A Comparison of Clustering and Missing Data Methods for Health Sciences

Ran Zhao†, Deanna Needell‡, Christopher Johansen¹, Jerry Grenad³
¹Claremont Graduate University ²Claremont McKenna College

Abstract
In this paper, we compare and analyze clustering methods with missing data in behavioral research. In particular, we propose and analyze the use of compressive sensing's matrix completion approach with spectral clustering to cluster health-related data. The empirical tests and real data results show that our methods can outperform standard methods like LPA and FIML in terms of lower misclassification rates in clustering and better matrix completion performance in missing data problems. According to our examination, a possible explanation of these improvements is that spectral clustering takes advantage of high data dimension and compressive sensing methods utilize the near-low-rank property of health data.

INTRODUCTION
A. Clustering Analysis
A vast array of literature has explored clustering techniques and missing data issues in both mathematics and public health research. Clustering refers to the separation of data into meaningful groups so that data within each group is similar.

B. Missing Data
In many large scale applications, data is incomplete. For example, participants may be unable or unwilling to complete an ongoing survey, or participants may be randomly assigned different blocks of questions to increase the variety of constructs assessed.

In public health data, especially the data from surveys or investigations, one expects the data to be low-rank or approximately low-rank for certain variables, because there are a small number of underlying factors that influence specific human opinions and behaviors. In this paper, we empirically investigate the use of matrix completion with spectral clustering to cluster incomplete data, and compare to standard FIML and LPA methods. From these studies, we find that the combination of compressive sensing and spectral clustering methods can offer better performance than standard methods currently used in health data research.

In this paper, we empirically investigate the use of matrix completion with spectral clustering to cluster incomplete data, and compare to standard FIML and LPA methods. From these studies, we find that the combination of compressive sensing and spectral clustering methods can offer better performance than standard methods currently used in health data research.

METHODS
A. Clustering Analysis
The Latest Profile Analysis (LPA) method is a common approach in health behavior research to identify unobserved classes of participants and explain the patterns of responses [11], [13], [15], [17]. Many current software packages use an iterative expectation maximization (EM) algorithm to estimate the parameters [15]. The EM algorithm and other variants have both advantages and drawbacks for estimation of the LPA parameters. The algorithms are sensitive to the initial values of the parameters with the potential for local solutions, and the EM approach does not estimate standard errors. Model identification, the issue of whether there is sufficient information to estimate the parameters [11], and subjective model fit selection are also drawbacks to these approaches.

Spectral clustering (SC) is a geometric method that can identify relationships in data (e.g., cluster individuals each with d variables) that are non-linear (13), [18], [19]. Here, one designs a similarity measure to form a Laplacian matrix from the data. A typical normalized Laplacian matrix L is defined by

\[ L = D^{-1/2} (D-W)^{-1/2} \]

where W is the symmetric weight matrix whose (i, j)th entry corresponds to the similarity between individuals i and j, and the degree matrix D has diagonal entries \( D_{ii} = \sum_j W_{ij} \). Spectral clustering computes the eigenvectors of this Laplacian which form a lower dimensional, linear separable representation of the dataset.

B. Missing Data
In many large scale applications, data is incomplete. For example, participants may be unable or unwilling to complete an ongoing survey, or participants may be randomly assigned different blocks of questions to increase the variety of constructs assessed.

1) FIML
Full Information Maximum Likelihood (FIML) [10], [18] is used to maximize the likelihood of the data by auditioning combinations of parameter estimates [9]. The procedure relies on assumptions such as normality, which when violated can result in biased parameter estimates. There is also a risk of convergence to local maxima resulting in poor parameter estimates.

The FIML estimator implemented in common statistical software packages maximizes a likelihood function that is the sum of n case-wise likelihood functions. A likelihood function is calculated for each observation or individual. The likelihood measures the discrepancy between the current parameter estimates and the observed data for the ith case. The function is maximized assuming multivariate normality:

\[ \log L_i = \frac{1}{2} \log (2\pi) + \frac{1}{2} \log |\Sigma| + \frac{1}{2} (x_i - \mu_i)' \Sigma^{-1} (x_i - \mu_i) \]

The vector of complete data for case i is represented by the term \( x_i \), and the vector of estimated means for those variables that are observed for case \( i \) is the term \( \mu_i \). Constant \( K_i \) depends upon the number of complete data points for case \( i \). Only those variables that are observed for case \( i \) are used to calculate the determinant and inverse of \( \Sigma \). The discrepancy function for the entire sample is calculated by summing over the n case-wise functions:

\[ \log L = \sum_i \log L_i \]

It is assumed that missing values for \( X \) are conditionally dependent on other observed variables in the data. Probability values for the missing data are estimated during the process of optimizing likelihood functions of vectors of partially complete data in the individual-level likelihood functions. The EM algorithm’s multiple regression of \( X \) on other variables to generate predicted scores for the missing data. The FIML estimate does not impute missing values, however, but uses all available raw data to directly estimate parameters and standard errors for the model.

2) Compressive Sensing
Compressive sensing (CS) is a new and fast growing field in applied mathematics. The CS application matrix completion demonstrates that a (near) low-rank matrix can be completed accurately and robustly from observation of only a few of its entries by solving a nuclear norm minimization problem [1], [13], [15]. A typical format of this optimization problem is:

\[ \min_{X} \| X \|_F \quad \text{subject to} \quad X_{ij} = \hat{X}_{ij}, \quad (i,j) \in O \]

where the nuclear norm \( \| X \|_F \) = \( \text{Tr}(X' X) \), M is the matrix we wish to recover, and \( O \) is the set of locations of observed matrix entries in M. This popular convex relax of the rank minimization problem is feasible and commonly used in matrix completion, since minimization of the rank of X is NP-hard due to its combinatorial nature.

EMPIRICAL RESULTS
We first use simulated data to compare the clustering performance of spectral clustering and LPA. First, we generate two-dimensional points whose x and y values follow a normal distribution with mean 0 (for one cluster) or a > 0 (for the other cluster) and variance 1. As a result, we expect clustering to be more successful, because the difference between clusters is more obvious. We define the correct classification rate (CCR) as the ratio of the number of correctly clustered points over the total number of points in each trial. We simulate 40 different data sets for each value of a, and use these 40 trials to compute the rates of each method. In Figure 5, we show the mean CCR for datasets that contain two equally sized clusters, with approximately 250 observations in each cluster. Figure 4, however, illustrates the CCR from datasets that contain unequal sizes of clusters, where one has approximately 25 (5%) observations and the other has approximately 225 (45%) observations. This experiment aims to test how well spectral clustering and LPA classify observations in situations with different clustering complexity and relative size of clusters.

Given this difference, it is hard to compare the performance of FIML and compressive sensing directly, so we instead use their results to cluster.

C. Application to Real Public Health Data
We next compare the above methods on real public health data. The data is obtained from the teen California Health Interview Survey (CHIS) from 2009. CHIS is one of the largest surveys in the nation and is conducted and maintained by the UCLA Center for Health Policy Research and its collaborators. CHIS obtains data via phone interviews on extensive health related items such as health status, health conditions, health-related behaviors, health insurance coverage, access to health care services, and other health and health related issues [2]. The averaged rates are illustrated in Figure 5. Though as expected the correct classification rates decrease monotonically, the CCR for compressive sensing/spectral clustering seems to reduce at a slower rate than that of FIML/LPA. Regardless, these two approaches overall generate quite reliable outcomes, even when the proportion of missing data reaches as large as 50%.

SUMMARY
Using two groups of simulated data, we observe that spectral clustering may be preferable to LPA, and that compressive sensing methods may have an advantage over FIML, in giving the recovered data matrix explicitly, taking advantage of nearly low-rank data. The contribution of this paper is to bring two methods from applied mathematics into health behavior research, and verify their advantages over traditionally used methods. Our future research direction is to further compare the performance of these methods on real health (and other types of) data, and aim to identify in what settings each type of method is preferred. This identification can aid in the design of health data surveys allowing for intentional missing data, thereby reducing participant burden and cost.

REFERENCES
Support for this research was provided by a BLAIRS Challenge Award (211-170045) received from Claremont Graduate University, Claremont, CA.

ACKNOWLEDGEMENT