i,j) \in \Omega_{\text{obs}}} \log p(Y_{i,j}|\tau \hat{w}_i^T c_j) + \lambda \sum_i \|\hat{w}_i^{(\bar{\Gamma})}\|_1 + \gamma \sum_i \frac{1}{2} \|\hat{w}_i^{(\Gamma)}\|_2^2
\]

subject to \( W \geq 0, \tau > 0, \|C\| \leq \eta \).

Here, \( \hat{w}_i^{(\Gamma)} \) is a vector of those entries in \( \hat{w}_i \) belonging to the set \( \Gamma \), while \( \hat{w}_i^{(\bar{\Gamma})} \) is a vector of entries in \( \hat{w}_i \) not belonging to \( \Gamma \). The \( \ell_2 \)-penalty term on \( \hat{w}_i^{(\Gamma)} \) regularizes the entries in \( W \) that are part of the (predefined) support of \( W \); we set \( \gamma = 10^{-6} \) in all our experiments. The \( \ell_1 \)-penalty term on \( \hat{w}_i^{(\bar{\Gamma})} \) induces sparsity on the entries in \( W \) that are not predefined but might be in the support of \( W \). Reducing the parameter \( \lambda \) enables one to discover new question–concept relationships (corresponding to new non-zero entries in \( W \)) that were not contained in \( \Gamma \).

The problem \( (P_{\Gamma}) \) is solved analogously to the approach described in Section 3.3, except that we split the \( W \) update step into two parts that operate separately on the entries indexed by \( \Gamma \) and \( \bar{\Gamma} \). For the entries in \( \Gamma \), the projection step corresponds to

\[
\hat{w}_i^{(\Gamma),\ell+1} \leftarrow \max\{\hat{w}_i^{(\Gamma),\ell+1}/(1 + \gamma t_{\ell}), 0\}.
\]
Figure 6.1: Performance comparison of Ordinal SPARFA-M vs. K-SVD+. “SP” denotes Ordinal SPARFA-M without given support $\Gamma$ of $W$, “SPP” denotes the variant with estimated precision $\tau$, and “SPT” denotes Ordinal SPARFA-Tag. “KS” stands for K-SVD+, and “KST” denotes its variant with given support $\Gamma$.

The step for the entries indexed by $\bar{\Gamma}$ is given by (3.4).

### 6.3 Ordinal SPARFA-Tag Experiments

We first showcase the performance of Ordinal SPARFA-Tag on synthetic data to demonstrate its convergence to a known ground truth. We then demonstrate the ease of interpretation of the estimated factors by leveraging instructor provided tags in combination with a Frobenius or nuclear norm constraint for two real educational datasets. We finally compare the performance of Ordinal SPARFA-M to state-of-
the-art collaborative filtering techniques on predicting unobserved ordinal learner responses.

### 6.3.1 Synthetic data

As in Section 5.1, we compare Ordinal SPARFA-Tag to K-SVD+. The performance evaluation criterion follows (5.1). We consider both the case when the precision $\tau$ is known a-priori and also when it must be estimated. In all synthetic experiments, the algorithm parameters $\lambda$ and $\eta$ are selected according to Bayesian information criterion (BIC) [20]. All experiments are repeated for 25 Monte-Carlo trials.

We generate the synthetic test data $\mathbf{W}$, $\mathbf{C}$, $\mathbf{\mu}$ as in [34, Eq. 10] with $K = 5$, $\mu_0 = 0$, $v_\mu = 1$, $\lambda_k = 0.66 \forall k$, and $\mathbf{V}_0 = \mathbf{I}_K$. $\mathbf{Y}$ is generated according to (2.2), with $P = 5$ bins and $\{\omega_0, \ldots, \omega_5\} = \{-\infty, -2.1, -0.64, 0.64, 2.1, \infty\}$, such that the entries of $\mathbf{Z}$ fall evenly into each bin. The number of concepts $K$ for each question is chosen uniformly in $\{1, 2, 3\}$. We first consider the impact of problem size on estimation error in Figure 6.2. To this end, we fix $Q = 100$ and sweep $N \in \{50, 100, 200\}$ for $K = 5$ concepts, and then fix $N = 100$ and sweep $Q \in \{50, 100, 200\}$.

**Impact of problem size:** We first study the performance of Ordinal SPARFA-M versus K-SVD+ while varying the problem size parameters $Q$ and $N$. The corresponding box-and-whisker plots of the estimation error for each algorithm are shown in Figure 6.1. In Figure 6.1(a), we fix the number of questions $Q$ and plot the errors $E_C$, $E_W$ and $E_\mu$ for the number of learners $N \in \{50, 100, 200\}$. In Figure 6.1(b), we
Figure 6.2: Performance comparison of Ordinal SPARFA-M vs. K-SVD+ by varying the number of quantization bins. “SP” denotes Ordinal SPARFA-M, “KSY” denotes K-SVD+ operating on $Y$, and “KSZ” denotes K-SVD+ operating on $Z$ in (2.2) (the unquantized data).

We fix the number of learners $N$ and plot the errors $E_C$, $E_W$ and $E_\mu$ for the number of questions $Q \in \{50, 100, 200\}$. It is evident that $E_W$, $E_C$, and $E_\mu$ decrease as the problem size increases for all considered algorithms. Moreover, Ordinal SPARFA-M has superior performance to K-SVD+ in all cases and for all error metrics. Ordinal SPARFA-Tag and the oracle support provided versions of K-SVD outperform Ordinal SPARFA-M and K-SVD+. We furthermore see that the variant of Ordinal SPARFA-M without knowledge of the precision $\tau$ performs as well as knowing $\tau$; this implies that we can accurately learn the precision parameter directly from data.

**Impact of the number of quantization bins:** We now consider the effect of the number of quantization bins $P$ in the observation matrix $Y$ on the performance of our algorithms. We fix $N = Q = 100$, $K = 5$ and generate synthetic data as before up to $Z$ in (6.3). For this experiment, a different number of bins $P$ is used to quantize $Z$ into $Y$. The quantization boundaries are set to $\{\Phi^{-1}(0), \Phi^{-1}(1/P), \ldots, \Phi^{-1}(1)\}$. 
Figure 6.3: Question–concept association graph for a high-school algebra test with $N = 99$ users answering $Q = 34$ questions. Boxes represent questions; circles represent concepts. We furthermore show the unique tag associated with each concept.
Figure 6.4: Question–concept association graph for a grade 8 Earth Science course with $N = 145$ learners answering $Q = 80$ questions ($Y$ is highly incomplete with only 13.5% entries observed). We furthermore show the unique tag associated with each concept.
To study the impact of the number of bins needed for Ordinal SPARFA-M to provide accurate factor estimates that are comparable to algorithms operating with real-valued observations, we also run K-SVD directly on the $Z$ values (recall (6.3)) as a base-line. Figure 6.2 shows that the performance of Ordinal SPARFA-M consistently outperforms K-SVD. We furthermore see that all error measures decrease by about half when using 6 bins, compared to 2 bins (corresponding to binary data). Hence, ordinal SPARFA-M clearly outperforms the conventional SPARFA model [34], when ordinal response data is available. As expected, Ordinal SPARFA-M approaches the performance of K-SVD operating directly on $Z$ (unquantized data) as the number of quantization bins $P$ increases.

6.3.2 Real-world data

We now demonstrate the superiority of Ordinal SPARFA-Tag compared to regular SPARFA as in [34]. In particular, we show the advantages of using tag information directly within the estimation algorithm and of imposing a nuclear norm constraint on the matrix $C$. For all experiments, we apply Ordinal SPARFA-Tag to the graded learner response matrix $Y$ with oracle support information obtained from instructor-provided question tags. The parameters $\lambda$ and $\eta$ are selected via cross-validation.

**Algebra test:** We analyze a dataset from a high school algebra test carried out on Amazon Mechanical Turk [38], a crowd-sourcing marketplace. The dataset consists of $N = 99$ users answering $Q = 34$ multiple-choice questions covering topics
such as geometry, equation solving, and visualizing function graphs. The questions were manually labeled with a set of 13 tags. The dataset is fully populated, with no missing entries. A domain expert manually mapped each possible answer to one of \( P = 4 \) bins, i.e., assigned partial credit to each choice as follows: totally wrong \((p = 1)\), wrong \((p = 2)\), mostly correct \((p = 3)\), and correct \((p = 4)\).

Figure 6.3 shows the question–concept association map estimated by Ordinal SPARFA-Tag using the Frobenius norm constraint \( \|C\|_F \leq \eta \). Circles represent concepts, and squares represent questions (labelled by their intrinsic difficulty \( \mu_i \)). Large positive values of \( \mu_i \) indicate easy questions; negative values indicate hard questions. Connecting lines indicate whether a concept is present in a question; thicker lines represent stronger question–concept associations. Black lines represent the question–concept associations estimated by Ordinal SPARFA-Tag, corresponding to the entries in \( W \) as specified by \( \Gamma \). Red, dashed lines represent the “mislabeled” associations (entries of \( W \) in \( \Gamma \)) that are estimated to be zero. Green solid lines represent new discovered associations, i.e., entries in \( W \) that were not in \( \Gamma \) that were discovered by Ordinal SPARFA-Tag.

By comparing Figure 6.3 with Figure 5.6, we can see that Ordinal SPARFA-Tag provides unique concept labels, i.e., one tag is associated with one concept; this enables precise interpretable feedback to individual learners, as the values in \( C \) represent directly the tag knowledge profile for each learner. This tag knowledge profile can be used by a PLS to provide targeted feedback to learners.
The estimated question–concept association matrix can also serve as a useful tool to domain experts or course instructors, as they indicate missing and inexistent tag–question associations.

**Grade 8 Earth Science course:** As a second example of Ordinal SPARFA-Tag, we analyze a Grade 8 Earth Science course dataset [19]. This dataset contains $N = 145$ learners answering $Q = 80$ questions and is highly incomplete (only 13.5% entries of $Y$ are observed). The matrix $Y$ is binary-valued; domain experts labeled all questions with 16 tags.

The result of Ordinal SPARFA-Tag with the nuclear norm constraint $\|C\|_* \leq \eta$ on $Y$ is shown in Figure 6.4. The estimated question–concept associations mostly matches those pre-defined by domain experts. Note that our algorithm identified some question–concept associations to be non-existent (indicated with red dashed lines). Moreover, no new associations have been discovered, verifying the accuracy of the pre-specified question tags from domain experts. Comparing to the question–concept association graph of the high school algebra test in Figure 6.3, we see that for this dataset, the pre-specified tags represent disjoint knowledge components, which is indeed the case in the underlying question set. Interestingly, the estimated concept matrix $C$ has rank 3; note that we are estimating $K = 13$ concepts. This observation suggests that all learners can be accurately represented by a linear combination of only 3 different “eigen-learner” vectors. Further investigation of this clustering phenomenon is part of on-going research.
Figure 6.5: Prediction performance on the Mechanical Turk algebra test dataset. We compare the collaborative filtering methods SVD++ and OrdRec to various Ordinal SPARFA-M based methods: “Nuc” uses the nuclear norm constraint, “Fro” uses the Frobenius norm constraint, “Bin” and “BinInd” learn the bin boundaries, whereas “Bin” learns one set of bin boundaries for the entire dataset and “BinInd” learns individual bin boundaries for each question.

6.3.3 Predicting unobserved learner responses

We now compare the prediction performance of ordinal SPARFA-M on unobserved learner responses against state-of-the-art collaborative filtering techniques: (i) SVD++ in [45], which treats ordinal values as real numbers, and (ii) OrdRec in [44], which relies on an ordinal logit model. We compare different variants of Ordinal SPARFA-M: (i) optimizing the precision parameter, (ii) optimizing a set of bins for all learners, (iii) optimizing a set of bins for each question, and (iv) using the nu-
clear norm constraint on $C$. We consider the Mechanical Turk algebra test, hold out 20% of the observed learner responses as test sets, and train all algorithms on the rest. The regularization parameters of all algorithms are selected using 4-fold cross-validation on the training set. Figure 6.5 shows the root mean square error (RMSE) 
\[ \sqrt{\frac{1}{|\Omega_{\text{obs}}|} \sum_{(i,j) \in \Omega_{\text{obs}}} \| \hat{Y}_{i,j} - Y_{i,j} \|_2^2} \] where $\hat{Y}_{i,j}$ is the predicted score for $Y_{i,j}$, averaged over 50 trials.

Figure 6.5 demonstrates that the nuclear norm variant of Ordinal SPARFA-M outperforms OrdRec, while the performance of other variants of ordinal SPARFA are comparable to OrdRec. SVD++ performs worse than all compared methods, suggesting that the use of a probabilistic model considering ordinal observations enables accurate predictions on unobserved responses.

We furthermore observe that the variants of Ordinal SPARFA-M that optimize the precision parameter or bin boundaries deliver almost identical performance.

We finally emphasize that Ordinal SPARFA-M not only delivers superior prediction performance over the two state-of-the-art collaborative filtering techniques in predicting learner responses, but it also provides interpretable factors, which is key in educational applications.
Chapter 7

Extension to SPARFA: SPARFA-Top

The SPARFA framework described above in Section 2 generates “abstract” latent concepts in the sense that these concepts are estimated from the data rather than dictated by a subject matter expert, providing limited interpretability to these concepts. Ordinal SPARFA-Tag framework introduced in Section 6 improves the interpretability of the estimated concept utilizing tags on questions provided by instructors. However, requiring domain experts to label the questions with tags is an obvious limitation to the approach, since such tags are often incomplete or inaccurate and thus provide insufficient or unreliable information.

Inspired by the recent success of modern text processing algorithms, such as latent Dirichlet allocation (LDA) [46], we posit that the text associated with each question can potentially reveal the meaning of the estimated latent concepts without the need of instructor-provided question tags. This approach would be advantageous as it would easily scale to domains with thousands of questions. Furthermore, directly incorporating textual information into the SPARFA statistical model (2.1) could potentially improve the estimation performance of the approach.

Therefore, in this section, we propose SPARFA-Top, which extends the SPARFA framework to jointly analyze both graded learner responses to questions and the text
of the question, response, or feedback. In addition to SPARFA outputs, SPARFA-Top also generates a list of most important keywords associated with each estimated concept, enabling a PLS to automatically generate a human readable interpretation for each estimated concept in a purely data-driven fashion.

In literature, the joint analysis of binary-valued data and associated textual information has been studied in the context of congressional voting patterns using Bayesian inference methods [47, 48]. However, these methods build on Bayesian data analysis and do not scale to large datasets. More computationally efficient optimization-based topic model methods have been proposed in [49, 50]. None of these methods, however, consider the joint analysis of textual information and other forms of observed data (such as graded learner responses).

### 7.1 The SPARFA-Top Model

We now introduce a novel probabilistic model to jointly consider graded learner response and associated textual information, in order to directly associate keywords with the estimated concepts.

Assume that we observe the word–question occurrence matrix $B \in \mathbb{N}^{Q \times V}$, where $V$ corresponds to the size of the vocabulary, i.e., the number of unique words that have occurred among the $Q$ questions. Each entry $B_{i,v}$ represents how many times the $v^{th}$ word occurs in the associated text of the $i^{th}$ question; as is typical in the topic model literature, common stop words (“the”, “and”, “in” etc.) are excluded from the
vocabulary. The word occurrences in $\mathbf{B}$ are modeled as follows:

$$ A_{i,v} = \mathbf{w}_i^T \mathbf{t}_v \quad \text{and} \quad B_{i,v} \sim Pois(A_{i,v}), \quad \forall i,v, $$

(7.1)

where $\mathbf{t}_v \in \mathbb{R}_+^K$ is a non-negative column vector that characterizes the expression of the $v^{th}$ word in every concept. Inspired by the topic model proposed in [50], the entries of the word-occurrence matrix $B_{i,v}$ in (7.1) are assumed to be Poisson distributed, with rate parameters $A_{i,v}$. In order to balance between observed graded learner response data and question text, we extend the SPARFA model for graded responses to:

$$ Z_{i,j} = \mathbf{w}_i^T \mathbf{c}_j + \mu_i, \quad \forall i,j, \quad \text{and} $$

$$ Y_{i,j} \sim Ber(\Phi(\tau_{i,j} Z_{i,j})), \quad (i,j) \in \Omega_\text{obs}, $$

(7.2)

The precision parameter $\tau_{i,j}$ models the reliability of the observed binary graded response $Y_{i,j}$. Larger values of $\tau_{i,j}$ indicate higher reliability on the observed graded learner responses, while smaller values indicate lower reliability.

We emphasize that the models (2.1) and (7.1) share the same question–concept association vector, which implies that the relationships between questions and concepts manifested in the learner responses are assumed to be exactly the same as the question–topic relationships expressed as word co-occurrences. Consequently, the question–concept associations generating the question–associated text are also sparse

*Since the Poisson rate $A_{i,v}$ must be strictly positive, we assume that $A_{i,v} \geq \varepsilon$ with $\varepsilon = 10^{-6}$ in all experiments.
and non-negative, coinciding with the standard assumptions made in the topic model literature [46, 51].

7.2 SPARFA-Top algorithm

We now develop the SPARFA-Top algorithm by using block multi-convex optimization, to jointly estimate \( W, C, \mu, \) and \( T = [t_1, \ldots, t_V] \) from the observed student–response matrix \( Y \) and the word-frequency matrix \( B \). Specifically, we seek to solve the following optimization problem:

\[
\begin{align*}
\text{minimize} & \quad W, C, T : W_{i,k} \geq 0 \forall i,k, C_{k,v} \geq 0 \forall k,v, T_k \geq 0 \forall k, v \\
& \sum_{(i,j) \in \Omega_{obs}} - \log p(Y_{i,j} | \bar{w}_i^T c_j + \mu_i, \tau) + \sum_{i,v} - \log p(B_{i,v} | \bar{w}_i^T t_v) \\
& + \lambda \sum_i \| \bar{w}_i \|_1 + \frac{\gamma}{2} \sum_j \| c_j \|_2^2 + \frac{\eta}{2} \sum_v \| t_v \|_2^2. 
\end{align*}
\]  

(7.3)

Here, the probabilities \( p(Y_{i,j} | \bar{w}_i^T c_j + \mu_i, \tau) \) and \( p(B_{i,v} | \bar{w}_i^T t_v) \) follow the statistical models in (7.2) and (7.1), respectively. As SPARFA, (7.3) is solved by a block coordinate descent approach as in Section 3, due to the tri-convex nature of the cost function.

The subproblem of optimizing over \( C \) with \( W \) and \( T \) fixed was detailed in Section 3. The subproblem of optimizing over \( T \) with \( W \) and \( C \) fixed is separable in each column of \( T \), with the problem for \( t_v \) being:

\[
\begin{align*}
\text{minimize} & \quad t_v : T_{k,v} \geq 0, \forall k, v \\
& \sum_i - \log p(B_{i,v} | \bar{w}_i^T t_v) + \frac{\eta}{2} \| t_v \|_2^2.
\end{align*}
\]  

(7.4)
The gradient of the objective function with respect to \( t_v \) is:

\[
\nabla_{t_v} \sum_i - \log p(B_{i,v} | \bar{w}_i^T t_v) + \frac{\eta}{2} \| t_v \|^2 = W^T r + \eta t_v, \tag{7.5}
\]

where \( r \) is a \( Q \times 1 \) vector with its \( i \)th element being 
\[
r_i = 1 - \frac{B_{i,v} \bar{w}_i^T t_v}{\bar{w}_i^T t_v}. \]

By setting this gradient to zero, we obtain the close form solution 
\( t_v = (W^T W + \eta I)^{-1} W^T b_v \), where 
\( b_v \) denotes the \( v \)th column of \( B \).

The subproblem of optimizing over \( W \) with \( C \) and \( T \) fixed is also separable in each row of \( W \). The problem for each \( \bar{w}_i \) is:

\[
\text{minimize}_{\bar{w}_i, \bar{w}_i, j \geq \theta_j} \sum_{j: (i,j) \in \Omega_{obs}} - \log p(Y_{i,j} | \bar{w}_i^T c_j + \mu_i, \tau) + \sum_v - \log p(B_{i,v} | \bar{w}_i^T t_v) + \lambda \| \bar{w}_i \|_1, \tag{7.6}
\]

which can be efficiently solved using FISTA. Specifically, analogous to (3.5), the gradient of the smooth part of the objective function with respect to \( \bar{w}_i \) corresponds to:

\[
\nabla_{\bar{w}_i} \left( \sum_{j: (i,j) \in \Omega_{obs}} - \log p(Y_{i,j} | \bar{w}_i^T c_j + \mu_i, \tau) + \sum_v - \log p(B_{i,v} | \bar{w}_i^T t_v) \right) = -C^T (y_i - p) + T^T s, \tag{7.7}
\]

where \( y_i \) represents the transpose of the \( i \)th row of \( Y \), \( p \) represents a \( N \times 1 \) vector with \( p_j = 1/(1 + e^{-\bar{w}_i^T c_j}) \) as its \( j \)th element, and \( s \) is a \( V \times 1 \) vector with \( s_v = 1 - \frac{B_{i,v} \bar{w}_i^T t_v}{\bar{w}_i^T t_v} \) as its \( v \)th element. The projection step is a soft-thresholding operation, as detailed in (3.4). The step-sizes are chosen via back-tracking line search as described in [32].

Note that we treat \( \tau \) as a fixed parameter. Alternatively, one could estimate this
parameter *within* the algorithm by introducing an additional step that optimizes over $\tau$. A throughout analysis of this approach is left for future work.

### 7.3 Experiments

We now demonstrate the efficacy of SPARFA-Top on both the STEMscopes dataset and the algebra test dataset. The question–associated text vocabulary for the STEMscopes dataset consists of 326 words, excluding common stop-words. The algebra test dataset consist vocabulary consist of 13 predefined keywords associated to each question.

The regularization parameters $\lambda$, $\gamma$ and $\eta$, together with the precision parameter $\tau$ of SPARFA-Top, are selected via cross-validation. In Figure 7.3, we show the prediction likelihood defined by $p(Y_{i,j}|\hat{w}_i^Tc_j + \mu_i, \tau), (i,j) \in \Omega_{\text{obs}}$ for SPARFA-Top on 20% holdout entries in $Y$ and for varying precision values $\tau$. We see that textual information can slightly improve the prediction performance of SPARFA-Top over SPARFA (which corresponds to $\tau \to \infty$), for both the STEMscopes dataset and the algebra test dataset. The reason for (albeit slightly) improving the prediction performance is the fact that textual information reveals additional structure underlying a given test/assessment.

Figures 7.1 and 7.2 show the question–concept association graphs along with the recovered intrinsic difficulties, as well as the top three words characterizing each concept. Compared to SPARFA (Figure 1.2 and Figure 5.6), we observe that SPARFA-
Top is able to relate all questions to concepts, including those questions that were found to be unrelated to any concept. Furthermore, Figures 7.1 and 7.2 demonstrate that SPARFA-Top is capable of automatically generating an interpretable summary of the true meaning of each concept.
Figure 7.1: Question–concept association graph and most important keywords recovered by SPARFA-Top for the STEMscopes dataset; boxes represent questions, circles represent concepts, and thick lines represent strong question–concept associations.
Table:

<table>
<thead>
<tr>
<th>Concept 1</th>
<th>Concept 2</th>
<th>Concept 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solving equations</td>
<td>Slope</td>
<td>Arithmetic</td>
</tr>
<tr>
<td>Quadratic function</td>
<td>Fractions</td>
<td>Trigonometry</td>
</tr>
</tbody>
</table>
| Fractions          | Simplifying expressions | System of equations |}

<table>
<thead>
<tr>
<th>Concept 4</th>
<th>Concept 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simplifying expressions</td>
<td>Inequality</td>
</tr>
<tr>
<td>Factoring polynomials</td>
<td>Plotting functions</td>
</tr>
<tr>
<td>Fractions</td>
<td>Geometry</td>
</tr>
</tbody>
</table>

Figure 7.2: Question–concept association graph and 3 most important keyword recovered by SPARFA-Top for the algebra test dataset; boxes represent questions, circles represent concepts, and thick lines represent strong question–concept associations.
Figure 7.3: Average predicted likelihood on 20% hold-out data in $Y$ using SPARFA-Top with different precision parameters $\tau$. For $\tau \to \infty$ SPARFA-Top corresponds to SPARFA.
A range of different machine learning algorithms have been applied in educational contexts. Bayesian belief networks have been successfully used to probabilistically model and analyze learner response data (e.g., [52, 53, 54]). Such models, however, rely on predefined question–concept dependencies (that are not necessarily the true dependencies governing learner responses) and primarily only work for a single concept. In contrast, SPARFA discovers question–concept dependencies from solely the graded learner responses to questions and naturally estimates multi-concept question dependencies.

Modeling question–concept associations has been studied in [55], [56], [57], and [58]. The approach in [55] characterizes the underlying question–concept associations using binary values, which ignore the relative strengths of the question–concept associations. In contrast, SPARFA differentiates between strong and weak relationships through the real-valued weights $W_{i,k}$. The matrix and tensor factorization methods proposed in [55], [56], and [57] treat graded learner responses as real but deterministic values. In contrast, the probabilistic framework underlying SPARFA provides a statistically principled model for graded responses; the likelihood of the observed
graded responses provides even more explanatory power.

Existing intelligent tutoring systems capable of modeling question–concept relations probabilistically include Khan Academy ([59, 60]) and the system of [61]. Both approaches, however, are limited to dealing with a single concept. In contrast, SPARFA is built from the ground up to deal with multiple latent concepts.

A probit model for graded learner responses is used in [58] without exploiting the idea of low-dimensional latent concepts. In contrast, SPARFA leverages multiple latent concepts and therefore can create learner concept knowledge profiles for personalized feedback. Moreover, SPARFA-M is compatible with the popular logit model.

The recent results developed in [62] and [63] address the problem of predicting the missing entries in a binary-valued graded learner response matrix. Both papers use low-dimensional latent factor techniques specifically developed for collaborative filtering, as, e.g., discussed in [64] and [65]. While predicting missing correctness values is an important task, these methods do not take into account the sparsity and non-negativity of the matrix $W$; this inhibits the interpretation of the relationships among questions and concepts. In contrast, SPARFA accounts for both the sparsity and non-negativity of $W$, which enables the interpretation of the value $C_{k,j}$ as learner $j$’s knowledge of concept $k$.

There is a large body of work on item response theory (IRT), which uses statistical models to analyze and score graded question response data (see, e.g., [66], [67], and
The main body of the IRT literature builds on the model developed by [69] and has been applied mainly in the context of adaptive testing (e.g., in the graduate record examination (GRE) and graduate management (GMAT) tests [70], [71], and [72]).

Our proposed statistical model shares some similarity to the Rasch model [69], the additive factor model [73], learning factor analysis [74]. The capability of SPARFA to model each learner in terms of a multi-dimensional concept knowledge vector is in stark contrast to the Rasch model, where each learner is characterized by a single, scalar ability parameter. Consequently, the SPARFA framework is able to provide stronger explanatory power in the estimated factors compared to that of the conventional Rasch model. Moreover, these models rely on pre-defined question features, do not support disciplined algorithms to estimate the model parameters solely from learner response data, or do not produce interpretable estimated factors. We finally note that multi-dimensional variants of IRT have been proposed in [75], [76], and [68]. We emphasize, however, that the design of these algorithms leads to poor interpretability of the resulting parameter estimates.
Chapter 9

Conclusions

In this thesis, we have formulated a new approach to learning and content analytics, which is based on a new statistical model that encodes the probability that a learner will answer a given question correctly in terms of three factors: (i) the learner’s knowledge of a set of latent concepts, (ii) how the question related to each concept, and (iii) the intrinsic difficulty of the question. We have proposed SPARFA-M, an efficient bi-convex optimization approach to produce maximum likelihood point estimates of the factors, to estimate the above three factors given incomplete observations of graded learner question responses. We have also introduced a novel method for incorporating user-defined tags on questions to facilitate the interpretability of the estimated factors. Experiments with both synthetic and real world education datasets have demonstrated both the efficacy and robustness of SPARFA.

The quantities estimated by SPARFA can be used directly in a range of PLS functions. For instance, we can identify the knowledge level of learners on particular concepts and diagnose why a given learner has incorrectly answered a particular question or type of question. Moreover, we can discover the hidden relationships among questions and latent concepts, which is useful for identifying questions that do and do not aid in measuring a learner’s conceptual knowledge. Outlier responses
that are either due to guessing or cheating can also be detected. In concert, these functions can enable a PLS to generate personalized feedback and recommendation of study materials, thereby enhancing overall learning efficiency.

Before closing, we point out a connection between SPARFA and dictionary learning that is of independent interest. This connection can be seen by noting that (2.2) for both the probit and inverse logit functions is statistically equivalent to (see [17]):

\[ Y_{i,j} = \text{sign}(WC + M + N)_{i,j}, \quad (i,j) \in \Omega_{\text{obs}}, \]

where \( \text{sign}(\cdot) \) denotes the entry-wise sign function and the entries of \( N \) are i.i.d. and drawn from either a standard Gaussian or standard logistic distribution. Hence, estimating \( W, C, \) and \( M \) (or equivalently, \( \mu \)) is equivalent to learning a (possibly overcomplete) dictionary from the data \( Y \). The key departures from the dictionary-learning literature ([33, 77]) and algorithm variants capable of handling missing observations ([78]) are the binary-valued observations and the non-negativity constraint on \( W \). Note that the algorithms developed in Section 3 to solve the sub-problems by holding one of the factors \( W \) or \( C \) fixed and solving for the other variable can be used to solve noisy binary-valued (or 1-bit) compressive sensing or sparse signal recovery problems, e.g., as studied in [79], [27], and [28]. Thus, the SPARFA algorithms can be applied to a wide range of applications beyond education, including the analysis of survey data, voting patterns, gene expression, and signal recovery from noisy 1-bit compressive measurements.
Bibliography


[37] ELEC 301, Rice University, “Introduction to signals and systems,” May 2011.


Appendix A

Proof of Theorem 1

We now establish the convergence of the FISTA algorithms that solve the SPARFA-M subproblems (RR)\textsuperscript{1} and (RR)\textsubscript{2}. We start by deriving the relevant Lipschitz constants.

Lemma 3 (Scalar Lipschitz constants) Let \( g_{\text{pro}}(x) = \frac{\Phi_{\text{pro}}(x)}{\Phi_{\text{pro}}(x)} \) and \( g_{\text{log}}(x) = \frac{\Phi_{\text{log}}(x)}{\Phi_{\text{log}}(x)} \), \( x \in \mathbb{R} \), where \( \Phi_{\text{pro}}(x) \) and \( \Phi_{\text{log}}(x) \) are the inverse probit and logit link functions defined in (2.3) and (2.4), respectively. Then, for \( y, z \in \mathbb{R} \) we have

\[
|g_{\text{pro}}(y) - g_{\text{pro}}(z)| \leq L_{\text{pro}}|y - z|, \tag{A.1}
\]

\[
|g_{\text{log}}(y) - g_{\text{log}}(z)| \leq L_{\text{log}}|y - z|, \tag{A.2}
\]

with the constants \( L_{\text{pro}} = 1 \) for the probit case and \( L_{\text{log}} = 1/4 \) for the logit case.

Proof For simplicity of exposition, we omit the subscripts designating the probit and logit cases in what follows. We first derive \( L_{\text{pro}} \) for the probit case by computing the derivative of \( g(x) \) and bounding its derivative from below and above. The derivative of \( g(x) \) is given by

\[
g'(x) = -\frac{\mathcal{N}(x)}{\Phi(x)} \left( x + \frac{\mathcal{N}(x)}{\Phi(x)} \right), \tag{A.3}
\]

where \( \mathcal{N}(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \) is the PDF of the standard normal distribution.

We first bound this derivative for \( x \leq 0 \). To this end, we individually bound the first and second factor in (A.3) using the following bounds listed in [80]:

\[
-\frac{x}{2} + \sqrt{\frac{x^2}{4} + \frac{2}{\pi}} \leq \frac{\mathcal{N}(x)}{\Phi(x)} \leq -\frac{x}{2} + \sqrt{\frac{x^2}{4} + 1}, \quad x \leq 0
\]
and
\[ \frac{x}{2} + \sqrt{\frac{x^2}{4} + \frac{2}{\pi}} \leq x + \frac{\mathcal{N}(x)}{\Phi(x)} \leq \frac{x}{2} + \sqrt{\frac{x^2}{4} + 1}, \quad x \leq 0. \]

Multiplying the above inequalities leads to the bounds
\[ -1 \leq g'(x) \leq -\frac{2}{\pi}, \quad x \leq 0. \tag{A.4} \]

We next bound the derivative of (A.3) for \( x > 0 \). For \( x > 0 \), \( \mathcal{N}(x) \) is a positive decreasing function and \( \Phi(x) \) is a positive increasing function; hence \( \frac{\mathcal{N}(x)}{\Phi(x)} \) is a decreasing function and \( \frac{\mathcal{N}(x)}{\Phi(x)} \leq \frac{\mathcal{N}(0)}{\Phi(0)} = \sqrt{2/\pi} \). Thus, we arrive at
\[ g'(x) = -\frac{\mathcal{N}(x)}{\Phi(x)} \left( x + \frac{\mathcal{N}(x)}{\Phi(x)} \right) \geq -\frac{\mathcal{N}(x)}{\Phi(x)} \left( x + \sqrt{2/\pi} \right), \]

where we have used the facts that \( \Phi(x) \geq 1/2 \) and \( \mathcal{N}(x) \leq \frac{1}{\sqrt{2\pi}} \) for \( x > 0 \). According to (2.3) and the bound of [81], we have
\[ \Phi(x) = \frac{1}{2} + \int_0^x \mathcal{N}(t \mid 0, 1)dt \geq \frac{1}{2} + \frac{1}{2} \sqrt{1 - e^{-x^2/2}} \geq 1 - \frac{1}{2} e^{-x^2/2}, \tag{A.5} \]

where the second inequality follows from the fact that \( (1 - e^{-x^2/2}) \in [0, 1] \). Using (A.5) we can further bound \( g'(x) \) from below as
\[ g'(x) \geq -\frac{\mathcal{N}(x)}{1 - \frac{1}{2} e^{-x^2/2}} \left( x + \sqrt{2/\pi} \right). \]

Let us now assume that
\[ -\frac{\mathcal{N}(x)}{1 - \frac{1}{2} e^{-x^2/2}} \left( x + \sqrt{2/\pi} \right) \geq -1. \]

In order to prove that this assumption is true, we rearrange terms to obtain
\[ \left( \frac{x}{\sqrt{2\pi}} + \left(1/\pi + 1/2\right) \right) e^{-x^2/2} \leq 1. \tag{A.6} \]

Now we find the maximum of the LHS of (A.6) for \( x > 0 \). To this end, we observe that \( \frac{x}{\sqrt{2\pi}} + \left(1/\pi + 1/2\right) \) is monotonically increasing and that \( e^{-x^2/2} \) monotonically
decreasing for $x > 0$; hence, this function has a unique maximum in this region. By taking its derivative and setting it to zero, we obtain

$$x^2 + \sqrt{2/\pi} + \sqrt{\pi/2} - 1 = 0$$

Substituting the result of this equation, i.e., $\hat{x} \approx 0.4068$, into (A.6) leads to

$$\left(\frac{\hat{x}}{\sqrt{2\pi}} + (1/\pi + 1/2)\right) e^{-\hat{x}^2/2} \approx 0.9027 \leq 1,$$

which certifies our assumption. Hence, we have

$$-1 \leq g'(x) \leq 0, \quad x > 0.$$

Combining this result with the one for $x \leq 0$ in (A.4) yields

$$-1 \leq g'(x) \leq 0, \quad x \in \mathbb{R}.$$  

We finally obtain the following bound on the scalar Lipschitz constant (A.1):

$$|g_{\text{pro}}(y) - g_{\text{pro}}(z)| \leq \left| \int_y^z |g_{\text{pro}}'(x)| \, dx \right| \leq \left| \int_y^z 1 \, dx \right| = |y - z|,$$

which concludes the proof for the probit case.

We now develop the bound $L_{\log}$ for the logit case. To this end, we bound the derivative of $g_{\log}(x) = \frac{1}{1 + e^x}$ as follows:

$$0 \geq g_{\log}'(x) = -\frac{e^x}{(1 + e^x)^2} = -\frac{1}{e^x + e^{-x} + 2} \geq -\frac{1}{4},$$

where we used the inequality of arithmetic and geometric means. Consequently, we have the following bound on the scalar Lipschitz constant (A.2):

$$|g_{\log}(y) - g_{\log}(z)| \leq \left| \int_y^z |g_{\log}'(x)| \, dx \right| \leq \left| \int_y^z \frac{1}{4} \, dx \right| = \frac{1}{4}|y - z|,$$

which concludes the proof for the logit case.

The following lemma establishes a bound on the (vector) Lipschitz constants for the individual regularized regression problems ($RR_1^+$) and ($RR_2$) for both the probit
and the logit case, using the results in Lemma 3. We work out in detail the analysis of \((RR^+)\) for \(\bar{w}_i\), i.e., the transpose of the \(i\)th row of \(W\). The proofs for the remaining subproblems for other rows of \(W\) and all columns of \(C\) follow analogously.

**Lemma 4 (Lipschitz constants)** For a given \(i\) and \(j\), let

\[
\begin{align*}
f_w(\bar{w}_i) &= -\sum_j \log p(Y_{i,j}|\bar{w}_i, c_j) + \frac{\mu}{2} \|\bar{w}_i\|_2^2 \\
&= -\sum_j \Phi((2Y_{i,j} - 1)\bar{w}_i^T c_j) + \frac{\mu}{2} \|\bar{w}_i\|_2^2, \\
f_c(c_j) &= -\sum_i \log p(Y_{i,j}|\bar{w}_i, c_j) = -\sum_i \Phi((2Y_{i,j} - 1)\bar{w}_i^T c_j),
\end{align*}
\]

where \(Y_{i,j}\), \(\bar{w}_i\), and \(c_j\) are defined as in Section 2.1. Here, \(\Phi(x)\) designates the inverse link function, which can either be (2.3) or (2.4). Then, for any \(x, y \in \mathbb{R}^K\), we have

\[
\begin{align*}
\|\nabla f_w(x) - \nabla f_w(y)\|_2 &\leq (L \sigma_{\max}(C) + \mu) \|x - y\|_2, \\
\|\nabla f_c(x) - \nabla f_c(y)\|_2 &\leq L \sigma_{\max}(W) \|x - y\|_2,
\end{align*}
\]

where \(L = L_{\text{pro}} = 1\) and \(L = L_{\text{log}} = 1/4\) are the scalar Lipschitz constants for the probit and logit cases from Lemma 3, respectively.

**Proof** For the sake of brevity, we only show the proof for \(f_w(x)\) in the probit case. The logit cases and the cases for \(f_c(x)\) follow analogously. In what follows, the PDF \(\mathcal{N}(x)\) and CDF \(\Phi(x)\) of the standard normal density (the inverse probit link function) defined in (2.3) are assumed to operate element-wise on the vector \(x \in \mathbb{R}^K\).

In order to simplify the derivation of the proof, we define the following effective matrix associated to \(\bar{w}_i\) as

\[
C_{\text{eff},i} = [(2Y_{i,1} - 1)c_1, \ldots, (2Y_{i,N} - 1)c_N],
\]

which is equivalent to a right-multiplication \(C = [c_1, \ldots, c_N]\) with a diagonal matrix containing the binary-valued response variables \((2Y_{i,j} - 1) \in \{-1, +1\} \forall j\). We can
now establish an upper bound of the $\ell_2$-norm of the difference between the gradients at two arbitrary points $x$ and $y$ as follows:

$$\|\nabla f_w(x) - \nabla f_w(y)\|_2 = \left\| C_{\text{eff},i} \frac{N(C_{\text{eff},i}^T x)}{\Phi(C_{\text{eff},i}^T x)} - C_{\text{eff},i} \frac{N(C_{\text{eff},i}^T y)}{\Phi(C_{\text{eff},i}^T y)} + \mu x - \mu y \right\|_2$$

$$\leq \sigma_{\text{max}}(C_{\text{eff},i}) \left\| \frac{N(C_{\text{eff},i}^T x)}{\Phi(C_{\text{eff},i}^T x)} - \frac{N(C_{\text{eff},i}^T y)}{\Phi(C_{\text{eff},i}^T y)} \right\|_2 + \mu \|x - y\|_2 \quad (A.7)$$

$$\leq L\sigma_{\text{max}}(C_{\text{eff},i}) \|C_{\text{eff},i}^T x - C_{\text{eff},i}^T y\|_2 + \mu \|x - y\|_2 \quad (A.8)$$

$$\leq L\sigma_{\text{max}}^2(C_{\text{eff},i}) \|x - y\|_2 + \mu \|x - y\|_2 \quad (A.9)$$

$$= (L\sigma_{\text{max}}^2(C) + \mu) \|x - y\|_2. \quad (A.10)$$

Here, (A.7) uses the triangle inequality and the Rayleigh-Ritz theorem of [82], where $\sigma_{\text{max}}(C_{\text{eff},i})$ denotes the principal singular value of $C_{\text{eff},i}$. The bound (A.8) follows from Lemma 3, and (A.9) is, once more, a consequence of the Rayleigh-Ritz theorem. The final equality (A.10) follows from the fact that flipping the signs of the columns of a matrix (as we did to arrive at $C_{\text{eff},i}$) does not affect its singular values, which concludes the proof. Note that the proof for $f_c(\cdot)$ follows by omitting $\mu$ and substitute $C$ by $W$ in (A.10). \hfill \blacksquare

Note that in all of the above proofs we only considered the case where the observation matrix $Y$ is fully populated. Our proofs easily adapt to the case of missing entries in $Y$, by replacing the matrix $C$ to $C_I$, where $C_I$ corresponds to the matrix containing the columns of $C$ corresponding to the observed entries indexed by the set $I = \{ j : (i, j) \in \Omega_{\text{obs}} \}$. We omit the details for the sake of brevity.
Appendix B

Proof of Theorem 2

Minimizing $F(x)$ as defined in Theorem 2 using SPARFA-M corresponds to a multi-block coordinate descent problem, where the subproblems $(RR)_1^+$ and $(RR)_2$ correspond to [1, Problem 1.2b and 1.2a], respectively. Hence, we can use the results of [1, Lemma 2.6, Corrollary 2.7, and Theorem 2.8] to establish global convergence of SPARFA-M. To this end, we must verify that the problem $(P)$ satisfies all of the assumptions in [1, Assumption 1, Assumption 2, and Lemma 2.6].

B.1 Prerequisites

We first show that the smooth part of the cost function in $(P)$, i.e., the negative log-likelihood plus both $\ell_2$-norm regularization terms, is Lipschitz continuous on any bounded set in $\mathbb{R}^{(N+Q)K}$. Then, we show that the probit log-likelihood function is real analytic. Note that the logit log-likelihood function is real analytic as shown in [1, Section 2.3]. Finally, we combine both results to prove Theorem 2, which establishes the global convergence of SPARFA-M.

Lemma 5 (Lipschitz continuity) Define $x = [\bar{w}_1^T, \ldots, \bar{w}_Q^T, c_1^T, \ldots, c_N^T]^T$, and let

$$f(x) = - \sum_{(i,j) \in \Omega_{obs}} \log p(Y_{i,j}|\bar{w}_i, c_j) + \frac{\mu}{2} \sum_i \|\bar{w}_i\|_2^2 + \frac{\gamma}{2} \sum_j \|c_j\|_2^2.$$ 

Then, $f(x)$ is Lipschitz continuous on any bounded set $D = \{x : \|x\|_2 \leq D\}$.
Proof. Let \( y, z \in D \), recall the notation of Lemma 4, and let \( \bar{w}^y_i, \bar{w}^z_i, c^y_j, \) and \( c^z_j \) denote the blocks of variables \( \bar{w}_i \) and \( c_j \) in \( y \) and \( z \), respectively. We now have

\[
\| \nabla f(y) - \nabla f(z) \|_2 = \left( \sum_{i,j} ( (\nabla f_w(\bar{w}^y_i) - \nabla f_w(\bar{w}^z_i))^2 \\
+ (\nabla f_c(c^y_j) - \nabla f_c(c^z_j))^2 + \gamma^2 \|c^y_j - c^z_j\|_2^2) \right)^{1/2} \leq \left( \sum_{i,j} \left( (L \sigma_{\max}(C) + \mu)^2 \|\bar{w}^y_i - \bar{w}^z_i\|_2^2 \\
+ (L^2 \sigma_{\max}(W) + \gamma^2) \|c^y_j - c^z_j\|_2^2) \right)^{1/2} \right) \leq \left( L (\|W\|_F^2 + \|C\|_F^2) + \max\{\mu, \gamma\} \right) \|y - z\|_2 \]

where (B.3) follows from Lemma 4, and (B.4) follows from the fact that the maximum singular value of a matrix is no greater than its Frobenius norm ([82]), which is bounded by \( D \) for \( x, y \in D \). We furthermore have \( L = 1 \) for the probit case and \( L = 1/4 \) for the logit case, shown in Lemma 3. Thus, \( f(x) \) is Lipschitz continuous on any bounded set.

Lemma 6 (Real analyticity) Define \( x = [\bar{w}_1^T, \ldots, \bar{w}_Q^T, c_1^T, \ldots, c_N^T]^T \), and let

\[
g(x) = - \sum_{(i,j) \in \Omega_{obs}} \log p(Y_{i,j}|\bar{w}_i, c_j) = - \sum_{(i,j) \in \Omega_{obs}} \log \Phi((2Y_{i,j} - 1)\bar{w}_i^T c_j),
\]

where \( \Phi_{pro}(\cdot) \) is the inverse probit link function defined in (2.3). Then, \( g(x) \) is real analytic.

Proof. We first show that the scalar negative log-likelihood function for the inverse probit link function is real analytic. To this end, recall the important property established by [83] that compositions of real analytic functions are real analytic. Therefore, the standard normal density \( \mathcal{N}(x) \) is real analytic, since the exponential function and
\( x^2 \) are both real analytical functions. Consequently, let \( \mathcal{N}^{(k)}(x) \) denote the \( k \)th derivative of \( \mathcal{N}(x) \), then \( \left( \frac{\mathcal{N}^{(k)}(x)}{k!} \right)^{\frac{1}{k}} \) is bounded for all \( k \), according to the definition of real analytic functions.

Now we show that \( \Phi(x) = \int_{-\infty}^{x} \mathcal{N}(t)dt \) is also real analytic. Its \( k \)th derivative is given by \( \Phi^{(k)}(x) = \mathcal{N}^{(k-1)}(x) \), and therefore \( \left( \frac{\Phi^{(k)}(x)}{k!} \right)^{\frac{1}{k}} \) is obviously bounded for all \( k \), since \( \left( \frac{\mathcal{N}^{(k)}(x)}{k!} \right)^{\frac{1}{k}} \) is bounded for all \( k \) as we have just shown. Thus, \( \Phi(x) \) is real analytic.

Given that \( \Phi(x) \) is real-analytic, it follows that the negative log-probit-likelihood \( -\log \Phi(x) \) is real analytic, since both the logarithm function and the inverse probit link function are real analytic. Finally, extending the proof from scalar functions to vector functions preserves analyticity according to [1, Section 2.2].

**B.2 Proof of Theorem 2**

We are finally armed to prove Theorem 2. We begin by showing that our problem (P) meets [1, Assumptions 1 and 2]. Then, we show that (P) meets all the additional assumptions needed for the convergence results in [1, Lemma 2.6], through which we can establish convergence of the sequence \( \{x^t\} \) from certain starting points to some finite limit point. Finally, we use [1, Theorem 2.8] to show global convergence of SPARFA-M from any starting point.

**B.2.1 Assumption 1**

We start by showing that (P) meets [1, Assumption 1]. Since every term in our objective function in (P) is non-negative, we have \( F(x) > -\infty \). It is easy to verify that (P) is also block multi-convex in the variable \( x \), with the rows of \( W \) and columns of \( C \) forming the blocks. Consequently, the problem (P) has at least one critical point, since \( F(x) \) is lower bounded by 0. Therefore, Assumption 1 is met.
B.2.2 Assumption 2

Problem (P) also meets [1, Assumption 2] regarding the strong convexity of the individual subproblems. Due to the presence of the quadratic terms $\frac{\mu}{2} \|\bar{w}_i\|_2^2$ and $\frac{7}{2} \|c_j\|_2^2$, the smooth part of the objective functions of the individual subproblems (RR$_1^+$) and (RR$_2$) are strongly convex with parameters $\mu$ and $\gamma$, respectively. Consequently, Assumption 2 is satisfied.

B.2.3 Assumptions in [1, Lem. 2.6]

Problem (P) also meets the assumptions in [1, Lem. 2.6] regarding the Lipschitz continuity of the subproblems and the Kurdyka-Łojasiewicz inequality. Lemma 5 shows that $f(x) = -\sum_{(i,j) \in \Omega_{obs}} \log p(Y_{i,j} | \bar{w}_i, c_j) + \frac{\mu}{2} \sum_i \|\bar{w}_i\|_2^2 + \frac{\gamma}{2} \sum_j \|c_j\|_2^2$, satisfies the Lipschitz continuous requirement in [1, Lemma 2.6]. As shown in Lemma 6 for the probit case and as shown in [1, Section 2.2] for the logit case, the negative log-likelihood term in (P) is real analytic, therefore also sub-analytic. All the regularizer functions in $F(x)$ defined in Theorem 2 are semi-algebraic and therefore sub-analytic, a consequence of [84, Section 2.1] and [1, Section 2.2]. Using [85, Theorems 1.1 and 1.2], the objective function $F(x)$ is also sub-analytic, since all of its parts are sub-analytic and bounded below (non-negative), therefore satisfying the Kurdyka-Łojasiewicz inequality at any point $x$, as shown in [84, Theorem 3.1]. Finally, the SPARFA-M algorithm uses $\omega_i^{k-1} \equiv 0$ and $\ell = \min\{\mu, \gamma\}$ where $\omega_i^{k-1} \epsilon$ and $\ell$ as defined in [1, Lemma 2.6].

Up to this point, we have shown that (P) satisfies all assumptions and requirements in [1, Lemma 2.6]. Now, SPARFA-M follows [1, Lemma 2.6] in the sense that, if $x^0$ is sufficiently close to some critical point $\hat{x}$ of (P), (more specifically, $x^0 \in B$ for some $B \subset U$ where $U$ is a neighborhood of $\hat{x}$ in which Kurdyka-Łojasiewicz inequality holds), then $\{x^k\}$ converges to a point in $B$. This establishes the convergence of SPARFA-M to a local minimum point from certain starting points.
B.2.4 Global Convergence

Finally, we can use [1, Lemma 2.6] to establish global convergence of SPARFA-M. It is obvious that the objective function \((P)\) is bounded on any bounded set. Hence, the sequence \(\{x^k\}\) will always have a finite limit point and meet the assumptions in [1, Theorem 2.8]. The final statement of Theorem 2 now directly follows from [1, Theorem 2.8]. Moreover, if the starting point is in close proximity to a global minimum, then SPARFA-M is guaranteed to converge to a global minimum. This is a consequence of [1, Corollary 2.7].