CONVEX CLUSTERING AND RECOVERY OF PARTIALLY OBSERVED DATA

Sunrita Poddar, Mathews Jacob

Department of Electrical and Computer Engineering
The University of Iowa, IA, USA

ABSTRACT

We propose a convex clustering and reconstruction algorithm for data with missing entries. The algorithm uses a similarity measure between every pair of points to cluster and recover the data. The cluster centres can be recovered reliably when the ground-truth similarity matrix is available. Moreover, the similarity matrix can also be reliably estimated from the partially observed data, when the clusters are well-separated and the coherence of the difference between points from different clusters is low. The algorithm performs well using the estimated similarity matrix on a simulated dataset. The method is also successful in reconstructing images from under-sampled Fourier data.

Index Terms— Clustering, Missing Data, Image Reconstruction, Matrix Completion.

1. INTRODUCTION

Clustering aims at finding groups within a collection of objects, based on similarity in their features. It is an important and well-studied problem in data analysis, as is evident from the vast amount of literature dedicated to it. Some traditional clustering techniques are K-means [1] and spectral clustering [2]. Recently, convex clustering methods [3] have been proposed which address many of the shortcomings of these traditional methods, such as sensitivity to initialization and prior knowledge of the number of clusters.

Most clustering algorithms assume full knowledge of all the features of each object. Relatively less work has been done on clustering data when incomplete feature information is available about some (or all) the objects. Such a problem might arise in the clustering of survey respondents who choose not to answer certain questions [4]. In Magnetic Resonance Imaging, data corresponding to different image frames is collected in the Fourier domain. Since the imaging process is slow, only a few Fourier samples can be collected. Similar images appearing at different time points may be clustered to aid image reconstruction [5].

The existing clustering algorithms can be directly applied to incompletely observed data through the methods of "deletion" or "imputation" [6, 7]. "Deletion" is the removal of the objects with missing features from the analysis. However, this might involve discarding a large amount of the collected data. "Imputation" is the estimation of the missing values prior to clustering. The clustering result then becomes very dependent on the accuracy of imputation. Further, the imputed values and measured values are treated equivalently. Recently, an alternative algorithm termed "k-POD" [8] has been proposed, which alternates between imputation and k-means clustering. However, it is prone to all the disadvantages of k-means.

We propose a convex clustering algorithm for data with missing entries. The algorithm estimates the cluster centres using the matrix of similarity measures between every pair of points. It is shown that using the ground-truth similarity matrix in the proposed algorithm results in good cluster centre estimates. However, the ground-truth similarity matrix is not available in practice due to missing entries in the data. Thus, we estimate the similarity matrix from the incomplete data itself. The similarity between a pair of points can be estimated using their partial distance, computed using the samples at commonly observed locations. It is assumed that the clusters are well-separated and the difference between pairs of points from different clusters have low coherence. Under these assumptions, the estimated similarity is shown to be accurate, if there are a sufficient number of commonly observed locations. It is observed that using the estimated similarity matrix does not result in very significant errors, and the performance improves with increase in the number of observed entries in the data. The technique is also used to reconstruct an image series from under-sampled Fourier measurements.

2. BACKGROUND

2.1. Notations

We consider the matrix $\mathbf{X} \in \mathbb{R}^{n\times N}$, where each column is an observation ($\mathbf{x}_i \in \mathbb{R}^n$) with n features. Each observation belongs to 1 out of k clusters. The i^{th} cluster contains N_i points, and therefore $\sum_{i=1}^{k} N_i = N$. The cluster to which $\mathbf{x}_j$ belongs is denoted by $C(j)$. Let $\mathbf{U} \in \mathbb{R}^{n\times N}$ be the matrix of cluster centres, such that the i^{th} column $\mathbf{u}_i \in \mathbb{R}^n$ represents the centre of cluster $C(i)$. $\mathbf{x}_i$ and $\mathbf{u}_i$ are related as:

$$\mathbf{x}_i = \mathbf{u}_i + \eta_i \tag{1}$$

Each element of the matrix $\mathbf{X}$ is sampled with probability p_0. The rectangular sampling matrix corresponding to $\mathbf{x}_i$ is denoted by $\mathbf{S}_i$. Our goal is to recover the cluster centres $\{\mathbf{u}_i\}$, given incomplete observations $\{\mathbf{S}_i\mathbf{x}_i\}$.

2.2. Convex Clustering of fully sampled data

Convex clustering methods have been proposed for the case of fully sampled data (i.e. $\mathbf{S}_i = \mathbf{I}, \forall i$) by solving the optimization problem:

$$\{\mathbf{u}_i^*\} = \arg\min_{\mathbf{u}_i} \sum_i \|\mathbf{u}_i - \mathbf{x}_i\|^2 + \lambda \sum_i \sum_j w_{ij} \|\mathbf{u}_i - \mathbf{u}_j\|^2 \tag{2}$$

Here, the weight w_{ij} represents the similarity between points $\mathbf{x}_i$ and $\mathbf{x}_j$, computed using a non-linear function such as:

$$w_{ij} = e^{-\frac{d_{ij}^2}{\sigma^2}} \tag{3}$$

where $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$. Unlike spectral clustering and k-means clustering, convex clustering methods do not require the prior knowledge of the desired number of clusters and are insensitive to initialization. The number of clusters change continuously with the regularization parameter λ. This allows for an observation of the "clustering path", by varying λ over a large range.

3. THEORY

3.1. Convex Clustering of data with missing entries

We propose to extend the convex clustering algorithm to account for missing data by solving the optimization problem:

$$\begin{aligned}\{\mathbf{u}_i^*\} = \arg\min_{\mathbf{u}_i} &\sum_i \|\mathbf{S}_i(\mathbf{u}_i - \mathbf{x}_i)\|^2 \\ &+\lambda \sum_i \sum_j w_{ij} \|\mathbf{u}_i - \mathbf{u}_j\|^2\end{aligned} \tag{4}$$

The weights w_{ij} cannot be estimated in this case using (3), since only a few entries of $\mathbf{x}_i$ are known, and the rest are missing. We first obtain the solution $\mathbf{u}^{*(gt)}$ assuming perfect knowledge of the ground-truth weight matrix $\mathbf{W}^{(gt)}$. We compare this to the solution $\mathbf{u}^{*(est)}$ using weights estimated from the partially observed data, denoted by $\mathbf{W}^{(est)}$. We note that the solution for each row of $\mathbf{U}^*$ is independent of other rows. For simplicity, we analyze the solution for only the 1^{st} row of $\mathbf{U}^*$. The 1^{st} rows of $\mathbf{X}$ and $\mathbf{U}$ are denoted by $\mathbf{x}$ and $\mathbf{u}$ respectively. x_j and u_j are the j^{th} elements of $\mathbf{x}$ and $\mathbf{u}$ respectively. We also have the following assumptions on the data:

A1 The maximum spread of any cluster is: $\delta = \max_i \|\eta_i\|$.

A2 The minimum distance between any 2 cluster centres is: $\epsilon = \min_{C(i)\neq C(j)} \|\mathbf{u}_i - \mathbf{u}_j\|$.

A3 There exists a constant $K > 2$ such that $\epsilon = K\delta$. A large value of K implies well-separated clusters.

A4 For any vector $\mathbf{x} \in \mathbb{R}^n$, the coherence is defined as $\frac{n\|\mathbf{x}\|_\infty^2}{\|\mathbf{x}\|_2^2}$. The coherence of the difference between each pair of columns from different clusters is upper bounded by μ. The intuition is to avoid situations where points belonging to different clusters differ only at a few sampling locations.

We derive an expression for $\mathbf{u}^{*(gt)}$. We observe that $\mathbf{u}^{*(gt)}$ is very close to the ground-truth cluster centres $\mathbf{u}$ and the difference reduces with increase in sampling probability p_0. We also derive an estimate for $\Delta\mathbf{u}^* = (\mathbf{u}^{*(gt)} - \mathbf{u}^{*(est)})$. Our experiments result in small values of $\Delta\mathbf{u}^*$ when K is large and μ is small. We observe that for relatively higher p_0, the estimate for $\Delta\mathbf{u}^*$ is quite accurate.

3.2. Performance using Ground-truth Weights

We define the ground-truth weight matrix $\mathbf{W}^{(gt)}$ entries as:

$$w_{ij}^{(gt)} = \begin{cases} 1 & \text{, if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same cluster.} \\ 0 & \text{, otherwise.} \end{cases} \tag{5}$$

We find an expression for the solution of (4) using $\mathbf{W}^{(gt)}$. The following definitions will be used to state the result:

- $s_j = 1$ if x_j has been observed and 0 otherwise.
- $\rho_{C(j)}$ is the fraction of entries belonging to cluster $C(j)$ that have been observed in $\mathbf{x}$.

Theorem 1. *The solution of the clustering algorithm* (4) *using the weight matrix* $\mathbf{W}^{(gt)}$ *is:*

$$u_j^{*(gt)} = \frac{1}{s_j + \lambda N_{C(j)}}[s_j x_j + \frac{\lambda}{\rho_{C(j)}} \sum_{m:\mathbf{x}_m \in C(j)} s_m x_m] \tag{6}$$

Corollary 1.1. *When there is zero intra-cluster variance (i.e.* $K \to \infty$*), then with a probability* $\geq 1 - (1-p_0)^{N_{C(j)}}$*, the solution to* (4) *using the weight matrix* $\mathbf{W}^{(gt)}$ *is:*

$$u_j^{*(gt)} = x_j = u_j \tag{7}$$

The probability in Corollary 1.1 is associated with the assumption that there is at least 1 known entry in $\mathbf{x}$ belonging to the cluster $C(j)$. If this assumption is satisfied, then we have perfect recovery of the centre of cluster $C(j)$.

3.3. Weight Estimation from Incomplete Data

Since $\mathbf{W}^{(gt)}$ is not available in practice, we estimate the weight matrix $\mathbf{W}^{(est)}$ from the partially observed data. Similar to [9], we will use the concept of partial distances. We denote the set of indices that are observed in both $\mathbf{x}_i$ and $\mathbf{x}_j$

by Ω_{ij}. We represent the vector of entries of $\mathbf{x}_i$ at locations in the set Ω by $\mathbf{x}_i^{\Omega}$. Let $|\Omega_{ij}| = q$. The partial distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as:

$$d_{ij}^{\Omega_{ij}} = \sqrt{\frac{n}{q}} \|\mathbf{x}_i^{\Omega_{ij}} - \mathbf{x}_j^{\Omega_{ij}}\|_2 \tag{8}$$

Using ideas from [9], we can conclude that if a pair of points has a sufficiently large number of commonly observed locations, then the partial distance between them is close to the actual distance between them with a high probability. The idea is formalized in the next theorem.

Theorem 2. *For any $0 < \delta_0, \delta_1 < 1$ and $q \geq q_0 = \frac{2\mu^2}{\delta_1^2} \log \frac{2}{\delta_0}$, we have with probability $\geq (1 - \delta_0)$:*

$$(1 - \delta_1) d_{ij}^2 \leq (d_{ij}^{\Omega_{ij}})^2 \leq (1 + \delta_1) d_{ij}^2 \tag{9}$$

Thus, for pairs of points having a sufficient number of commonly observed locations, the weight w_{ij} can be estimated reliably from the partial distances with high probability. Motivated by (3), we compute the weight matrix $\mathbf{W}^{(est)}$ from partial distances as:

$$w_{ij}^{(est)} = \begin{cases} e^{-\frac{(d_{ij}^{\Omega_{ij}})^2}{\sigma^2}} & \text{, if } |\Omega_{ij}| \geq q_0 \text{ and } (d_{ij}^{\Omega_{ij}})^2 < t. \\ 0 & \text{, otherwise.} \end{cases} \tag{10}$$

Corollary 2.1. *The weight $w_{ij}^{(est)}$ is computed for a pair of points $\mathbf{x}_i$ and $\mathbf{x}_j$, where $|\Omega_{ij}| \geq q_0$ and $t = (1 + \delta_1)\delta^2$.*

- *If $C(i) \neq C(j)$, then $w_{ij}^{(est)} = 0$ with probability $\geq (1 - \frac{\delta_0 e^{-(K-2)^4}}{2})$.*
- *If $C(i) = C(j)$, then $w_{ij}^{(est)} \geq e^{-\frac{(1+\delta_1)\delta^2}{\sigma^2}}$ with probability $\geq (1 - \frac{\delta_0}{2})$.*

3.4. Performance using Estimated Weights

We perform some preliminary analysis on the performance of algorithm (4) using an estimated weight matrix. A more thorough analysis will be the subject of future work. For simplicity, we study the performance of the clustering algorithm (4) when we have the following weight matrix:

$$w_{ij}^{(th)} = \begin{cases} 1 & \text{, if } C(i) = C(j). \\ e^{-\frac{(d_{ij}^{\Omega_{ij}})^2}{\sigma^2}} & \text{, if } C(i) \neq C(j), |\Omega_{ij}| \geq q_0, \\ & (d_{ij}^{\Omega_{ij}})^2 < t. \\ 0 & \text{, otherwise.} \end{cases} \tag{11}$$

The matrices $\mathbf{W}^{(est)}$ and $\mathbf{W}^{(th)}$ differ only in the definition of the intra-cluster weights. We note from our simulations that under favourable conditions such as high sampling probability and low intra-cluster variance, the effect of $\mathbf{W}^{(est)}$ and $\mathbf{W}^{(th)}$ on the clustering algorithm are comparable. The difference in the solutions $\mathbf{u}^{*(th)}$ and $\mathbf{u}^{*(gt)}$ (using $\mathbf{W}^{(th)}$ and $\mathbf{W}^{(gt)}$ respectively) is due to the presence of non-zero inter-cluster weights.

We next analyze the effect of the inter-cluster weights. We introduce the term "1^{st} order interaction" to refer to the effect of w_{ij} (where $C(i) \neq C(j)$) on u_m^* where $m \in C(i)$, or $m \in C(j)$. Higher order interactions refer to the effect of w_{ij} (where $C(i) \neq C(j)$) on u_m^* where $m \notin C(i), C(j)$. We denote the difference $(u_j^{*(gt)} - u_j^{*(th)})$ by Δu_j^*, and approximate it as the sum of all 1^{st} order interactions. Before stating the next result, which gives a closed-form approximation for Δu_j^*, we define the following:

- $\mathbf{L}^{(gt)}$ and $\mathbf{L}^{(th)}$ are the Laplacian matrices corresponding to $\mathbf{W}^{(gt)}$ and $\mathbf{W}^{(th)}$ respectively.
- $\mathbf{S}$ is the square diagonal sampling matrix for $\mathbf{x}$.

Theorem 3. *The difference between $u_j^{*(th)}$ and $u_j^{*(gt)}$, approximated as the sum of all 1^{st} order interaction errors is:*

$$\Delta u_j^* \approx \lambda [(\mathbf{S} + \lambda \mathbf{L}^{(gt)})^{-1} (\mathbf{L}^{(gt)} - \mathbf{L}^{(th)}) \mathbf{u}^{*(gt)}]_j \tag{12}$$

This approximation ignores all higher order inter-class interactions. We expect that under favourable conditions such as well-estimated weights, well-separated clusters and low intra-cluster variance, this approximation should be accurate.

Corollary 3.1. *If we assume zero intra-cluster variance (i.e. $K \to \infty$) and set $t = 0$, then $\mathbf{W}^{(th)} = \mathbf{W}^{(gt)}$ with probability ≈ 1. In this case, using either of the weight matrices $\mathbf{W}^{(est)}$ or $\mathbf{W}^{(th)}$ in the algorithm (4) results in $\Delta u_j^* = 0$ with probability $\geq (1 - (1 - p_0)^{N_{C(j)}})$.*

4. RESULTS

4.1. Validation on simulated data

The proposed algorithm was tested on a simulated matrix $\mathbf{X} \in \mathbb{R}^{20 \times 500}$ with columns lying in $k = 5$ convex clusters with $N_i = 100$. Columns $1 - 100$ of $\mathbf{X}$ were assigned to cluster 1, $101 - 200$ were assigned to cluster 2 and so on. Fig 1 compares the results $\mathbf{u}^{*(gt)}$ and $\mathbf{u}^{*(est)}$ to the actual cluster-centres $\mathbf{u}$, for $\lambda = 10^{-6}$, $p_0 = 0.8, 0.5, 0.3$ and $K = \infty, 10, 3.5$. As expected, higher K and p_0 result in superior performance. The error $\Delta \mathbf{u}^* = \mathbf{u}^{*(gt)} - \mathbf{u}^{*(est)}$ is shown in Fig 2 for the same parameters. This is compared to the 1^{st} order error approximation given by (12). The approximation is accurate for higher K and p_0.

4.2. Application to MR image reconstruction

Image reconstruction from a few Fourier samples is a common problem in Magnetic Resonance Imaging. We used the

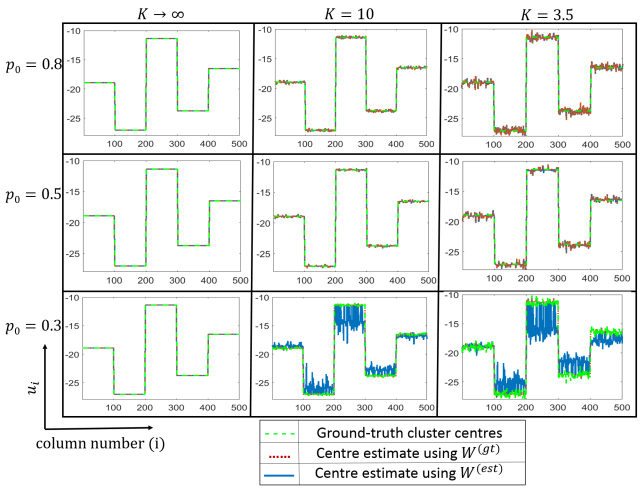

Fig. 1. Clustering performance: The 1^{st} row of the matrix of cluster centres is shown here, obtained from (1) Ground-truth data (green) (2) Proposed algorithm using $\mathbf{W}^{(gt)}$ (red) (3) Proposed algorithm using $\mathbf{W}^{(est)}$ (blue). The results are shown for 3 values of p_0 and K, and λ was fixed at 10^{-6}.

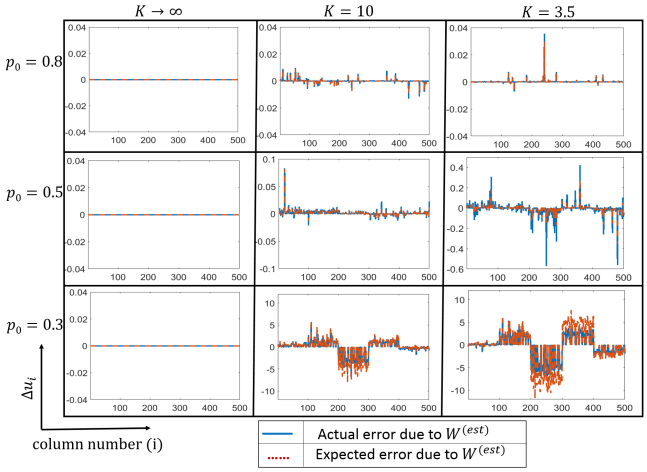

Fig. 2. Clustering error due to imperfect weights: We show the difference between the 1^{st} row of estimated cluster centres when computed using $\mathbf{W}^{(gt)}$ and $\mathbf{W}^{(est)}$. The experimentally obtained error is shown in blue. The theoretically obtained 1^{st} order error approximation is shown in red. The results are shown for the same parameters as in Fig 1.

proposed algorithm to reconstruct a time series of cardiac PINCAT [10] images, from under-sampled Fourier data. The images were generated in the breath-held, short-axis mode. There are $N = 200$ image frames, each of size 128×128 and $k = 20$ cardiac cycles, each consisting of $N_i = 10$ frames. The Fourier domain data corresponding to each image frame can be reshaped into a vector of size $128^2 \times 1$ and arranged as a column of the matrix $\mathbf{X} \in \mathbb{C}^{128^2 \times 200}$. In the original

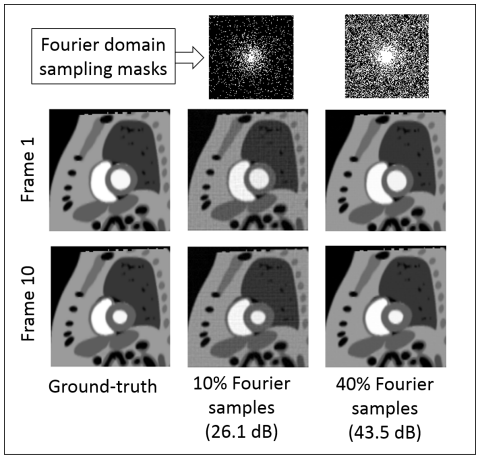

Fig. 3. Image reconstruction from under-sampled Fourier data: Cardiac PINCAT phantom images were under-sampled in the Fourier domain. The images were clustered and reconstructed from 10% and 40% of the Fourier samples, using the proposed algorithm. The under-sampling masks and reconstructed images are shown here along with the ground-truth.

dataset, we have $K \to \infty$. We add zero-mean Gaussian random noise in the image domain, resulting in $K = 3.25$. The Fourier data corresponding to the noisy images was under-sampled using a variable density random sampling mask, as shown in Fig 3. Reconstructions are shown from 10% and 40% of the Fourier samples. It can be seen that the reconstructed images are very similar to the ground-truth.

5. CONCLUSION

A method for convex clustering and reconstruction of data with missing entries is proposed. The algorithm uses a weight matrix which depends on the similarity between every pair of points. An expression is derived for the solution using a ground-truth weight matrix. It is then shown that the weight matrix can also be estimated from the data itself. An estimate is obtained for the difference in the solutions using the two weight matrices. The results obtained at 20% and 50% missing samples on a simulated dataset are quite promising. If the ratio between minimum inter-cluster distance and maximum intra-cluster distance is kept high, then good results are also obtained at 70% missing entries. The algorithm is shown to be successful in clustering and reconstructing cardiac images from highly under-sampled (60% and 90% missing) Fourier data. Cluster size and number of clusters are other factors whose effects are to be studied in future work. Further research needs to be performed to explore the utility of the algorithm in various other image processing and reconstruction applications.

6. REFERENCES

[1] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Applied statistics*, pp. 100–108, 1979.

[2] A. Y. Ng, M. I. Jordan, Y. Weiss *et al.*, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.

[3] F. Lindsten, H. Ohlsson, and L. Ljung, "Just relax and come clustering!: A convexification of k-means clustering," 2011.

[4] J. M. Brick and G. Kalton, "Handling missing data in survey research," *Statistical methods in medical research*, vol. 5, no. 3, pp. 215–238, 1996.

[5] L. Feng, L. Axel, H. Chandarana, K. T. Block, D. K. Sodickson, and R. Otazo, "XD-GRASP: Golden-angle radial MRI with reconstruction of extra motion-state dimensions using compressed sensing," *Magnetic Resonance in Medicine*, 2015.

[6] K. L. Wagstaff and V. G. Laidler, "Making the most of missing values: Object clustering with partial data in astronomy," in *Astronomical Data Analysis Software and Systems XIV*, vol. 347, 2005, p. 172.

[7] J. K. Dixon, "Pattern recognition with partly missing data," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 9, no. 10, pp. 617–621, 1979.

[8] J. T. Chi, E. C. Chi, and R. G. Baraniuk, "k-pod: A method for clustering partially observed data," *arXiv preprint arXiv:1411.7013*, 2014.

[9] B. Eriksson, L. Balzano, and R. Nowak, "High-rank matrix completion and subspace clustering with missing data," *arXiv preprint arXiv:1112.5629*, 2011.

[10] B. Sharif and Y. Bresler, "Physiologically improved NCAT phantom (PINCAT) enables in-silico study of the effects of beat-to-beat variability on cardiac MR," in *Proceedings of the Annual Meeting of ISMRM, Berlin*, vol. 3418, 2007.