4-22-2014

A Comparison of Clustering and Missing Data Methods for Health Sciences

Ran Zhao
Claremont Graduate University

Deanna Needell
Claremont McKenna College

Christopher Johansen
Claremont Graduate University

Jerry L. Grenard
Claremont Graduate University

A Comparison of Clustering and Missing Data Methods for Health Sciences

R. Zhao
Institute of Mathematical Sciences
Claremont Graduate University
Claremont, CA, 91711
Email: ran.zhao@cgu.edu

D. Needell
Department of Mathematics
and Computer Science
Claremont McKenna College
Claremont, CA, 91711

C. Johansen and J. L. Grenard
School of Community
and Global Health
Claremont Graduate University
Claremont, CA, 91711

Abstract—In this paper, we compare and analyze clustering methods with missing data in health behavior research. In particular, we propose and analyze the use of compressive sensing's matrix completion along with spectral clustering to cluster health related data. The empirical tests and real data results show that these 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-to-low-rank property of health data.

I. 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.

The Latent Profile Analysis (LPA) method is a common approach in health behavior research to identify unobserved classes of participants and explain the pattern of responses [11], [14], [5], [12]. 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 the data (here we consider n 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 $\mathbf{L} \in \mathbb{R}^{n \times n}$ is defined by

$$L = D^{-1/2}(D - W)D^{-1/2}, \qquad (1)$$

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 $D_i i = \sum_j W_{ij}$. Spectral clustering computes the eigenvectors of this Laplacian which form a lower dimensional, linear separable representation of the dataset [17].

In the dataset from Section II-A, the unsorted and sorted eigenvector (from the second largest eigenvalue) entries are shown in Figure 1. Since each entry of the eigenvector corresponds to an individual, we may use the values of these entries to separate the individuals into clusters. In this case, the threshold to designate different clusters seems to be $y = 0$, as plotted in red. For more than two clusters, one can instead run k-means [13], [19] on this data to identify the clusters.

In Figure 2, we show the comparison of the spectral clustering results and that of the actual (randomly generated) clusters on right.

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.
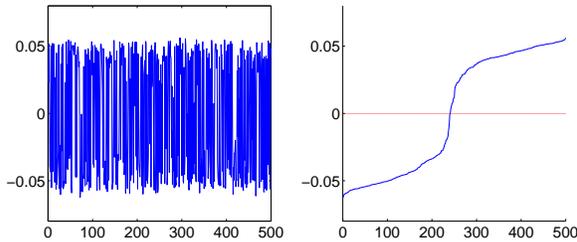
Fig. 1 Left: unsorted eigenvector. Right: Sorted eigenvector. The red horizontal line indicates the separation.
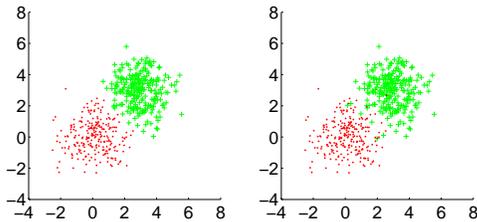

Fig. 2 Results from spectral clustering (on the left) actual real clusters (on the right). Red and green show different clusters.

1) FIML: Full Information Maximum Likelihood (FIML) [10], [8] aims 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 parameters. 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 function measures the discrepancy between the current parameter estimates and the observed data for the ith case. The function is maximized assuming multivariate normality:

$$\log L_i = K_i - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(x_i - \mu_i)'\Sigma_i^{-1}(x_i - \mu_i)$$

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 μ_i. A 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 Σ_i. The discrepancy function for the entire sample is calculated by summing over the n case-wise functions:

$$\log L(\mu, \Sigma) = \sum_{i=1}^{N}\log L_i.$$

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 implied during the parameter estimation process by incorporating vectors of partially complete data in the individual-level likelihood functions. This is analogous to using 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 (nearly) 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], [3], [6]. A typical format of this optimization problem is

$$\begin{aligned} \text{minimize} \quad & \|X\|_* \\ \text{subject to} \quad & X_{ij} = M_{ij} \quad (i,j) \in \Omega \end{aligned}$$

where the nuclear norm $\|X\|_* = \sum_{k=1}^{n}\sigma_k(X)$, M is the matrix we wish to recover, and Ω is the set of locations of observed matrix entries in M. This popular convex relaxation 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.

When the underlying matrix is low-rank, matrix completion completes the data matrix provably well from a small number of possibly noisy observations [4], [7], [16]. To quantify how exact the method recovers the matrix, we generate two 1000×300 matrices with rank 2 and 10 and remove 20%, 40%, 60%, and 80% of the data purposefully. The entries are random values that follow a standard normal distribution. The matrix completion results are presented in Table I and Table II. We measure the recovery error between the actual matrix X and the recovered matrix $\hat{X}$ by the Frobenius norm $\|X - \hat{X}\|_F$, the relative Frobenius norm $\|X - \hat{X}\|_F / \|X\|_F$, and the spectral norm $\|X - \hat{X}\|$. As

TABLE I Rank 2 Matrix Completion Results

Rank 2	Frobenius	Relative Frob.	Spectral
missing 20%	0.0687	8.84E-05	0.0505
missing 40%	0.0376	4.85E-05	0.0246
missing 60%	0.0651	8.38E-05	0.0422
missing 80%	0.0959	1.23E-04	0.0645

TABLE II Rank 10 Matrix Completion Results

Rank 10	Frobenius	Relative Frob.	Spectral
missing 20%	0.0896	5.37E-05	0.0297
missing 40%	0.145	8.69E-05	0.0512
missing 60%	0.186	1.12E-04	0.0643
missing 80%	0.350	2.10E-04	0.1890

is evident and not surprising, the error increases slightly with more missing data, and the higher rank matrix has slightly higher recovery error.

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.

II. EMPIRICAL RESULTS

A. Experiments on Clustering Analysis

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 increases, 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 3, 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 475 (95%) 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.
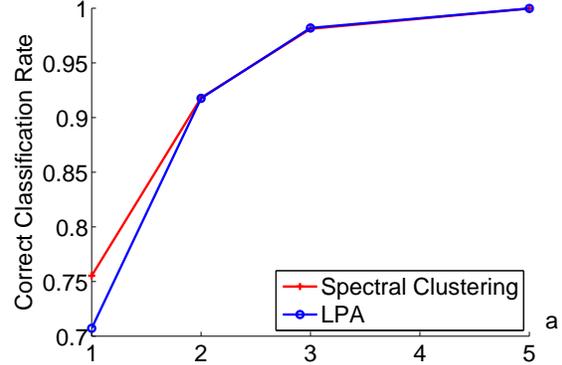

Fig. 3 Clustering results of spectral clustering and LPA methods for equally sized clusters.
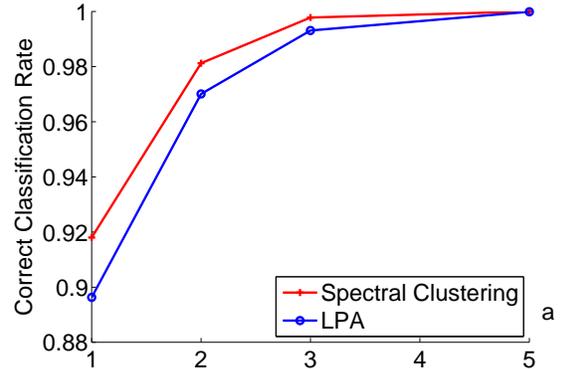

Fig. 4 Clustering results of spectral clustering and LPA methods for unequally sized clusters.

We observe that both methods have increasing CCR for larger a, and that spectral clustering has a higher correction rate when the distance between centroids is small for equally sized clusters and for unequally sized clusters. Overall, spectral clustering seems to offer improvements over LPA in this setting. A detailed summary of results for equal cluster sizes is shown in Table III. When $a \geq 2$, where the (spatial) distance between centroids is greater than 2.828, the CCR of both methods exceed 90%, but in most cases spectral clustering has lower standard deviation of estimation.

3

TABLE III Correct Classification Rate Results

In this table, a represents the centroid (a, a) other than $(0, 0)$. CCR is the abbreviation of correct classification rate, and N represents the number of observations. In this test we have sample size of 500 and approximately 250 obs. for each cluster. The rest of items are summary statistics of correct classification rate estimated from 40 trials.

a	CCR	N	Mean	S.D.	Min	Med.	Max
1	SC	500	0.755	0.0198	0.692	0.756	0.812
	LPA	500	0.707	0.0550	0.508	0.072	0.796
2	SC	500	0.918	0.0132	0.868	0.918	0.958
	LPA	500	0.918	0.0130	0.882	0.916	0.962
3	SC	500	0.981	0.0065	0.958	0.982	0.998
	LPA	500	0.982	0.0060	0.960	0.982	0.998
5	SC	500	1.00	0.0008	0.994	1.00	1
	LPA	500	1.00	0.0010	0.996	1.00	1.00

B. Experiments on Missing Data

Missing data is an important part of real-world health data analysis. There is an obvious difference in how FIML and compressive sensing handle this problem. FIML does not recover missing data or interpolate unknown values to incomplete data entries, while matrix completion does precisely this. Given this difference, it is hard to compare the performance of FIML and compressive sensing directly, so we instead use their results to cluster. We compare the correction rates of the following three methods, a) FIML combined with LPA, b) matrix completion followed by LPA, and c) matrix completion followed by spectral clustering.

By comparing methods a) and b), we can compare the relative performance of FIML and matrix completion, because the clustering methods are the same (LPA). Therefore the correct classification rates will reflect how well these two methods handle missing data. The results of b) and c) will strengthen the conclusion from part a), where the two different clustering methods are compared.

We generate 40 data sets of 1000 (the number of individuals) by 100 (the number of variables) matrices. One can imagine the 500×100 top half of the matrix corresponding to one cluster, and the bottom half to the other. The bottom half has standard normally distributed entries, whereas the top half has normally distributed entries with variance 1 but with varying means; the first ten columns have mean 0.1, the next 10 have mean 0.2, and so on, so that the last ten have mean 1.0 (we do this to introduce more variety within the cluster). We remove entries from the matrix uniformly at random to create missing data.

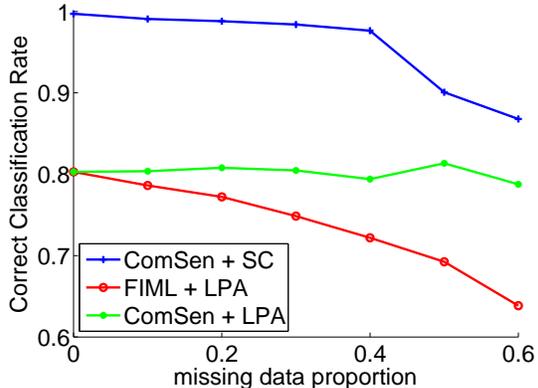In Figure 5, we observe that the correct clas-

Fig. 5 Mean correct classification rate as a function of the proportion of missing data for method a), method b), and method c), (red, green, and blue, respectively). The description of the methods is elaborated in II-B.

sification rate is higher for method b) than for method a). This indicates that matrix completion may handle missing data better than FIML in this setting. Additionally, method c) is better than both, suggesting that matrix completion coupled with spectral clustering may offer even better performance. There is of course a computational tradeoff, but for methods where accuracy is the priority, matrix completion coupled with spectral clustering may offer improved performance over standard methods.

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].

One major difficulty of analyzing the clustering techniques on real data is that there is not an obvious ground truth to which to compare. To overcome this, we first eliminate irrelevant variables such as individual's serial number and zip code. This yields a data matrix with 3379 individuals and 144 variables. Then we apply both spectral cluster-

ing and LPA, and identify those individuals who were clustered in the same way by both methods. This left 2836 individuals as the "consistent population", which is 83.93% of the original data. Next, we sample 1000 individuals (without replacement) among this consistent population, and repeat this process 40 times. In each of these 40 trials, we randomly remove 10%, 30%, and 50% of the entries to mimic missing data. Finally, we apply matrix completion/spectral clustering and FIML/LPA and compute the mean CCR using the consistent clusters as ground truth for each approach.

The averaged rates are illustrated in Figure 6. Though as expected the correct classification rates decrease monotonically, the CCR for compressive sensing/spectral clustering seems to decay at a much slower speed 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%.
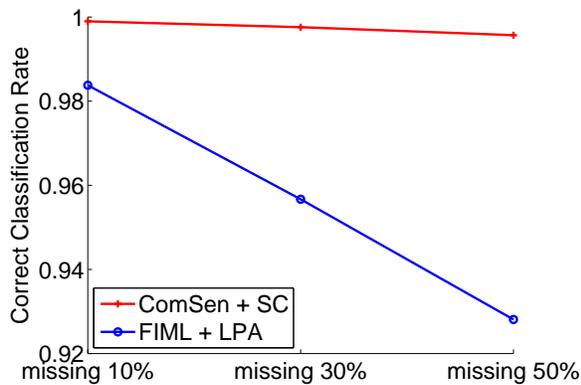

Fig. 6 Missing data completion and clustering analysis on CHIS data.

III. 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

[1] Compressed sensing webpage. http://dsp.rice.edu/cs.
[2] California health interview survey. chis 2009 adolescent public use file. release 1. Technical report, UCLA Center for Health Policy Research, Los Angeles, CA, 2009.
[3] E. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.
[4] E. J. Candes and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.
[5] S.J.H. Biddle W.C. Liu C.K.J. Wang and B.S.C. Lim. A latent profile analysis of sedentary and physical activity paterns. *Journal of Public Health*, 20:367–373, 2012.
[6] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. Published online.
[7] Y. C. Eldar, D. Needell, and Y. Plan. Uniqueness conditions for low-rank matrix recovery. *Applied Computational Harmonic Analysis*, 31(1):59–73, 2011.
[8] C. K. Enders. *Applied missing data analysis*. New York: Guilford Press, 2010.
[9] C.K. Enders. The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement*, 61(5):713–740, 2001.
[10] J. W. Graham. *Missing data analysis and design*. New York: Springer, 2012.
[11] J. A. Hagenaars and A. L. McCutcheon. *Applied latent class analysis*. Cambridge University Press, New York, 2002.
[12] U. Tautwein H.W. Marsh, L. Ludtke and A.J.S. Morin. Classical latent profile analysis of academic self-concept dimensions: Synergy of person- and variable-centered approaches to theoretical models of self-concept. *Structural Equation Modeling*, 16(2):191–225, 2009.
[13] Stuart Lloyd. Least squares quantization in pcm. *Information Theory, IEEE Transactions on*, 28(2):129–137, 1982.
[14] G.H. Lubke and B. Muthen. Investigating population heterogeneity with factor mixture models. *Psychological Methods*, 10(1):21–39, 2005.
[15] L.K. Muthen and B.O. Muthen. *Mplus User's Guide*. Los Angeles, CA: Muthen & Muthen, 1998-2012.
[16] B. Recht, M. Fazel, and P.A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.
[17] J. Shi and J.Malik. Normalized cuts and image segmentation. *Pattern Analysis andMachine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.
[18] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
[19] Xindong Wu and Vipin Kumar. *The top ten algorithms in data mining*. CRC Press, 2009.