Self-Expressive Clustering of Binary Data via Group Sparsity

Andrew S. Lan  
Rice University  
e-mail: sl29@rice.edu

Christoph Studer  
Cornell University  
e-mail: studer@cornell.edu

Richard G. Baraniuk  
Rice University  
e-mail: richb@rice.edu

Abstract—We present a novel, computationally efficient approach to cluster binary-valued data using self-expressive representations. Given a binary-valued data matrix, we use sparse, self-expressive representations to cluster similar rows and similar columns. We formulate our method as a convex logistic matrix-factorization problem that relies on group sparsity to identify the key rows and columns that explain the observed data. We demonstrate the effectiveness of our approach on two educational datasets where we cluster similar learners and test questions.

I. INTRODUCTION

Self-expressive data representations seek to explain observed data points as a linear combination of other observed data points [1]–[4]. This approach resembles dictionary learning (DL) [5], [6], where the dictionary simply corresponds to the observed data points. The advantages of such an approach are that (i) the resulting learning model only maps data points as a linear combination of other observed data points [1]–[4].

This approach resembles dictionary learning (DL) [5], [6], where the dictionary simply corresponds to the observed data points. The advantages of such an approach are that (i) the resulting learning model only maps data points as a linear combination of other observed data points [1]–[4].

We consider the case of a (partially) observed data matrix $Y \in \{\pm 1\}^{P \times N}$ consisting of response data of $N$ users to $P$ items. We assume that only the entries $Y_{i,j}$ in the index set $\Omega$ are observed. In contrary to the conventional, real-valued and one-sided self-representation models in [1], [2], we propose the following two-sided binary-valued observation model:

$$Y = \text{sign}(A \overline{Y} + \overline{Y}B^T + N).$$

Here, the entries in $N$ are i.i.d. standard logistic distributed. We construct the matrix $\overline{Y}$ from the observed entries of $Y$ as follows: $\overline{Y}_{i,j} = Y_{i,j}$ if $(i,j) \in \Omega$ and $\overline{Y}_{i,j} = 0$ otherwise. The $P \times P$ left matrix $A$ and $N \times N$ right matrix $B$ characterize how the data in $Y$ is self-represented. Specifically, the entries $A_{i,j}$ with $i,j \in \{1, 2, \ldots, P\}$ indicate how row $i$ is self represented by row $j$ in $\overline{Y}$, and the entries $B_{m,n}$ with $m,n \in \{1, 2, \ldots, N\}$ indicate how column $m$ is self represented by column $n$ in $\overline{Y}$.

III. OPTIMIZATION PROBLEM AND ALGORITHM

To identify a small number of rows and columns that explain the entire binary-valued data matrix sufficiently well, we minimize the negative log-likelihood $- \log p(Y|A, B)$ of the data model (1) together with (group) sparsity-inducing penalties on $A$ and $B$ [8], [9]. In particular, we solve the following optimization problem:

$$\begin{align*}
\text{minimize} & \quad - \log p(Y|A, B) \\
& \quad + \lambda \sum_{i=1}^{P} ||a_i||_2 + \sum_{j=1}^{N} ||b_j||_2 \\
& \quad + \gamma (||A||_1 + ||B||_1), \\
\text{subject to} & \quad A \geq 0, B \geq 0 \\
& \quad \text{diag}(A) = 0, \text{diag}(B) = 0.
\end{align*}$$

Here, $a_i$ and $b_j$ denote the $i^{th}$ column of $A$ and $j^{th}$ column of $B$, respectively. The group-sparsity penalties $\sum_{i=1}^{P} ||a_i||_2$ and $\sum_{j=1}^{N} ||b_j||_2$ enforce that only a small number of "eigen-columns" of $A$ and $B$ are selected. The $\ell_1$-norm penalties on $A$ and $B$ are the sums of the absolute values of their entries. The constraints $A \succeq 0$ and $B \succeq 0$ induce entry-wise non-negativity and encourage representations of similar columns, and that $\text{diag}(A) = 0$ and $\text{diag}(B) = 0$ inhibit pure self-expressive representations. The parameters $\lambda$ and $\gamma$ determine the amount of group sparsity and entry-wise sparsity, respectively, and are selected via cross-validation [10].

The problem (P) is convex, and we use a fast iterative shrinkage-thresholding algorithm (FISTA)-based algorithm [11], [12] to estimate $A$ and $B$. In each algorithm iteration, we perform a gradient step with respect to the smooth negative log-likelihood term and a projection step with respect to the non-smooth, sparsity-inducing penalties. The projection requires us to solve problems of the form

$$\text{minimize}_{X \succeq 0, X \geq 0} \frac{1}{2} \|X - Z\|_F^2 + \lambda \sum_{i=1}^{N} \|x_i\|_2 + \gamma \|X\|_1,$$

for which we use the alternating direction method of multipliers [13].

IV. NUMERICAL EXPERIMENTS

In order to demonstrate the efficacy of SEC, we evaluate its predictive performance on unobserved entries in two educational datasets. In each experiment, we learn $A$ and $B$ from 75% randomly selected entries in $Y$, test on the other 25%, and report the prediction accuracy. Each experiment is averaged over 10 Monte Carlo trials. Table I shows the predictive performance (in percent) on both datasets. The results show that using both rows and columns to self-represent the data results in superior performance as opposed to using only A or B. The proposed approach performs almost as well as the state-of-the-art quantized matrix completion (Q-MC) method in [7], which does not enable interpretability.

Figures 1 and 2 visualize the explanatory power of SEC using Dataset 1 consisting of $N = 43$ learners answering $P = 143$ questions in an undergraduate signal processing course. For simplicity, we visualize the results of our approach with only $A$ or only $B$. Circles represent the sparse set of selected "eigen-questions/learners," while boxes represent the other questions/learners. Thicker edges represent stronger similarity. We observe that a small number of eigen-learners or eigen-questions are capable of explaining the entire dataset. In Figure 1, for example, we see that most questions cluster around question 17 and 119; a careful inspection of these questions shows that they cover a large proportion of all concepts in the course.
TABLE I

PREDICTIVE PERFORMANCE OF SELF-EXPRESSIVE CLUSTERING

<table>
<thead>
<tr>
<th>Dataset</th>
<th>A only</th>
<th>B only</th>
<th>A and B</th>
<th>Q-MC [7]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>82.4 ± 0.4%</td>
<td>81.8 ± 0.2%</td>
<td>86.3 ± 0.5%</td>
<td>87.0 ± 0.2%</td>
</tr>
<tr>
<td>2</td>
<td>78.8 ± 0.3%</td>
<td>78.4 ± 0.3%</td>
<td>80.1 ± 0.4%</td>
<td>80.2 ± 0.3%</td>
</tr>
</tbody>
</table>

REFERENCES


