Information Sciences 494 (2019) 278–293

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Multi-view cluster analysis with incomplete data to understand treatment effects

Guoqing Chao[a], Jiangwen Sun[b], Jin Lu[a], An-Li Wang[c], Daniel D. Langleben[c], Chiang-Shan Li[d], Jinbo Bi[a,*]

[a] Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA
[b] Department of Computer Science Old Dominion University, Norfolk, Virginia, USA
[c] University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
[d] Department of Psychiatry Yale University, New Haven, CT, USA

ARTICLE INFO

Article history:
Received 25 December 2018
Revised 13 March 2019
Accepted 21 April 2019
Available online 22 April 2019

Keywords:
Missing value
Multi-view data analysis
Co-clustering
Granular computing
Opioid addiction

ABSTRACT

Multi-view cluster analysis, as a popular granular computing method, aims to partition sample subjects into consistent clusters across different views in which the subjects are characterized. Frequently, data entries can be missing from some of the views. The latest multi-view co-clustering methods cannot effectively deal with incomplete data, especially when there are mixed patterns of missing values. We propose an enhanced formulation for a family of multi-view co-clustering methods to cope with the missing data problem by introducing an indicator matrix whose elements indicate which data entries are observed and assessing cluster validity only on observed entries. In comparison with common methods that impute missing data in order to use regular multi-view analytics, our approach is less sensitive to imputation uncertainty. In comparison with other state-of-the-art multi-view incomplete clustering methods, our approach is sensible in the cases of either missing any entry in a view or missing the entire view. We first validated the proposed strategy in simulations, and then applied it to a treatment study of *opioid* dependence which would have been impossible with previous methods due to a number of missing-data patterns. Patients in the treatment study were naturally assessed in different feature spaces such as in the pre-, during- and post-treatment time windows. Our algorithm was able to identify subgroups where patients in each group showed similarities in all of the three time windows, thus leading to the identification of pre-treatment (baseline) features predictive of post-treatment outcomes. We found that cue-induced heroin craving predicts adherence to XR-NTX therapy. This finding is consistent with the clinical literature, serving to validate our approach.

1. Introduction

Granular computing, as defined by Bargiela and Pedrycz in [3], is a computational principle for effectively using granules in data such as subsets or groups of samples, or intervals of parameters to build an efficient computational model for

[*] Corresponding author.
E-mail addresses: jsun@odu.edu (J. Sun), langlebe@pennmedicine.upenn.edu (D.D. Langleben), chiang-shan.li@yale.edu (C.-S. Li), jinbo.bi@uconn.edu (J. Bi).

https://doi.org/10.1016/j.ins.2019.04.039

complex systems with massive quantities of data, information and knowledge. It provides an umbrella to cover any theories, methodologies, techniques, and tools that make use of granules - components or subspaces of a space - in problem solving [42]. It can consist of a structured combination of algorithmic abstraction of data and non-algorithmic, empirical verification of the semantics of these abstraction [3,43]. Cluster analysis is such an important technique aiming to identify subgroups in a population so that subjects in the same group are more similar to each other than to those in other groups. It has been extensively used in computer vision [29,45], natural language processing [6,7,24] and bioinformatics [21,26]. In this paper, we propose a method to identify the cluster granules in a patient population to analyze treatment study data where missing values occur. In particular, we take into account the nature of the treatment studies, i.e., multiple views of input variables with incomplete data to model treatment effects.

Multi-view data exist in many real-world applications. For instance, a web page can be represented by words on the page or by all the hyper-links pointing to it from other pages. Similarly, an image can be represented by the visual features extracted from it or by the text describing it. Multi-view data analytics aims to make the full use of the multiple views of data, and has attracted wide interests in recent years such as in those works of semi-supervised learning with unlabeled data [2,4,11], or unsupervised multi-view data analytics [8–10,20,35]. In this paper, we focus on the unsupervised multi-view clustering methods [14,15,18,28,41,44], specifically multi-view co-clustering [33–35]. Consider a dataset in which data matrices have rows representing subjects and columns representing features. They share the same set of subjects but each matrix has a different set of features. Multi-view co-clustering is a technique to cluster the rows (subjects) consistently across multiple data matrices (sets of features). A family of such methods [33–35] can find subspaces in each different view (rather than using all features in each view) to group subjects consistently across the views. However, existing multi-view co-clustering methods cannot deal with incomplete datasets. Subjects with missing values often need to be removed or imputation has to be done before clustering. Eliminating data weakens the results by reducing the sample size. On the other hand, imputation may bring a separate layer of uncertainty, especially when some data are missing at random but others are not.

The issue of missing value is common in real-world applications. Data may be missing at random or due to selection bias. For example, in the study of an asthma education intervention [27], some missing values were caused by the participants who forgot to visit the school clinic to fill out the form; some were caused by the students whose asthma was too serious to visit the school clinic to report. The former values are missing at random and the latter are not. According to different reasons, the strategies to handle missing values are different. If the data are missing at random, researchers either use only the samples with complete variables [40] or impute the missing values [12] from the available data; if data are missing systematically, there can be a variety of difficulties for researchers to recognize and capture the missing patterns.

In longitudinal studies [16], the missing patterns are very complicated and difficult to deal with. A prospective treatment study usually begins with a baseline assessment and follows up through time, and missing values are commonly encountered because study subjects may not be available at all time points. Just as in our heroin dependence treatment study, both random and non random missing values exist. Because of the mixed missing value patterns, we choose a simple yet effective strategy to handle this problem: introducing an indicator matrix to indicate which feature is observed for which subjects and then omitting the calculation of the loss ocurring on missing locations while clustering. Since the missing values is unknown, imputation cannot guarantee the right values. Ignoring the loss in the missing locations should be a better choice.

In multi-view data, if there are many missing values in different views, then it is useful but challenging to make the different views compensate each other on the missing information to obtain consistent subject grouping. The most recent multi-view co-clustering methods cannot handle incomplete data that potentially occur in all of the views. Moreover, although imputation methods have been studied for decades, our simulation studies show that even the latest imputation method might not effectively handle the nature of mixed missing patterns, and create another layer of uncertainty in the imputed data. A few recent methods handle incomplete data [22,30,31,37], but they commonly assume that there is at least one complete view for all the sample subjects or each subject should have one or more complete views, which is however not the case in treatment studies (we can have incomplete features in every view).

For each view of the data, all the methods mentioned so far require either having the complete features in a view or having no features in the view. Two kernel based methods [31,37] borrowed the idea of graph Laplacian to complete the incomplete kernel matrix. The partial multi-view clustering (PVC) method [22] reorganized the data into three parts (in the case of two views): subjects with both complete views, subjects with complete view 1, and subjects with complete view 2, and then projected them into a latent space and finally conducted a standard clustering algorithm in the latent space. When multiple incomplete views are present, clustering via weighted nonnegative matrix factorization with $L21$ regularization (the so-called WNMF21) is the most similar to our method which also introduces an indicator matrix. That method used only one weighted matrix to indicate which instance misses which view while we introduce an indicator matrix for each view to indicate the observed entries in the corresponding view. Among all the multi-view clustering methods with incomplete data, only ours is not restricted to any specific missing data pattern. In comparison with the common strategy of removing subjects with missing values, our approach can use all observed data in a cluster analysis. In comparison with common methods that impute missing values and then use regular multi-view analytics, our approach is less sensitive to the imputation uncertainty. In comparison with other state of the art multi-view incomplete clustering methods, our approach is applicable to any pattern of missing data. We first validate the proposed algorithm in a simulation study, and then use it in a longitudinal treatment study to better understand the differential responses of heroin users to the medication naltrexone.

The main contributions of our work include the following two aspects:

1. In terms of methodology, we propose an enhanced multi-view co-clustering algorithm that is capable of dealing with complex patterns of incomplete data, and validate its performance by comparing against other state of the art methods.
2. In terms of application, we have successfully applied the proposed method to an opioid dependence treatment study and identified meaningful patient subgroups, which would be implausible otherwise. By analyzing the study data, we produce an important finding that features such as changes in craving for heroin in response to cues at baseline could be a useful predictor for patient adherence to naltrexone.

The rest of this paper is organized as follows: we describe the longitudinal multi-view data collected in our treatment study in Section 2; an enhanced multi-view co-clustering method is introduced in Section 3 to deal with missing values; Section 4 presents the performance comparison on the synthetic datasets and the statistical analysis results in the case study; we then conclude and discuss in Section 5.

2. Incomplete data in treatment study

Opioid addiction is a resurgent public health problem in the United States [36]. There exist three Food and Drug Administration (FDA) approved medications for the treatment of opioid use disorder in general and heroin addiction in particular. Two of these options are opioid agonists, acting on the principle of opioid substitution and one - naltrexone, is an opioid antagonist. Naltrexone is an important treatment option because it is pharmacologically analogous to abstinence. However, the clinical efficacy of oral naltrexone is limited by non-adherence [23]. To address this limitation, an injectable extended-release preparation of naltrexone (XR-NTX) has been developed. In the following section we briefly describe a prospective study of XR-NTX in heroin addicted individuals [38] and the missing values encountered in this study.

2.1. Subjects and assessment

Thirty-two opioid-dependent individuals who used intravenous heroin were recruited. Heroin was a drug of choice in all participants. Most of the patients also used other illicit drugs, such as prescription opioids, cocaine and marijuana. All of them smoked tobacco cigarettes. Although the sample is relatively small, it represents the common sample size in a treatment study that includes repeated magnetic resonance imaging (MRI) tests. Participants received up to three monthly injections of XR-NTX (manufactured by Alkermes, Cambridge, MA, USA). The urine drug screens (UDSs) and the Beck Depression Inventory (BDI) survey were administered weekly. Plasma concentrations of naltrexone and 6-beta-naltrexol (an active metabolite) were measured 13 ± 7 days after the first injection, 22 ± 13 days after the second injection and 21 ± 5 days after the third injection with established liquid chromatography and tandem mass spectrometry techniques described in an early study [19].

To measure the level of craving for heroin and other drugs such as cocaine, MRI sessions were conducted before, during and after the XR-NTX treatment. Two comparable sets of previously reported cue reactivity tasks [19,38] were presented in each MRI session and counter-balanced across participants. For the cue reactivity task, a stimuli set comprised 48 heroin-related and 48 neutral pictures. Stimuli were separated by a variable interval (0 – 18 s) during which a crosshair was displayed. Presentation software (Neurobehavioral Systems, San Francisco, CA, USA) was used to present the stimuli in a random, event-related fashion. Subjects were asked to rate their craving for heroin and other drugs on a scale of 0 (not at all) to 9 (extremely), before and after the cue reactivity task. Post-session craving was managed clinically by debriefing and “talk down” until craving was fully subsided.

To explore the correlations between different types of variables, and evaluate if baseline variables correlate with any variables during or after treatment, we introduce two ways to organize the views.

The data variables were naturally grouped into three views by variable type:

- View 1 - survey variables: The study collected participants’ responses to a set of surveys, resulting in craving scores for heroin (*Cra_Heroin*), heroin withdrawal symptoms (*WD_Heroin*), feeling high for heroin (*high_Heroin*), as well as craving scores (*Cra_Oth*), other drugs withdraw symptoms (*WD_ Oth*), and feeling high for other drugs (*high_Oth*). Besides these, three other survey instruments: BDI, timeline followback (TLFB) measures for smoking and subjective opiate withdrawal scale (SOWS) also provided a set of variables. We used prefix Pre, On, and Post, to represent the three periods of pre-, during-, and post-treatment. We computed the difference between the two craving scores before and after the cue exposure for each of the three sessions. The resultant variables were named in a specific format. For instance, Δ*Pre_Cra_Heroin* referred to the change in self-reported craving ratings for heroin after exposure to drug-related stimuli (i.e. Δ*Pre_Cra_Heroin* = Post exposure craving - Pre exposure craving). We similarly computed the differences from all the raw craving variables.
- View 2 - lab test variables: consisted of naltrexone (Nal) and 6-beta-naltrexol plasmas (Beta) and qualitative urine test results for opioid (OPI), teltrahydrocannabinol (THC), cocaine (COC) in the three different sessions. The variables were named as follows. For instance, *On_OPI_2nd* represents the urine test result for the opioid level after the 2nd XR-NTX injection. Note that the prefix “On” indicates that it is during the treatment sessions and distinguished by the 1st, 2nd, or 3rd injection periods. We also had these variables measured at pre-treatment (also known as baseline) and post-treatment time points.

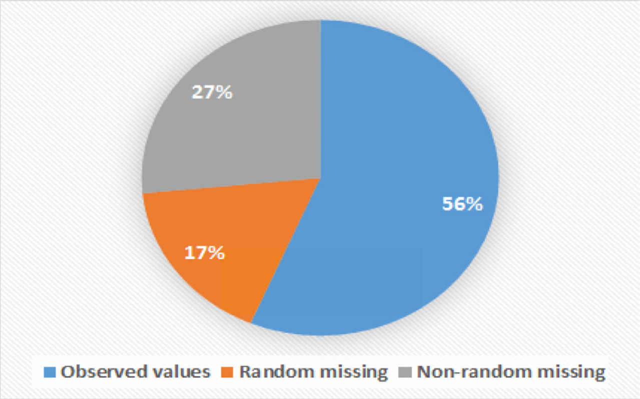

Fig. 1. The distribution of the data entries in the heroin treatment dataset.

- View 3 - vital sign variables: We computed the difference of vital signs, such as heart rate (HR), systolic pressure (SP), mean arterial pressure (MAP) between before and after the cue exposure during the MRI tests for all of the three sessions. For instance, Δ*Pre_HR* specified the heart rate difference after exposure to visual drug-related cues at baseline.

The data variables could also be categorized into three views according to the time windows:

- Window 1 - baseline variables: all the variables obtained at baseline formed this view of data for a patient. For instance, *Pre_BDI* measured the depression level at baseline. Other variables included *Cra_Heroin, WD_Heroin, high_Heroin, Cra_Oth, WD_Oth, high_Oth*, HR, SP, MAP, SOWS, OPI, THC, COC, BDI, and TLFB. We included "Pre" in the names of these variables to denote that they were measured at baseline.
- Window 2 - variables collected during treatment: were used in this view. For instance, Δ*On_Cra_Heroin* represented craving elevation for heroin after drug-related cues at a time point within the treatment period. Besides the variables described in Window 1, this view had additional variables: Nal and Beta which were naltrexone plasma levels measured at a time point after each injection. Therefore, some variables in this view may include 1st, 2nd, or 3rd in their names. Note that Nal and Beta were collected only during and after treatment, not at baseline.
- Window 3 - post-treatment variables: were collected at a follow-up time point after the treatment. For instance, Δ*Post_WD_Oth* referred to the difference in withdrawal levels for other drugs (Post cue exposure - Pre exposure) after the treatment process was terminated. These variables included "Post" in their names.

Note that if we use time windows to define the views, each view could have some variables that are not related to cue exposure, thus these variables are not the difference Δ variables. Readers can refer to a supplemental table available at https://healthinfo.lab.uconn.edu/mvbc-incomplete/ for a complete list of variables included in this study. One of the fundamental questions in treatment studies is that how the variables are related, e.g., whether a lab test result is associated with a opioid withdrawal score in a subgroup of patients, and whether any baseline variable could be predictive of a post-treatment outcome in a subgroup of patients. These questions motivated us to group variables into views according to two settings by variable type or by time.

2.2. Missing data

Missing values are commonly encountered in treatment studies. In this heroin treatment study, there existed mixed types of missing values where some were clearly associated with an event, e.g. drop-out from the study, but for others, the cause was unknown. Fig. 1 demonstrates a pie chart of the data distribution, showing 44% of total missing values.

There were two types of non-random missing values in this study. When a patient dropped out of the treatment study, all of the variables collected afterwards would be missing for this patient. For instance, nine subjects decided that they would not tolerate naltrexone and hence received no injection, so there were only baseline variables for these subjects. After the 1*st* injection period, five other subjects dropped out, so they missed values for the remaining treatment variables. After the 2*nd* injection, three more subjects left. Totally, fifteen subjects received all three injections. The number of injections each patient received served as a good overall treatment outcome measure, especially because XRNTX provides a pharmacological abstinence state regardless of whether subjects attempt to use opioids. In such cases, UDS would be positive for opioids but patient remains abstinent. Furthermore, the variables for some subjects might have missing values due to the "obligated missing" situation. For instance, some patients might not pass the naloxone challenge test before receiving an injection, so the missing values related to that injection follow the "obligated missing" pattern. Table 1 provides a more comprehensive view of the non-random missing values spanning across different views and windows.

In this heroin treatment study, you cannot ignore the subjects with missing values since almost all of the subjects have missing values. It may not be a good choice to impute data before the multi-view cluster analysis, because there exist both

Table 1
The distribution of non-random missing values over all the variable groups on the 32 subjects. The three views of variables are indicated by “V1”, “V2” and “V3”, respectively, and “ – ” indicates that the values are observed whereas “ + ” indicates that missing values present. “NoS” indicates the number of subjects.

			Baseline			During-treatment					Post-treatment		
	NoS	NoS	V1	V2	V3	V2 after 1st	V2 after 2nd	V2 after 3rd	V1	V3	V1	V2	V3
Received 0 injections	9	3	+	–	+	+	+	+	+	+	+	+	+
		3	–	–	+	+	+	+	+	+	+	+	+
		3	–	+	–	+	+	+	+	+	+	+	+
Received 1 injections	5	1	–	–	+	–	+	+	–	+	–	–	+
		1	–	–	–	–	+	+	–	–	+	–	+
		3	–	–	–	–	+	+	–	–	+	+	+
Received 2 injections	3	1	–	–	–	–	+	+	–	+	+	+	+
		1	–	+	+	+	+	+	+	+	+	+	+
		1	–	–	–	–	–	+	–	–	+	+	+
Received 3 injections	15	3	–	–	–	–	–	–	–	–	+	+	+
		2	–	–	+	–	–	–	–	+	–	–	+
		10	–	–	–	–	–	–	–	–	–	–	–

random and non-random missing patterns. In such a situation, it will be a better option to ignore just the missing values rather than remove the entire record of a subject with missing values that will result in a reduced sample size. In the following section, we will develop a multi-view co-clustering algorithm that can ignore missing values.

3. Multi-view co-clustering with incomplete data

Multi-view co-clustering aims to group subjects in the same way across multiple views and identify the important variables from each view. In other words, multi-view co-clustering can group the subjects into some subgroups and at the same time the selected variables from different views play an important role in the grouping process. Since the selected variables from different views identify the same subject groups, the characteristics of each group helps show the correlation of the variables between different views. Although such a method is desirable, the first step is to extend it to incomplete data. Next, we introduce the proposed multi-view co-clustering method for incomplete data.

3.1. The optimization formulation

Given the data matrices $\mathbf{X}^k \in \mathcal{R}^{n\times d_k} (k = 1, \cdots, m)$ which can describe the same set of n subjects from m different views. For each matrix $\mathbf{X}^k$, two vectors $\mathbf{u}^k$ and $\mathbf{v}^k$ can be obtained by a rank-one matrix approximation, i.e., $\mathbf{X}^k \approx \mathbf{u}^k\mathbf{v}^{kT}$. If we require both $\mathbf{u}$ and $\mathbf{v}$ to be sparse, then rows in $\mathbf{X}^k$ corresponding to non-zero components in $\mathbf{u}^k\mathbf{u}^k$ form a subject cluster and columns in $\mathbf{X}^k$ corresponding to non-zero components in $\mathbf{v}^k$ are the selected variable from the kth view. However, in a multi-view setting, $\mathbf{u}^k$'s from the different views do not guarantee to form the same subject cluster. To create subject clusters consistent across different views, in a recent work [35], a binary vector $\boldsymbol{\omega}$ is introduced to make the clusters across different views consistent where each component of $\mathbf{u}^k$ is multiplied by the corresponding component of $\boldsymbol{\omega}$, i.e., $u_i^k = u_i^k\omega_i$. If $\omega_i = 0$, regardless the value of the ith component of every $\mathbf{u}^k$, the ith row will be excluded from the cluster in all views. Hence, to get the subject cluster, rather than finding sparse $\mathbf{u}^k$'s, the method seeks a sparse $\boldsymbol{\omega}$. After identifying the subject cluster, we will deflat the data matrix by $\mathbf{X}^k = \mathbf{X}^k - \text{diag}(\boldsymbol{\omega})\mathbf{u}^k\mathbf{v}^{\mathbf{k}^T}$ and then seek for the next subject cluster and iterate this process untill no subjects are left. Our method is similar to the power method with deflation [13] that sequentially finds each pair of singular vectors at a time and then deflates the data matrix to compute next eigenvectors.

Thus, the optimization problem is formulated as follows:

$$\begin{array}{ll} \min\limits_{\boldsymbol{\omega},\mathbf{u}^k,\mathbf{v}^k,k=1,\cdots,m} & \sum\limits_{k=1}^{m} \frac{1}{2}\|\mathbf{X}^k - \text{diag}(\boldsymbol{\omega})\mathbf{u}^k\mathbf{v}^{k^T}\|_F^2 \\ \text{subject to} & \|\boldsymbol{\omega}\|_0 \leq s_\omega, \ \ \|\mathbf{v}^k\|_0 \leq s_{v^k}, \\ & k = 1, \cdots, m, \\ & \boldsymbol{\omega} \in \mathcal{B}_n. \end{array} \tag{1}$$

where the objective function measures the approximation error in terms of the Frobenius norm of the difference matrices in every view, and $\|\cdot\|_0$ is the so-called ℓ_0 vector norm (which is not really a vector norm) that returns the number of non-zeros in a vector. The set $\mathcal{B}_n$ contains all binary vectors of length n, and s_ω and s_{v^k}'s are hyper-parameters that are pre-chosen to determine, respectively, how many subjects in a group and how many variables will be selected from each view.

This multi-view co-clustering algorithm provides a sensible way to analyze the heroin treatment data, but in practice it cannot be applied because of the missing values. To create a general strategy to recover cluster structure from the observed data, we introduce an indicator matrix $\boldsymbol{A}^k$ whose entry $\mathbf{A}_{ij}^k$ indicates whether $\mathbf{X}_{ij}^k$ is observed, i.e.,

$$\mathbf{A}_{ij}^k = \begin{cases} 1 & \mathbf{X}_{ij}^k \text{ is observed} \\ 0 & \mathbf{X}_{ij}^k \text{ is missing.} \end{cases} \tag{2}$$

The indicator matrix prompts the clustering algorithm to ignore the missing values. This way we can use more information than ignore the subjects with any missing value, and can be better if there are non-random missing values and imputation quality is not desirable.

Now, to minimize the error only occurred on the observed entries, the loss function of Eq. (1) becomes

$$\sum_{k=1}^{m} \frac{1}{2} \|\mathbf{A}^k \odot \left(\mathbf{X}^k - \text{diag}(\boldsymbol{\omega})\mathbf{u}^k \mathbf{v}^{k^T}\right)\|_F^2 \tag{3}$$

where $\odot$ computes the Hadamard (element-wise) product of two matrices and returns a matrix of the same size as $\mathbf{X}^k$ (or $\boldsymbol{A}^k$). When $\mathbf{A}_{ij}^k = 0$, the algorithm does not care about the actual value of the term $\boldsymbol{\omega}_i u_i^k v_j^k$ because we do not observe $\mathbf{X}_{ij}^k$ and thus the difference on this item should not be penalized. Note that for different views, the loss in Eq. (3) can be weighed so that it can deal with different views differently, herein we simply treat each view equally because the variable number in each view of our data set is more or less in the same quantity. Eq. (3) is equal to the following formula:

$$h(\boldsymbol{\omega}, \mathbf{u}^k, \mathbf{v}^k) = \sum_{k=1}^{m} \frac{1}{2} \|\mathbf{A}^k \odot \mathbf{X}^k - \mathbf{A}^k \odot \left(\text{diag}(\boldsymbol{\omega})\mathbf{u}^k \mathbf{v}^{k^T}\right)\|_F^2,$$

and then the optimization problem becomes:

$$\begin{aligned} \min_{\boldsymbol{\omega}, \mathbf{u}^k, \mathbf{v}^k, k=1,\cdots,m} \quad & h(\boldsymbol{\omega}, \mathbf{u}^k, \mathbf{v}^k) \\ \text{subject to} \quad & f(\boldsymbol{\omega}) \le s_\omega, \ \ g(\mathbf{v}^k) \le s_{v^k}, \\ & k = 1, \cdots, m, \\ & \boldsymbol{\omega} \in \mathcal{B}_n. \end{aligned} \tag{4}$$

where $f(\boldsymbol{\omega}) = \|\boldsymbol{\omega}\|_0$, and $g(\mathbf{v}^k) = \|\mathbf{v}^k\|_0$. In Eq. (4), the objective function consisting of a Frobenius norm is smooth, and convex with respect to each group of variables $\boldsymbol{\omega}$, $\mathbf{u}^k$, and $\mathbf{v}^k$ $(k = 1, \cdots, m)$, but the two constraints $\|\boldsymbol{\omega}\|_0 \le s_\omega$ and $\|\mathbf{v}^k\|_0 \le s_{v^k}$ are nonconvex and nonsmooth. In addition, ℓ_0 vector norm constraints make Eq. (4) NP-hard.

3.2. The optimization algorithm

We adopt a proximal alternating linearized minimization (PALM) algorithm [5] to solve Eq. (4). It has been established that each bounded sequence generated by the PALM globally converges to a critical point of the problem (4).

Our algorithm alternates between optimizing each block of the variables $\boldsymbol{\omega}$, $\boldsymbol{u}$'s and $\boldsymbol{v}$'s (see Algorithm 1). The central idea is to, for each block of variables, perform a gradient step on the smooth part, but a proximal step on the nonsmooth part. Let $\boldsymbol{\omega}^t$, $(\mathbf{u}^k)^t$ and $(\mathbf{v}^k)^t$ be the current values at iteration t. For instance, to optimize $\boldsymbol{\omega}$, the algorithm minimizes a linearized approximation of the objective function, which is the gradient step $<\boldsymbol{\omega} - \boldsymbol{\omega}^t, \nabla_{\boldsymbol{\omega}} h>$ where $\nabla_{\boldsymbol{\omega}} h$ is the partial derivative of h with respect to $\boldsymbol{\omega}$. Then, $\text{argmin}\{<\boldsymbol{\omega} - \boldsymbol{\omega}^t, \nabla_{\boldsymbol{\omega}} h> + \frac{\gamma_\omega L_\omega}{2} \|\boldsymbol{\omega} - \boldsymbol{\omega}^t\|_2 : \ \|\boldsymbol{\omega}\|_0 \le s_\omega\}$ is a *well-defined* proximal map for f where $\gamma_\omega > 1$ is a pre-chosen constant and $L_{\boldsymbol{\omega}}$ is the Lipschitz modulus of $\nabla_{\boldsymbol{\omega}} h$.

Algorithm 1 Multi-view co-clustering with incomplete data.

```
Input: $\mathbf{X}^{k:=1,\cdots,m}$, $s_\omega$ and $s_{v_{k:=1,\cdots,m}}$
Output: $\omega$, $\mathbf{u}^k$ and $\mathbf{v}^k$ for $k = 1, \cdots, m$
1. Initialize $\boldsymbol{\omega}^0$, and $(\mathbf{v}^k)^0$, $(\mathbf{u}^k)^0$ for all $k = 1, \cdots, m$ and calculate the indicator matrix $\mathbf{A}^{k:=1,\cdots,m}$ from $\mathbf{X}^{k:=1,\cdots,m}$.
2. Compute $(\mathbf{u}^k)^t$, $\forall k = 1, \cdots, m$ according to Eq. (7).
3. Compute $(\mathbf{v}^k)^t$, $\forall k = 1, \cdots, m$ according to Eq. (10).
4. Compute $\boldsymbol{\omega}^t$ according to Eq. (11)
Repeat steps 2 - 4 until convergence (e.g., until $||\boldsymbol{\omega}^{t+1} - \boldsymbol{\omega}^t|| \le \varepsilon$, $||(\mathbf{u}^k)^{t+1} - (\mathbf{u}^k)^t|| \le \varepsilon$, and $||(\mathbf{v}^k)^{t+1} - (\mathbf{v}^k)^t|| \le \varepsilon$.)
```

Note that all partial derivatives of the objective function h are Lipschitz continuous so there exists a Lipschitz modulus. If $\|\nabla_{\boldsymbol{\omega}} h(\boldsymbol{\omega}_1) - \nabla_{\boldsymbol{\omega}} h(\boldsymbol{\omega}_2)\| \le L_{\boldsymbol{\omega}} \|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|$ for some constant $L_{\boldsymbol{\omega}} \ge 0$ and any $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$, then $\nabla_{\boldsymbol{\omega}} h$ is Lipschitz continuous and $L_{\boldsymbol{\omega}}$ is called the Lipschitz modulus of $\nabla_{\boldsymbol{\omega}} h$.

Given the values of $\mathbf{u}^k$, $\mathbf{v}^k$ and $\boldsymbol{\omega}$ at the iteration t, we now describe the procedure to update the variables at the iteration $t + 1$ as follows:

(a) Compute $(\mathbf{u}^k)^{t+1}$ using ω^t, $(\mathbf{u}^k)^t$ and $(\mathbf{v}^k)^t$

Because the update of $\boldsymbol{u}$'s are independent from each other, each $(\mathbf{u}^k)^{(t+1)}$ can be calculated separately. Let $\triangledown_{\mathbf{u}^k} h$ be the partial derivative of h at point $(\boldsymbol{\omega}^t, (\mathbf{u}^k)^t, (\mathbf{v}^k)^t)$ with respect to $\mathbf{u}^k$, In order to calculate $\triangledown_{\mathbf{u}^k} h$, we denote the ith rows of the matrices $\boldsymbol{A}^k$ and $\mathbf{X}^k$, and vectors $(\mathbf{u}^k)^t$ and $\boldsymbol{\omega}^t$ by $\mathbf{A}^k_{(i,\cdot)}$, $\mathbf{X}^k_{(i,\cdot)}$, u^k_i, and ω^t_i. Then each entry of $\triangledown_{\mathbf{u}^k} h$ can be calculated by

$$\triangledown_{u^k_i} h = \left(u^k_i \omega^t_i \mathbf{A}^k_{(i,\cdot)} \odot (\mathbf{v}^k)^{t^T} - \mathbf{A}^k_{(i,\cdot)} \odot \mathbf{X}^k_{(i,\cdot)}\right) \omega^t_i \left(\mathbf{A}^k_{(i,\cdot)} \odot (\mathbf{v}^k)^{t^T}\right)^T \tag{5}$$

The Lipschitz modulis of $\triangledown_{\boldsymbol{u}^k} h$ can be calculated using the following proposition.

Proposition 1. *Let $L_{\mathbf{u}^k}$ be the Lipschitz modulis of $\triangledown_{\mathbf{u}^k} h$ as defined in Eq. (5), then*

$$L_{\mathbf{u}^k} = \left\|\left({\omega^t_1}^2 \|\mathbf{A}^k_{(1,\cdot)} \odot (\mathbf{v}^k)^{t^T}\|_2^2, \cdots, {\omega^t_n}^2 \|\mathbf{A}^k_{(n,\cdot)} \odot (\mathbf{v}^k)^{t^T}\|_2^2\right)\right\|_2.$$

Proof. For any two given vectors $\mathbf{u}^k_1$ and $\mathbf{u}^k_2$, we can have

$$\begin{aligned} \|\triangledown_{\mathbf{u}^k} h(\mathbf{u}^k_1) - \triangledown_{\mathbf{u}^k} h(\mathbf{u}^k_2)\|_2 &= \left\| \begin{pmatrix} {\omega^t_1}^2 \|\mathbf{A}^k_{(1,\cdot)} \odot (\mathbf{v}^k)^{t^T}\|_2^2 & \dots & 0 \\ \vdots & \dots & \vdots \\ 0 & \dots & {\omega^t_n}^2 \|\mathbf{A}^k_{(n,\cdot)} \odot (\mathbf{v}^k)^{t^T}\|_2^2 \end{pmatrix} \begin{pmatrix} u^k_{11} - u^k_{21} \\ \vdots \\ u^k_{1n} - u^k_{2n} \end{pmatrix} \right\|_2 \\ &\leq \|({\omega^t_1}^2 \|\mathbf{A}^k_{(1,\cdot)} \odot (\mathbf{v}^k)^{t^T}\|_2^2, \cdots, {\omega^t_n}^2 \|\mathbf{A}^k_{(n,\cdot)} \odot (\mathbf{v}^k)^{t^T}\|_2^2)\|_2 \|\mathbf{u}^k_1 - \mathbf{u}^k_2\|_2 \end{aligned} \tag{6}$$

According to the definition of the Lipschitz moduli, $L_{\mathbf{u}^k} = \|({\omega^t_1}^2 \|\mathbf{A}^k_{(1,\cdot)} \odot (\mathbf{v}^k)^{t^T}\|_2^2, \cdots, {\omega^t_n}^2 \|\mathbf{A}^k_{(n,\cdot)} \odot (\mathbf{v}^k)^{t^T}\|_2^2)\|_2$ is the Lipschitz moduli of function $\triangledown_{\mathbf{u}^k} h$. □

We compute $(\mathbf{u}^k)^{t+1}$ by solving the following optimization problem:

$$\min_{\mathbf{u}^k} \quad < \mathbf{u}^k - (\mathbf{u}^k)^t, \triangledown_{\mathbf{u}^k} h > + \frac{\gamma_u L_{\mathbf{u}^k}}{2} \|\mathbf{u}^k - (\mathbf{u}^k)^t\|_2.$$

where $\gamma_u > 1$ is a constant and note that there is no nonsmooth part due to no regularizer on $\boldsymbol{u}$'s. This problem has an analytical solution as:

$$(\mathbf{u}^k)^{t+1} = (\mathbf{u}^k)^t - \frac{1}{\gamma_u L_{\mathbf{u}^k}} \triangledown_{\mathbf{u}^k} h \tag{7}$$

(b) Compute $(\mathbf{v}^k)^{t+1}$ using $\boldsymbol{\omega}^t$, $(\mathbf{u}^k)^{t+1}$ and $(\mathbf{v}^k)^t$

Similarly, each $\mathbf{v}^k$ can also be computed separately. Let $\mathbf{A}^k_{(\cdot,i)}$ and $\mathbf{X}^k_{(\cdot,i)}$ denote the ith columns of the matrices $\boldsymbol{A}^k$ and $\mathbf{X}^k$. Following a similar derivation to that in (a), we compute each entry of the partial derivatives $\triangledown_{\mathbf{v}^k} h$ and the Lipschitz modulis $L_{\mathbf{v}^k}$ as follows:

$$\triangledown_{\mathbf{v}^k_i} h = \left(\mathbf{A}^k_{(\cdot,i)} \odot \operatorname{diag}(\boldsymbol{\omega}^t)(\mathbf{u}^k)^{t+1}\right)^T \left(\mathbf{v}^{k^T}_i \mathbf{A}^k_{(\cdot,i)} \odot \operatorname{diag}(\boldsymbol{\omega}^t)(\mathbf{u}^k)^{t+1} - \mathbf{A}^k_{(\cdot,i)} \odot \mathbf{X}^k_{(\cdot,i)}\right)$$

and $L_{\mathbf{v}^k} = \|(l_1, \cdots, l_{d_k})\|_2$ where $l_s = \|\mathbf{A}^k_{(\cdot,s)} \odot \operatorname{diag}(\boldsymbol{\omega}^t)(\mathbf{u}^k)^{t+1}\|_2^2$ for $s = 1, \cdots, d_k$. In order to obtain the update for $\mathbf{v}^k$, we solve the proximal map:

$$\begin{aligned} &\min_{\mathbf{v}^k} < \mathbf{v}^k - (\mathbf{v}^k)^t, \triangledown_{\mathbf{v}^k} h > + \frac{\gamma_v L_{\mathbf{v}^k}}{2} \|\mathbf{v}^k - (\mathbf{v}^k)^t\|_2 \\ &\text{subject to} \quad \|\mathbf{v}^k\|_0 \leq s_{v^k}. \end{aligned}$$

Let $\delta_{s_{v^k}}(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ be a step function defined by:

$$\delta_{s_{v^k}}(\mathbf{x}) = \begin{cases} 0 & \|\mathbf{x}\|_0 \leq s_{v^k} \\ +\infty & \|\mathbf{x}\|_0 > s_{v^k}. \end{cases} \tag{8}$$

The above minimization problem can be converted to:

$$\min_{\mathbf{v}^k} \quad < \mathbf{v}^k - (\mathbf{v}^k)^t, \triangledown_{\mathbf{v}^k} h > + \frac{\gamma_v L_{\mathbf{v}^k}}{2} \|\mathbf{v}^k - (\mathbf{v}^k)^t\|_2 + \delta_{(s_{v^k})}(\mathbf{v}^k).$$

This problem can be proved to be equivalent to the following problem:

$$\min_{\mathbf{v}^k} \frac{\gamma_v L_{\mathbf{v}^k}}{2} \left\|\mathbf{v}^k - \left((\mathbf{v}^k)^t - \frac{1}{\gamma_v L_{\mathbf{v}^k}} \triangledown_{\mathbf{v}^k} h\right)\right\|_2 + \delta_{(s_{v^k})}(\mathbf{v}^k). \tag{9}$$

Let

$$(\tilde{\mathbf{v}}^k)^{t+1} = (\mathbf{v}^k)^t - \frac{1}{\gamma_v L_{\mathbf{v}^k}} \triangledown_{\mathbf{v}^k} h.$$

It can be shown that the optimal solution to Eq. (9) is the vector that keeps the original values in $(\tilde{\mathbf{v}}^k)^{t+1}$ at the positions whose absolute values are among the largest s_{v^k} of them. For instance, if s_{v^k} is 3, we rank the components in $(\tilde{\mathbf{v}}^k)^{t+1}$ in descending order according to their absolute values, and then choose the top three to maintain their values and set the rest to 0. We denote the corresponding threshold by α which is the minimum value among the s_{v^k} largest absolute values in $(\tilde{\mathbf{v}}^k)^{t+1}$, and compute $(\mathbf{v}^k)^{t+1}$ as follows:

$$(\mathbf{v}^k)_i^{t+1} = \begin{cases} (\tilde{\mathbf{v}}^k)_i^{t+1} & |(\tilde{\mathbf{v}}^k)_i^{t+1}| \geq \alpha, \\ 0 & |(\tilde{\mathbf{v}}^k)_i^{t+1}| < \alpha. \end{cases} \tag{10}$$

(c) Compute $\boldsymbol{\omega}^{t+1}$ using ω^t, $(\mathbf{u}^k)^{t+1}$ and $(\mathbf{v}^k)^{t+1}$

We compute each entry of the partial derivatives $\nabla_{\boldsymbol{\omega}} h$ and the Lipschitz modulis $L_{\boldsymbol{\omega}}$ as follows:

$$\nabla_{\omega_i} h = \textstyle\sum_{k=1}^m \left((u_i^k)^{(t+1)} \omega_i \mathbf{A}_{(i,\cdot)}^k \odot (\mathbf{v}^k)^{(t+1)^T} - \mathbf{A}_{(i,\cdot)}^k \odot \mathbf{X}_{(i,\cdot)}^k \right) (u_i^k)^{t+1} \left(\mathbf{A}_{(i,\cdot)}^k \odot (\mathbf{v}^k)^{(t+1)^T} \right)^T,$$

and $L_{\boldsymbol{\omega}} = \|(l_1, \cdots, l_n)\|_2$ where $l_s = \sum_{k=1}^m \left((u_s^k)^2\right)^{(t+1)} \|\mathbf{A}_{(s,\cdot)}^k \odot (\mathbf{v}^k)^{(t+1)^T}\|_2^2$ for $s = 1, \cdots, n$. We solve the following optimization problem for $\boldsymbol{\omega}^{t+1}$:

$$\begin{aligned} &\min_{\boldsymbol{\omega}} < \boldsymbol{\omega} - \boldsymbol{\omega}^t, \nabla_{\boldsymbol{\omega}} h > + \frac{\gamma_\omega L_{\boldsymbol{\omega}}}{2} \|\boldsymbol{\omega} - \boldsymbol{\omega}^t\|_2 \\ &\text{subject to} \quad \|\boldsymbol{\omega}\|_0 \leq s_\omega. \end{aligned}$$

By introducing the indicator function δ as in Eq. (8), and following the process for solving $\mathbf{v}^k$, we get the update formula for $\boldsymbol{\omega}$

$$\tilde{\boldsymbol{\omega}}^{t+1} = \boldsymbol{\omega}^t - \frac{1}{\gamma_\omega L_{\boldsymbol{\omega}}} \nabla_{\boldsymbol{\omega}} h.$$

Let β be the minimum value among the largest s_ω absolute values in $\tilde{\boldsymbol{\omega}}^{t+1}$. We compute $\boldsymbol{\omega}^{t+1}$ as follows:

$$\boldsymbol{\omega}_i^{t+1} = \begin{cases} \tilde{\boldsymbol{\omega}}_i^{t+1} & |\tilde{\boldsymbol{\omega}}_i^{t+1}| \geq \beta \\ 0 & |\tilde{\boldsymbol{\omega}}_i^{t+1}| < \beta. \end{cases} \tag{11}$$

Algorithm 1 summarizes the above steps. By applying this algorithm, we obtain a set of row and column clusters. The rows corresponding to non-zero values in ω indicate a subject cluster and the columns corresponding to non-zero values in each $\mathbf{v}^k$ indicate the selected variables in the kth view. In order to obtain the next set of row and column clusters, we need to deflate the data matrix by removing the rows corresponding to the subjects already identified in a row cluster. We then repeat Algorithm 1 on the updated data matrix. There are other ways to deflate the data matrix, such as computing $\mathbf{X} - \text{diag}(\boldsymbol{\omega})\mathbf{u}\mathbf{v}^T$ for each view, which will however create clusters with overlapping subjects.

The computational complexity of Algorithm 1 at each iteration is $O(nmd)$ where d is the maximum of $d_1, \cdots, d_m$. Since this computational complexity is in a linear order of the problem dimensions, it is very efficient. Due to the independence in calculating the updates of $\mathbf{u}^k$ and $\mathbf{v}^k$, Algorithm 1 is ready to be parallelized and distributed if more processors are available. Compared with the objective function of problem (1), the additional indicator term $\boldsymbol{A}^k$ in the objective function of problem (4) is a known constant matrix that does not affect the convergence property of Algorithm 1. Algorithm 1 still globally converges to a critical point of the problem (4).

4. Experiments

We validated the proposed approach in both simulation studies and the analysis of the clinical data collected in our heroin treatment study.

4.1. Simulations

We generated three sets of data. We first created a dataset without missing values, and removed data values in a way that simulated the missing patterns observed in the treatment data. Then for the incomplete dataset we created, we used an imputation method to impute the removed entries. In this paper, we adopted a widely used multiple imputation method in [32] where it was compared against and outperformed a suite of other algorithms.

We generated a dataset with implanted diagonal block structures that corresponded to row and column clusters. Two views of data for 1000 subjects were created. There were 12 variables in view 1, and 15 variables in view 2. The data matrix of each view was created by randomly setting an entry to 0 or 1 with different probabilities that were determined according to the prefixed block structures. Precisely, we started from a data matrix of all zeros. Then we reset data entries inside and outside the blocks to 1 with a probability of 0.8 and 0.2, respectively. For better illustration, we aligned the subjects in the two views and indexed them from 1 to 1000; and indexed the variables using consecutive numbers starting from 1. Fig. 2 demonstrates the block structures in the two views. View 1 was designed to have two blocks. The first block consisted of subjects from 1 to 400 and variables from 1 to 3. The second block included 200 subjects indexed from 481 to 680 and

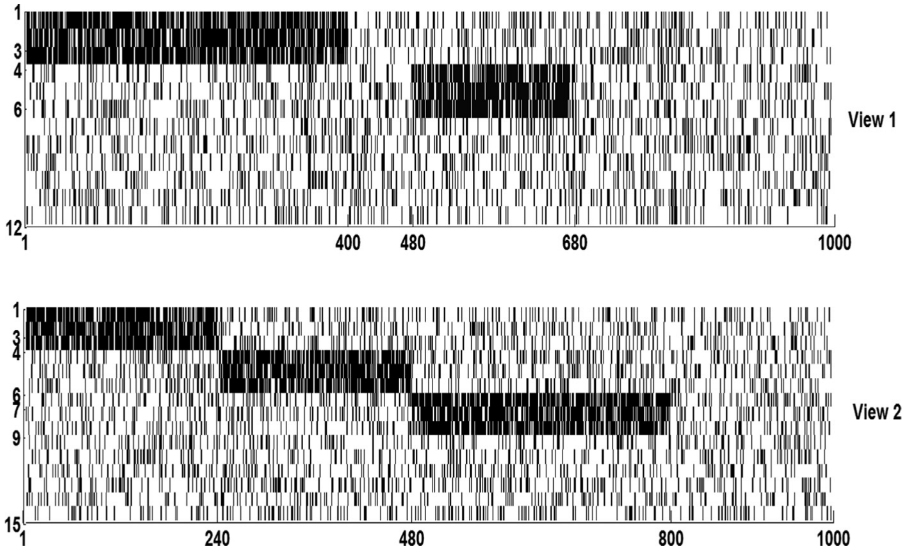

Fig. 2. The simulated block data structure, the numbers in the vertical axis represent the associated variables, the number in the horizontal axis represent the subject index.

the 4th to the 6th variables. Three blocks were included in view 2. The first block contained subjects from 1 to 240 and was associated with the first three variables. The second block consisted of subjects from 241 to 480 and was associated with the 4th to the 6th variables. The last block included 320 subjects indexed from 481 to 800 and was relevant to the 7th to the 9th variables. By comparing the two views, there would be three consistent clusters (i.e. containing the same subjects) across the two views. We created an incomplete data matrix by removing some entries from the synthesized full data matrix. We removed the data on the features 4–12 for the subjects 1 to 400 and the data on the features 7–12 for the subjects 481 to 680 in the first view. For the second view, we removed data on the features 4–15 for the subjects 1–240, and data on the features 7–15 for the subjects 241–480, and data on the features 10–15 for the subjects 481–800. These steps created non-random missing patterns. We then removed ρ% ($\rho =$ 10, 20, 30, 40, and 50) of entries from the remaining data randomly (with a uniform distribution over the remaining data), which created data of missing at random. In the subsequent experiments, for each ρ value, we repeated the data removal process five times and reported the average performance of each algorithm.

We used the multiple imputation methods discussed in [32] to impute the missing values for the incomplete synthetic data. Overall, three types of datasets were created in our simulations: (1) the full synthetic data matrix; (2) the incomplete data matrix; and (3) the data matrices where missing values were imputed from the incomplete matrix by multiple imputation. The multi-view co-clustering method in [35] was applied to the full and imputed data matrices to identify clusters whereas our proposed method - Algorithm 1 - that can directly handle missing values, was applied to the incomplete data matrix. We set each method to returning three clusters, and used three standard metrics to evaluate the clustering performance.

- **Normalized Mutual Information (NMI).** For two random variables X and Y, the NMI is defined as:

$$\mathrm{NMI}(\mathbf{X},\mathbf{Y}) = \frac{I(\mathbf{X},\mathbf{Y})}{\sqrt{H(\mathbf{X})H(\mathbf{Y})}} \tag{12}$$

  where $I(\boldsymbol{X}, \boldsymbol{Y})$ is the mutual information between $\boldsymbol{X}$ and $\boldsymbol{Y}$, while $H(\boldsymbol{X})$ and $H(\boldsymbol{Y})$ are the entropies of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Clearly, NMI takes the value from [0,1], the higher NMI means the better the clustering performance.
- **Accuracy (Acc).** Accuracy discovers the one-to-one relationship between clusters and classes and measures the extent to which each cluster contained data points from the corresponding class. It sums up the whole matching degree between all pair class-clusters.

$$\mathrm{Acc} = \frac{1}{N}\max\left(\sum_{X_i,Y_j} N(X_i,Y_j)\right) \tag{13}$$

  where X_i denotes the ith cluster in the final results, and Y_j is the jth class (true cluster) in the synthetic data. $N(X_i, Y_j)$ is the number of entities which belong to class j but are assigned to cluster i. Accuracy computes the maximum sum of $N(X_i, Y_j)$ for all pairs of clusters and classes, and these pairs have no overlaps. The greater clustering accuracy means the better clustering performance.

Table 2
Comparison of clustering performance in terms of the normalized mutual information (NMI) measure with different ρ values. "Full" indicates the result obtained on the full data matrix, "MI" indicates the clustering results on imputed data, "InCo" corresponds to the results of our method on the incomplete data. Note the digits before / indicate the NMI values (%) while the digits after / indicate the standard deviations (%).

	Trial 1		Trial 2		Trial 3		Trial 4		Trial 5	
Full	39.91		44.71		44.17		46.22		44.85	
ρ	MI	InCo	MI	InCo	MI	InCo	MI	InCo	MI	InCo
0.1	36.18/1.92	**36.91**/3.54	**36.38**/1.65	36.32/1.68	35.96/1.44	**40.84**/2.85	36.03/1.74	**38.38**/2.39	37.38/1.68	**40.64**/2.62
0.2	29.87/1.20	**36.23**/1.87	30.34/2.17	**36.60**/0.88	29.33/1.58	**41.29**/1.75	30.02/1.35	**37.57**/2.16	28.03/1.47	**39.32**/1.61
0.3	23.63/2.65	**37.16**/0.77	23.81/1.24	**37.52**/1.47	25.05/3.09	**39.78**/2.40	24.07/2.92	**38.78**/1.47	26.96/1.97	**39.07**/1.96
0.4	20.89/1.12	**34.27**/0.96	21.84/1.18	**35.44**/1.72	21.01/1.35	**36.36**/1.16	21.90/1.26	**33.96**/1.35	24.42/1.45	**36.76**/1.74
0.5	18.24/2.04	**29.91**/3.03	17.32/1.00	**31.07**/1.84	17.11/3.03	**30.21**/2.70	17.77/2.41	**30.31**/1.18	17.41/1.92	**31.01**/4.00

Table 3
Comparison of clustering performance in terms of the normalized mutual information (NMI) measure with different ρ values. "Full" indicates the result obtained on the full data matrix, "MI" indicates the clustering results on imputed data, "InCo" corresponds to the results of our method on the incomplete data, "PVC" indicates the result of partial multi-view clustering, "WMNF21" indicates the result of multiple incomplete views clustering via weighted nonnegative matrix factorization with $L2$, 1 regularization, CoKL indicates the result of clustering on multiple incomplete datasets via collective kernel learning. Note the digits before / indicate the NMI values (%) while the digits after / indicate the standard deviations (%).

Full	44.85				
ρ	MI	InCo	PVC	WMNF21	CoKL
0.1	37.38/1.68	**40.64**/2.62	28.10/7.13	29.60/6.25	26.26/6.43
0.2	28.03/1.47	**39.32**/1.61	24.81/3.23	28.98/6.26	30.23/2.77
0.3	26.96/1.97	**39.07**/1.96	20.60/0.51	22.62/3.85	25.15/6.13
0.4	24.42/1.45	**36.76**/1.74	17.23/4.11	21.57/0.83	19.96/5.89
0.5	17.41/1.92	**31.01**/4.00	13.52/1.59	14.09/1.77	24.01/2.07

- **Rand Index.** Rand index can be considered as an alternative to the information-theoretic interpretation of clustering NMI, it views clustering as a series of decisions, one for each of the $n(n-1)/2$ pairs of subjects in the sample. We want to assign two subjects to the same cluster if and only if they are similar. A true positive (TP) decision assigns two similar subjects to the same cluster, a true negative (TN) decision assigns two dissimilar subjects to different clusters. There are two types of errors we can commit. A (FP) decision assigns two dissimilar subjects to the same cluster. A (FN) decision assigns two similar subjects to different clusters. Rand index measures the percentage of decisions that are correct.

$$\mathrm{RI} = \frac{TP + TN}{TP + FP + TN + FN} \tag{14}$$

Due to the randomness in the data creation process, we generated five datasets following the above procedure to evaluate if the comparison was consistent. Tables 2, 4, 6 summarize and compare the performance of our method on the incomplete datasets against the multi-view co-clustering method after imputation. For each of the five datasets, we reported the mean NMI, Acc, RI values and the standard deviations for each different ρ value. The five synthetic datasets corresponded to the columns of Tables 2, 4, 6. Since obviously the best results were obtained on the full dataset, we highlighted the second best results with bold font. We can find that, in terms of the NMI measure, the proposed method worked better on the incomplete data than the multi-view co-clustering method on the imputed data by multiple imputation in all of the comparison settings except when $\rho = 0.1$. In terms of the Acc measure, for $\rho = 0.1, 0.2, 0.3,$ our proposed method performed better than multi-view co-clustering on the data after multiple imputation, but for some cases in $\rho = 0.4$ or 0.5, our method did not demonstrate its superiority, this may be because the useful information available was too limited. In terms of RI measure, our method outperformed the counterpart in every noise level. In summary, for the three metrics, they performed almost consistently and our method demonstrated its effectiveness.

Furthermore, we compared our approach against three recent multi-view incomplete clustering methods including PVC [22], WMNF21 [31], and the method of clustering on multiple incomplete datasets via collective kernel learning (CoKL) [30].These methods were run on the above generated incomplete datasets. Since all of the three methods could not deal with the case of missing any number of values in each view, we imputed data for half of the subjects to have both views, a quarter of them to have one of the views using multiple imputation [32]. On the same dataset (Trial 5), we reported the NMI values of all these methods in Tables 3, 5, 7. Compared with PVC, WMNF21, and CoKL, we see that our method performed consistently the best over all choices of ρ in terms of the metrics NMI and RI. In terms of the Acc measure, CoKL performed better than our method when $\rho = 0.1$ and 0.4, but the margin was rather small. Since PVC, WMNF21 and CoKL required the

Table 4
Comparison of clustering performance in the accuracy measure with different ρ values. “Full” indicates the result obtained on the full data matrix, “MI” indicates the clustering results on imputed data, “InCo” corresponds to the results of our method on the incomplete data. Note the digits before / indicate the accuracy values (%) while the digits after / indicate the standard deviations (%).

	Trial 1		Trial 2		Trial 3		Trial 4		Trial 5	
Full	43.2		44.5		44.7		45.5		44.80	
ρ	MI	InCo	MI	InCo	MI	InCo	MI	InCo	MI	InCo
0.1	32.83/0.57	**34.50**/5.5	32.00/0.68	**34.12**/5.74	33.34/0.69	**38.34**/5.58	30.80/0.83	**38.58**/1.42	33.58/0.75	**40.20**/1.29
0.2	34.30/0.74	**37.21**/1.81	33.74/1.43	**34.36**/5.67	31.92/0.84	**38.36**/5.36	33.04/1.16	**36.58**/4.97	33.39/0.98	**39.72**/1.04
0.3	32.45/0.20	**32.52**/5.66	32.39/2.14	**33.68**/5.21	32.64/2.75	**38.30**/5.44	32.50/2.57	**37.80**/4.98	33.25/1.43	**37.98**/5.19
0.4	**33.93**/0.42	29.32/4.22	33.72/2.87	**33.84**/4.78	30.32/4.38	**33.11**/1.97	32.18/3.94	**33.86**/4.98	**33.06**/3.82	32.88/5.71
0.5	32.67/1.10	**32.92**/4.71	32.60/1.23	**32.72**/5.80	30.62/5.19	**35.03**/2.06	**33.70**/1.89	33.58/3.93	31.88/1.59	**33.96**/5.51

Table 5
Comparison of clustering performance in terms of the Acc measure with different ρ values. “Full” indicates the result obtained on the full data matrix, “MI” indicates the clustering results on imputed data, “InCo” corresponds to the results of our method on the incomplete data, “PVC” indicates the result of partial multi-view clustering, “WMNF21” indicates the result of multiple incomplete views clustering via weighted nonnegative matrix factorization with $L2$, 1 regularization, CoKL indicates the result of clustering on multiple incomplete datasets via collective kernel learning. Note the digits before / indicate the Acc values (%) while the digits after / indicate the standard deviations (%).

Full	44.85				
ρ	MI	InCo	PVC	WMNF21	CoKL
0.1	33.58/0.75	40.20/1.29	21.71/6.56	40.26/1.23	**40.73**/3.37
0.2	33.39/0.98	**39.72**/1.04	18.93/3.08	38.93/6.29	38.83/2.74
0.3	33.25/1.43	**37.98**/5.19	15.13/0.98	35.03/5.88	37.53/5.32
0.4	33.06/3.82	32.88/5.71	12.31/3.40	31.60/1.99	**32.96**/4.88
0.5	31.88/1.59	**33.96**/5.51	10.56/1.31	30.93/2.24	32.10/1.13

Table 6
Comparison of clustering performance in terms of the RI measure with different ρ values. “Full” indicates the result obtained on the full data matrix, “MI” indicates the clustering results on imputed data, “InCo” corresponds to the results of our method on the incomplete data. Note the digits before / indicate the RI values (%) while the digits after / indicate the standard deviations (%).

	Trial 1		Trial 2		Trial 3		Trial 4		Trial 5	
Full	74.69		75.93		75.83		76.91		76.35	
ρ	MI	InCo	MI	InCo	MI	InCo	MI	InCo	MI	InCo
0.1	67.82/0.71	**71.88**/2.89	67.64/0.87	**71.54**/2.69	68.49/0.70	**73.64**/2.99	66.75/0.71	**72.98**/1.47	67.38/0.54	**74.34**/1.46
0.2	68.80/0.52	**72.56**/0.94	68.18/0.99	**70.98**/2.41	66.94/0.75	**74.25**/2.89	67.53/0.83	**71.81**/2.50	67.51/0.95	**73.62**/0.88
0.3	67.50/1.24	**71.02**/2.59	67.87/0.64	**70.77**/2.40	66.95/1.29	**73.94**/2.64	67.08/1.25	**73.51**/2.20	67.70/0.65	**73.55**/2.41
0.4	68.53/2.31	**69.14**/1.63	68.28/1.48	**71.00**/1.75	67.48/2.24	**70.02**/2.57	66.70/2.19	**69.85**/2.67	67.41/2.23	**70.60**/2.44
0.5	67.63/1.02	**70.48**/2.07	67.96/0.90	**69.03**/2.53	68.66/1.46	**69.34**/3.22	67.80/0.85	**69.30**/2.47	66.88/0.70	**70.55**/3.02

Table 7
Comparison of clustering performance in terms of the RI measure with different ρ values. “Full” indicates the result obtained on the full data matrix, “MI” indicates the clustering results on imputed data, “InCo” corresponds to the results of our method on the incomplete data, “PVC” indicates the result of partial multi-view clustering, “WMNF21” indicates the result of multiple incomplete views clustering via weighted nonnegative matrix factorization with $L2$, 1 regularization, CoKL indicates the result of clustering on multiple incomplete datasets via collective kernel learning. Note the digits before / indicate the accuracy values (%) while the digits after / indicate the standard deviations (%).

Full	76.35				
ρ	MI	InCo	PVC	WMNF21	CoKL
0.1	67.38/0.54	**74.34**/1.46	54.75/7.46	71.10/3.67	68.71/4.14
0.2	67.51/0.95	**73.62**/0.88	53.54/4.42	70.58/2.49	66.86/0.79
0.3	67.70/0.65	**73.55**/2.41	47.18/4.39	68.78/1.37	65.88/0.53
0.4	67.41/2.23	**70.60**/2.44	44.26/2.99	67.63/0.67	65.77/1.01
0.5	66.88/0.70	**70.55**/3.02	44.10/1.22	65.10/1.18	63.94/0.38

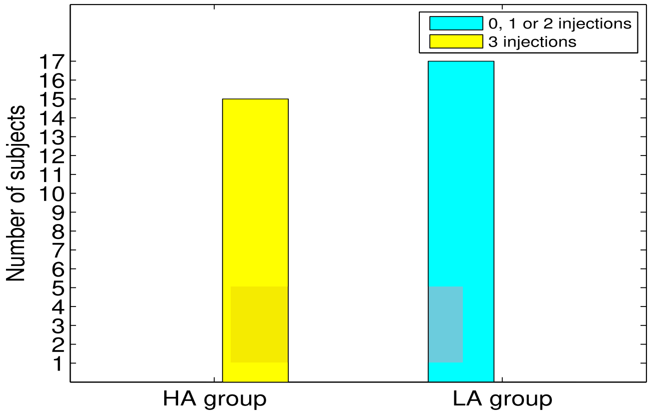

Fig. 3. The adherence characteristics of the two clusters (high adherence (HA) versus low adherence (LA)) obtained by our algorithm when variables were grouped in the three views according to variable type.

complete feature sets in a view, they performed even worse than just running the multi-view co-clustering with imputed data (i.e., MI).

4.2. Case study: XR-NTX treatment of heroin use disorder

We performed two experiments using our XR-NTX treatment dataset: one with the views defined by variable types, and the other with the views defined via time windows as discussed in Section 2. For each experiment, we will discuss subject characteristics in the resultant clusters and the features selected by our algorithm for each cluster. We first grouped all features in the dataset according to variable types in order to study the connections and correlations between survey variables, lab test results and vital signs. In the second experiment, variables were grouped by different time window to study if baseline or during treatment variables had connections to the outcomes observed in the post-treatment window.

In both experiments, we set $s_\omega = 15$ and $s_{v_k} = 3$ for all $1 \le k \le m$ in Eq. (1) by a cross validation tuning process using half of the sample. We initialized ω^0, $(\mathbf{u}^k)^0$, and $(\mathbf{v}^k)^0$ in a way such that all entries were equal to 1 for all $1 \le k \le m$. In each experiment, two clusters were generated by our algorithm. One of the widely-used ways to define the XR-NTX adherence [39] is to evaluate if the number of injections a participant received is out of the maximum available. The concurrent validity of the clusters was validated in terms of the number of injections which was not used as a basis of the cluster analysis. In our study, the subjects who received three total injections were considered highly adherent to the treatment comparing to those who received less.

4.2.1. Connections between different variable types

We partitioned the subjects jointly on the basis of three views: surveys variables, lab test results, and vital signs. The first cluster we obtained consisted of 15 subjects, all of whom happened to receive all three injections whereas the second cluster consisted of 17 subjects, none of them finished up the three injections as shown in Fig. 3. We hence named the two clusters, respectively, the high adherence group (HA) and the low adherence group (LA).

From the variables that we used as the basis of cluster analysis, our algorithm automatically selected the most relevant ones in each of the views. We plotted the mean values of the selected variables in each view for each group in Fig. 4. We observed that subjects in the HA group increased their craving level for other drugs (Δ*Pre_Cra_Oth*) after the cue exposure at baseline whereas subjects in LA group craved less instead. Although it was the only variable selected in this view, meaning that this variable itself was enough to distinguish the two clusters in this view, we also observed that subjects in the HA group increased their craving for heroin (Δ*Pre_Cra_Heroin*) (by a rating of 2.93 on average) more than the subjects in the LA group (by a rating of 1.07 on average) after exposing to cues. In the lab test view, subjects in the HA group showed on average higher tetrahydrocannabinol (*Pre_THC*) and cocaine (*Pre_COC*) levels in the urine drug screen at baseline. This result demonstrated that subjects in HA group tended to take these two drugs as well. In the vital signs view, the HA subjects showed decreased mean arterial pressure (Δ*Pre_MAP*) after the cue exposure at baseline. A study with a larger sample may be needed to validate these observations.

To further examine whether or not the selected variables had statistical significance in distinguishing between categories of XR-NTX treatment adherence, we performed an additional association test. Because most of these variables were categorical variables, we first applied the multiple correspondence analysis (MCA) [1] to all the selected variables to reduce dimension and identify the first principal component. MCA is similar to the principal component analysis that is applicable to continuous variables, but it is able to cope with categorical variables with missing values. We then used the resultant principal dimension as the predictor to predict whether a subject would complete all three injections, and we observed a p value of 0.0258, which demonstrated that the association was statistically significant.

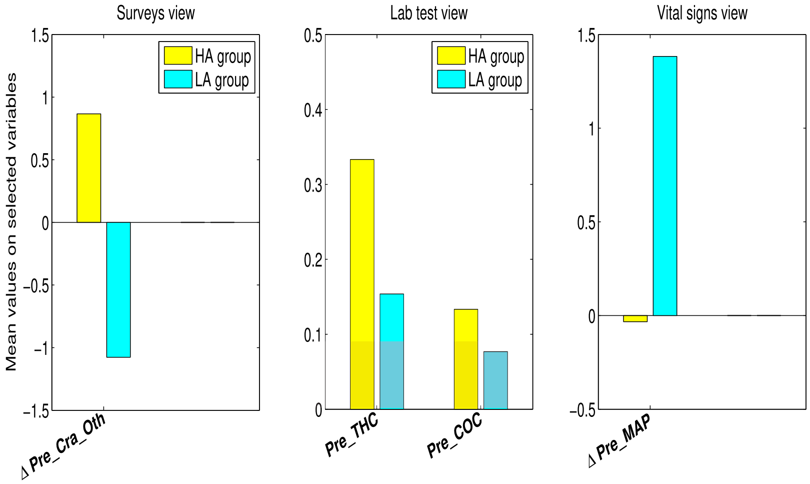

Fig. 4. The mean values of the selected variables by cluster when data were grouped in the three views according to variable type. **Abbreviation**: △Pre_Cra_Oth, change in craving for other drugs after cue exposure at baseline; Pre_THC, tetrahydrocannabinol level in urine drug screen at baseline; Pre_COC, cocaine level in urine drug screen at baseline; △Pre_MAP, change in mean arterial pressure after cue exposure at baseline.

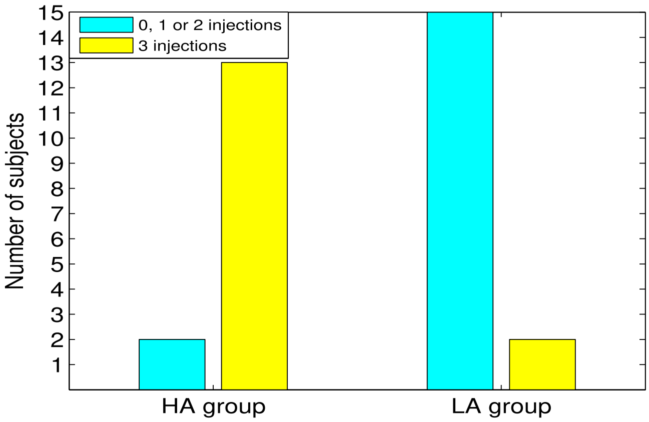

Fig. 5. The adherence characteristics of the two clusters (high adherence (HA) versus low adherence (LA)) obtained by our algorithm when variables were grouped in three time windows.

4.2.2. Connections between different time windows

We then identified consistent clusters across the three time windows: Pre-treatment (baseline), On-treatment, and Post-treatment. As shown in Fig. 5, the first cluster also contained 15 subjects, 13 of them receiving all three injections. There were two subjects in this group who only finished one injection. The second cluster contained two subjects with three injections and the remaining subjects did not finish all injections. We hence still named these two clusters as the HA group and the LA group, respectively, without notation confusion (although they were two different clusters from those in the first setting).

To study the variables used in the analysis, we plotted the mean values of the selected variables in the different time windows in Fig. 6. Four variables in the Pre-treatment window, two variables in the On-treatment window, and two variables in the Post-treatment window, were selected. Subjects in the HA group increased their craving level for heroin (△*Pre_Cra_Heroin*) after exposing to cues at baseline but decreased their subjective heroin withdrawal scale (△*Pre_SOWS*). Similar to the first setting, we also observed that the HA subjects showed elevated craving for other drugs (△*Pre_Cra_Oth*), and elevated withdrawal scores for other drugs (△*Pre_WD_Oth*), after cues at baseline. However, these subjects showed decreased craving (△*On_Cra_Oth*) and withdrawal scores (△*On_WD_Oth*) for other drugs once the treatment began. At post-treatment, subjects in the HA group showed a lower average level of telrahydrocannabinol (*Post_THC*) in their urine tests than that in the LA group, and they felt less “high” for heroin when exposing to cues after completing the treatment. Although not shown in Fig. 6, for the subjects in the HA group, craving for heroin was increased by 2.93 after exposing to stimuli at baseline but only by 1.18 at post-treatment.

To gain more insights into whether the selected variables were statistically significantly associated with the XR-NTX treatment adherence, we first applied the MCA to identify the first principal component of these variables and used it as

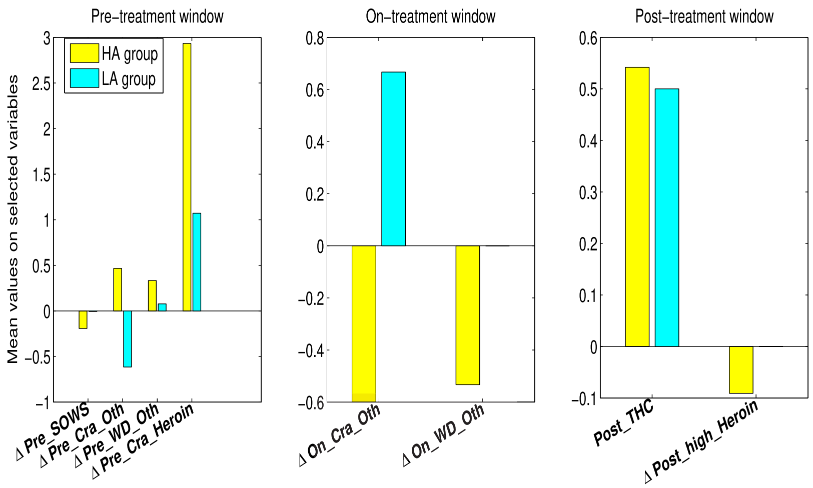

Fig. 6. The mean values of the selected variables by cluster when data were grouped in three time windows. *Abbreviation*: ΔPre_SOWS, change in the subjective opioid withdrawal scale after cue exposure at baseline; ΔPre_Cra_Oth, change in craving for other drugs after cue exposure at baseline; ΔPre_WD_Oth, change in withdrawal for other drugs after cue exposure at baseline; ΔPre_Cra_Heroin, change in craving for heroin after cue exposure at baseline; ΔOn_Cra_Oth, change in craving for other drugs after cue exposure during treatment; ΔOn_WD_Oth, change in withdrawal for other drugs after cue exposure during treatment; Post_THC, tetrahydrocannabinol urine drug screen after treatment; ΔPost_high_Heroin, change in feeling "high" for heroin after cue exposure after treatment.

Table 8
The statistical test results of the top five significant (the smallest 5 p values) variables between the two groups for the time window setting.

The top 5 variables	p value
Change in craving for heroin after cue exposure at baseline	**0.0316**
Beck depression inventory at baseline	0.1020
Change in feeling high for other drugs after cue exposure at baseline	0.1857
Change in subjective opioid withdrawal scale after cue exposure at baseline	0.2501
Change in craving for other drugs after cue exposure at baseline	0.2737

the predictor to predict if a subject would finish all three injections in a logistic regression association test. We obtained a p value of 0.00829, showing the statistical significance of these selected variables as a whole. Because in this setting, we were more interested in finding if baseline variables can predict treatment outcome at the post-treatment stage, we performed another set of association tests. In particular, we carried out a t-test on all the individual variables between the HA and LA subjects. Table 8 shows the top five variables together with their corresponding p values. Based on this table, the difference in craving for heroin after cue exposure at baseline (Δ*Pre_Cra_Heroin*) was significantly different between the subjects in the HA group and those in the LA group at a significance level of $p < 0.05$, demonstrating that changes in craving for heroin in response to cues at baseline could be a useful predictor for patient adherence to XR-NTX.

5. Discussion and conclusion

As data acquisition technologies advance, more and more data collected in real-world applications are from heterogeneous sources, resulting in multi-view datasets. Different views may provide complementary information. Cluster analysis in any single view may miss important cluster characteristics from other views. Simply concatenating all views together cannot guarantee finding clusters recognizable in individual views. To exploit such multiple view information, we have adopted the much-needed multi-view learning methods.

A challenge in practical multi-view applications comes from missing values in the different views. It is common that there are missing values in each view of the data. Usually researchers use imputation methods [27] such as simple ones that use mean value or zero to complete the data. Among the existing imputation methods, multiple imputation [12] is a popular choice. In our simulation study, we have compared the multiple imputation method with our generalized multi-view co-clustering method that can directly cluster subjects based on incomplete data. We observed that when very few missing values were present, multiple imputation performed as well as or even better than our method. However, when the number of missing values increases, our method clearly showed its superiority. Another way to deal with an incomplete dataset is to simply exclude subjects (or samples) with missing values. However, this method dramatically reduces the available sample size, causing insufficient data for subsequent clustering, which was the case in both our synthetic datasets and our heroin treatment study. There would be none left for analysis if we had removed subject with any missing values in the heroin treatment dataset. Most state-of-the-art multi-view incomplete clustering methods [22,30,31] can only deal with the situation where the subjects miss some of the views entirely rather missing any number of variables in one view. Hence, the proposed method can be a better alternative.

In the heroin treatment study, the newly proposed method has helped us to determine that the difference in craving for heroin after exposure to visual heroin cues at baseline can be a good predictor of whether a patient will adhere to the treatment. Besides the change in craving for heroin, craving for other drugs can be another important feature to predict adherence to heroin treatment. These features may have combined effects on adherence to XR-NTX. When studying the connections between features in the pre-, during- and post- treatment time windows, we found that subjects who tended to complete the treatment actually had elevated craving and withdrawal scores when exposing to cues at baseline but they decreased these features after cue exposure once the treatment started. This finding might be counter-intuitive against medical practice where it is recognized that the patients who have less craving tend to adhere to XR-NTX, thus warranting further investigations using larger samples in independent studies.

By cross referencing the post-stimuli craving for heroin at baseline and at post-treatment, we observed that XR-NTX can effectively reduce craving if patients adhere to the treatment. Furthermore, the adherent patients showed increased withdrawal scores after stimuli at baseline but once treatment started, their withdrawal scores became decreased after stimuli. These observations agree with prior findings such as in [36] where XR-NTX shows efficacy for decreasing craving, maintaining abstinence, improving retention, and preventing relapse among opioid dependent patients following detoxification. Prior studies also show that concurrent use of cocaine is quite common in opioid dependent individuals [17,25,38]. In our experiments, we found that craving for other drugs (e.g., cocaine) at baseline was a major predictor for XR-NTX treatment adherence, which may reflect the concurrent use of these drugs.

One of our previous studies used the present dataset to identify neural correlates of XR-NTX treatment adherence [38]. There were two major differences between the current and prior studies. We used an advanced multi-view cluster analysis method that was able to explore which baseline variables were predictive of the adherence beforehand by concurrently grouping subjects at the three time windows. The second difference was that we proposed a new multi-view method which could directly handle the significant amount of missing data entries commonly encountered in treatment studies, which might provide a powerful alternative to coping with incomplete data in future investigations.

In conclusion, we proposed and tested a novel multi-view co-clustering formulation that can handle incomplete data by introducing an indicator matrix to the original formulation without the need of data imputation. This enhanced multi-view approach has been carefully evaluated in simulation studies, showing advantages over several other alternative methods such as removing incomplete samples or multiple imputation. Then this approach was applied to the incomplete data collected in the heroin treatment study. We used the proposed approach in two separate settings: in three variable-type views as well as in three time window views, and obtained very similar patient groupings. In each setting, a group of subjects that were highly adherent to the XR-NTX treatment was identified. We found several variables, such as craving for heroin in response to visual drug stimuli at baseline, that were predictive of treatment adherence. These results come with some limitations. Although the size of our sample is common in treatment studies, especially with repeated brain imaging, it is relatively small, which might limit the statistical power of many analytics. With larger samples, taking into account the gender, race, age of patients may give us more insights into the predictors of patient adherence to XR-NTX and possibly other treatments of heoin dependence. Other related factors may also be included as variables in our proposed analysis, such as monetary and nonmonetary incentives as reported in [17] that incentives for naltrexone adherence increased opiate abstinence in heroin-dependent adults.

Conflict of interest

None.

Acknowledgment

This work was supported by National Institutes of Health (NIH) grants R01DA037349 and K02DA043063, and National Science Foundation (NSF) grants DBI-1356655, CCF-1514357, and IIS-1718738. Jinbo Bi was also supported by NSF grants IIS-1320586, IIS-1407205, and IIS-1447711. An-Li Wang was also supported by NIH grant R00HD84746.

References

[1] H. Abdi, D. Valentin, Multiple correspondence analysis, Encycl. Meas. Stat. (2007) 651–657.
[2] M.F. Balcan, A. Blum, K. Yang, Co-training and expansion: towards bridging theory and practice, in: L.K. Saul, Y. Weiss, L. Bottou (Eds.), Advances in Neural Information Processing Systems 17, MIT Press, Cambridge, MA, 2005, pp. 89–96.
[3] A. Bargiela, W. Pedrycz, The roots of granular computing, in: 2006 IEEE International Conference on Granular Computing, IEEE, 2006, pp. 806–809.
[4] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: Proceedings of the 11th Annual Conference on Computational Learning Theory, ACM, New York, NY, USA, 1998, pp. 92–100, doi:10.1145/279943.279962.
[5] J. Bolte, S. Sabach, M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, Math. Program. 146 (1) (2014) 459–494.
[6] X. Cai, W. Li, R. Zhang, Enhancing diversity and coverage of document summaries through subspace clustering and clustering-based optimization, Inf. Sci. 279 (2014) 764–775.
[7] G. Chao, Discriminative k-means laplacian clustering, Neural Process. Lett. (2018) 1–13.

[8] G. Chao, S. Sun, Alternative multi-view maximum entropy discrimination, IEEE Trans. Neural Netw. Learn. Syst. 27 (7) (2016) 1445–1456.

[9] G. Chao, S. Sun, Consensus and complementarity based maximum entropy discrimination for multi-view classification, Inf. Sci. 367–368 (2016) 296–310.

[10] G. Chao, S. Sun, Multi-kernel maximum entropy discrimination for multi-view learning, Intell. Data Anal. 20 (2016) 481–493.

[11] G. Chao, S. Sun, Semi-supervised multi-view maximum entropy discrimination with expectation laplacian regularization, Inf. Fusion 45 (2019) 296–306.

[12] N. Eisemann, A. Waldmann, A. Katalinic, Imputation of missing values of tumour stage in population-based cancer registration, BMC Med. Res. Method. 11 (2011), doi:10.1186/1471-2288-11-129.

[13] H. Hoffmann, Unsupervised Learning of Visuomotor Associations, vol. 11, Bielefeld University, 2005.

[14] S. Huang, Z. Kang, I.W. Tsang, Z. Xu, Auto-weighted multi-view clustering via kernelized graph learning, Pattern Recognit. 88 (2019) 174–184.

[15] S. Huang, Z. Kang, Z. Xu, Self-weighted multi-view clustering with soft capped norm, Knowl. Based Syst. 158 (2018) 1–8.

[16] J.G. Ibrahim, G. Molenberghs, Missing data methods in longitudinal studies: a review, Test 18 (2009) 1–43.

[17] B. Jarvis, A. Holtyn, A. Defulio, K. Dunn, J. Everly, J. Leoutsakos, A. Umbricht, M. Fingerhood, G. Bigelow, Effects of incentives for naltrexone adherence on opiate abstinence in heroin-dependent adults, Addiction (2017) 830–837.

[18] Z. Kang, X. Lu, J. Yi, Z. Xu, Self-weighted multiple kernel learning for graph-based clustering and semi-supervised classification, arXiv:1806.07697(2018).

[19] D.D. Langleben, K. Ruparel, I. Elman, J.W. Loughead, E.L. Busch, J. Cornish, K.G. Lynch, E.S. Nuwayser, A.R. Childress, C.P. O'Brien, Extended-release naltrexone modulates brain response to drug cues in abstinent heroin-dependent patients, Addict. Biol. 19 (2) (2014) 262–271.

[20] M. Lee, H. Shen, J.Z. Huang, J.S. Marron, Biclustering via sparse singular value decomposition, Biometrics 66 (2010) 1087–1095.

[21] X. Lei, F. Wang, F.-X. Wu, A. Zhang, W. Pedrycz, Protein complex identification through Markov clustering with firefly algorithm on dynamic protein–protein interaction networks, Inf. Sci. 329 (2016) 303–316.

[22] S.-Y. Li, Y. Jiang, Z.-H. Zhou, Partial multi-view clustering, in: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014, pp. 1968–1974.

[23] P. Lobmaier, H. Kornor, N. Kunoe, A. Bjorndal, Sustainedrelease naltrexone for opioid dependence, Cochrane Database Syst. Rev. 33 (2008) CD006140, doi:10.1002/14651858.CD006140.

[24] M. Lu, X.-J. Zhao, L. Zhang, F.-Z. Li, Semi-supervised concept factorization for document clustering, Inf. Sci. 331 (2016) 86–98.

[25] I. Maremmani, P. Pani, A. Mellini, M. Pacini, G. Marini, M. Lovrecic, G. Perugi, M. Shinderman, Alcohol and cocaine use and abuse among opioid addicts engaged in a methadone maintenance treatment program, J. Addict. Dis. 26 (2007) 61–70.

[26] E. Masciari, G.M. Mazzeo, C. Zaniolo, Analysing microarray expression data through effective clustering, Inf. Sci. 262 (2014) 32–45.

[27] T.D. Pigott, A review of methods for missing data, Educ. Res. Eval. 7 (4) (2001) 353–383.

[28] G. Pio, F. Serafino, D. Malerba, M. Ceci, Multi-type clustering and classification from heterogeneous networks, Inf. Sci. 425 (2018) 107–126.

[29] M.M.G. Samira, F.Z. Mohammad Hossein, T. Ismail Burhan, Multi-central general type-2 fuzzy clustering approach for pattern recognitions, Inf. Sci. 328 (2016) 172–188.

[30] W. Shao, L. He, P.S. Yu, Clustering on multiple incomplete datasets via collective kernel learning, in: Proceedings of the IEEE 13th International Conference on Data Mining, 2013, pp. 1181–1186.

[31] W. Shao, L. He, P.S. Yu, Multiple incomplete views clustering via weighted nonnegative matrix factorization with l2,1 regularization, in: Proceedings of the 2015th European Conference on Machine Learning and Knowledge Discovery in Databases, 2015, pp. 318–334.

[32] Y.S. Su, A. Gelman, J. Hill, M. Yajima, Multiple imputation with diagnostics (mi) in r:opening windows into the black box, J. Stat. Softw. 45 (2011).

[33] J. Sun, J. Bi, H.R. Kranzler, Multi-view biclustering for genotype-phenotype association studies of complex diseases, in: Proceedings of IEEE International Conference on Bioinformatics and Biomedicine, 2013, pp. 316–321.

[34] J. Sun, J. Bi, H.R. Kranzler, Multi-view singular value decomposition for disease subtyping and genetic associations, BMC Genet. 15 (73) (2014) 1–12.

[35] J. Sun, J. Lu, T. Xu, J. Bi, Multi-view sparse co-clustering via proximal alternating linearized minimization, in: ICML, 2015, pp. 757–766.

[36] C. TJ., E. MS., H. J., Shifting patterns of prescription opioid and heroin abuse in the United States, N. Engl. J. Med. 373 (2015) 1789–1790.

[37] A. Trivedi, P. Rai, D.I. H., S. DuVall, Muliview clusterting with incomplete views, NIPS 2010: Workshop on Machine Learning for Social Computing, Whistler, Canda, 2010.

[38] A. Wang, I. Elman, S. Lowen, S. Blady, K. Lynch, J. Hyatt, C. O'brien, D. Langleben, Neural correlates of adherence to extended-release naltrexone pharmacotherapy in heroin dependence, Transl. Psychiatry 5 (3) (2015) e531.

[39] A. Wang, I. Elman, S. Lowen, S. Blady, K. Lynch, J. Hyatt, C. O'Brien, D. Langleben, Neural correlates of adherence to extended-release naltrexone pharmacotherapy in heroin dependence, Transl. Psychiatry 5 (2015), doi:10.1038/tp.2015.20.

[40] S. Xiang, L. Yuan, W. Fan, W. Yalin, P.M. Thompson, J. Ye, Bi-level multi-source learning for heterogeneous block-wise missing data, Neuroimage 102 (2014) 192–206.

[41] Z. Xue, J. Du, D. Du, S. Lyu, Deep low-rank subspace ensemble for multi-view clustering, Inf. Sci. 482 (2019) 210–227.

[42] J. Yao, Y. Yao, A granular computing approach to machine learning., FSKD 2 (2002) 732–736.

[43] J.T. Yao, A.V. Vasilakos, W. Pedrycz, Granular computing: perspectives and challenges, IEEE Trans. Cybern. 43 (6) (2013) 1977–1989.

[44] X. Zhang, H. Sun, Z. Liu, Z. Ren, Q. Cui, Y. Li, Robust low-rank kernel multi-view subspace clustering based on the schatten p-norm and correntropy, Inf. Sci. 477 (2019) 430–447.

[45] P. Zhou, Y. Hou, J. Feng, Deep adversarial subspace clustering, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.