CLUSTERING OF DATA WITH MISSING ENTRIES

Sunrita Poddar, Mathews Jacob

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

ABSTRACT

The analysis of large datasets is often complicated by the presence of missing entries, mainly because most of the current machine learning algorithms are designed to work with full data. The main focus of this work is to introduce a clustering algorithm, that will provide good clustering even in the presence of missing data. The proposed technique solves an ℓ_0 fusion penalty based optimization problem to recover the clusters. We theoretically analyze the conditions needed for the successful recovery of the clusters. We also propose an algorithm to solve a relaxation of this problem using saturating non-convex fusion penalties. The method is demonstrated on simulated and real datasets, and is observed to perform well in the presence of large fractions of missing entries.

Index Terms— clustering, missing entries, non-convex penalties

1. INTRODUCTION

Clustering is a popular unsupervised data analysis technique for finding natural groupings in the absence of training data. Specifically, it assigns each data point to a group, such that all points within a group are similar and points in different groups are dissimilar in some sense. Clustering methods are widely used in the analysis of gene expression data, image segmentation, identification of lexemes in handwritten text, search result grouping and recommender systems [1, 2].

Most clustering algorithms cannot be directly applied to datasets with missing entries. For example, gene expression data often contains missing entries due to image corruption, fabrication errors or contaminants [3], rendering gene cluster analysis difficult. Likewise, large databases used by recommender systems (e.g Netflix) usually have a huge amount of missing data, which makes pattern discovery challenging [4]. Similar issues are reported in the context of missing responses in surveys [5] and failing imaging sensors in astronomy [6] are reported to make the analysis in these applications challenging. The most obvious way to apply existing clustering algorithms to data with missing entries is to convert the data to a complete one. This can be done using deletion or imputation [7]. An extension of the weighted sum-of-norms algorithm [8] has been proposed where the weights are estimated from the data points by using some imputation techniques on the missing entries [9]. A majorize minimize algorithm was introduced to solve for the cluster-centres and cluster memberships in [10], which offers proven reduction in cost with iteration. However, these is no theoretical analysis of these algorithms, which makes it difficult to determine what fraction of entries need to be sampled to recover the correct clusters.

In this paper, we introduce an algorithm to cluster data when some of the features are missing in each point. The method is inspired by the recently proposed sum-of-norms clustering technique [8]. This technique assigns a surrogate variable to each data point, which is an estimate of the cluster centre to which that point belongs. When a fusion penalty is used, it is observed that the surrogate variables belonging to the same cluster coalesce to that centre point. These values denote the estimated cluster centres. Guarantees for correct clustering using this technique are available for the case without missing entries [11]. In prior work, we used a weighted convex fusion penalty to recover under-sampled MRI images lying on a manifold [12, 13], where the weights were estimated using a special navigator acquisition. In this work, we propose an optimization problem with an ℓ_0 norm based fusion penalty, since we have observed that non-convex fusion penalties provide better clustering performance. We theoretically analyze the conditions for clustering data using the proposed optimization technique, when several features are missing. This analysis reveals that the clustering performance is determined by factors such as cluster-separation, cluster variance and feature coherence. When two clusters are distinguishable by very few features, then it is difficult to distinguish between them if these features are not observed, making feature coherence important. We also obtain a higher probability of successful clustering in the presence of fewer missing entries. We propose an algorithm to efficiently solve a relaxation of this optimization problem, using saturating non-convex fusion penalties. It is experimentally demonstrated that the proposed algorithm successfully clusters data in the presence of large fractions of missing entries.

2. CLUSTERING USING ℓ_0 FUSION PENALTY

2.1. Background

We consider the clustering of points drawn from one of K distinct clusters $C_1, C_2, \ldots, C_K$. We denote the center of the

This work is supported by NIH 1R01EB019961-01A1 and ONR-N000141310202.

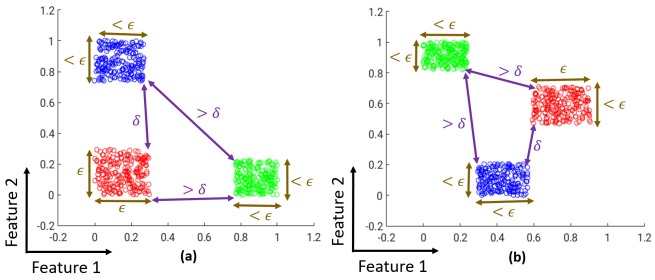

Fig. 1: Central Assumptions: (a) and (b) show different datasets of points $\in \mathbb{R}^2$ lying in 3 clusters (denoted by red, green and blue). A.1 and A.2 are illustrated in both (a) and (b). The importance of A.3 can be appreciated by comparing (a) and (b). In (a), points in the red and blue clusters cannot be distinguished using only feature 1, while the red and green clusters cannot be distinguished using only feature 2. Due to low coherence in (b), this problem does not arise.

clusters by $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K \in \mathbb{R}^P$. For simplicity, we assume that there are M points in each of the clusters. The individual points in the k^{th} cluster are modelled as:

$$\mathbf{z}_k(m) = \mathbf{c}_k + \mathbf{n}_k(m); \;\; m = 1, .., M, \; k = 1, \ldots, K \quad (1)$$

Here, $\mathbf{n}_k(m)$ is the noise or the variation of $\mathbf{z}_k(m)$ from the cluster center $\mathbf{c}_k$. The set of input points $\{\mathbf{x}_i\}, i = 1, .., KM$ is obtained as a random permutation of the points $\{\mathbf{z}_k(m)\}$. The objective of a clustering algorithm is to estimate the cluster labels, denoted by $\mathcal{C}(\mathbf{x}_i)$ for $i = 1, .., KM$.

The sum-of-norms (SON) method is a recently proposed convex clustering algorithm [8]. Here, a surrogate variable $\mathbf{u}_i$ is introduced for each point $\mathbf{x}_i$, which is an estimate of the centre of the cluster to which $\mathbf{x}_i$ belongs. In order to find the optimal $\{\mathbf{u}_i^*\}$, the following optimization problem is solved:

$$\{\mathbf{u}_i^*\} = \arg\min_{\{\mathbf{u}_i\}} \sum_{i=1}^{KM} \|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \lambda \sum_{i=1}^{KM} \sum_{j=1}^{KM} \|\mathbf{u}_i - \mathbf{u}_j\|_p \quad (2)$$

The fusion penalty ($\|\mathbf{u}_i - \mathbf{u}_j\|_p$) can be enforced using different ℓ_p norms, out of which the ℓ_1, ℓ_2 and ℓ_∞ norms have been used in literature [8]. The use of sparsity promoting fusion penalties encourages sparse differences $\mathbf{u}_i - \mathbf{u}_j$, which facilitates the clustering of the points $\{\mathbf{u}_i\}$.

2.2. Central Assumptions

We make the following assumptions (illustrated in Fig 1), which are key to the successful clustering of the points:

A.1: Cluster separation: Points from different clusters are separated by $\delta > 0$ in the ℓ_2 sense, i.e:

$$\min_{\{m,n\}} \|\mathbf{z}_k(m) - \mathbf{z}_l(n)\|_2 \geq \delta; \; \forall \; k \neq l \quad (3)$$

A.2: Cluster size: The maximum separation of points within any cluster in the ℓ_∞ sense is $\epsilon \geq 0$, i.e:

$$\max_{\{m,n\}} \|\mathbf{z}_k(m) - \mathbf{z}_k(n)\|_\infty = \epsilon; \; \forall k = 1, \ldots, K \quad (4)$$

A.3: Feature concentration: The coherence of a vector $\mathbf{y} \in \mathbb{R}^P$ is defined as: $\mu(\mathbf{y}) = \frac{P\|\mathbf{y}\|_\infty^2}{\|\mathbf{y}\|_2^2}$. We bound the coherence of the difference between points from different clusters as:

$$\max_{\{m,n\}} \mu(\mathbf{z}_k(m) - \mathbf{z}_l(n)) \leq \mu_0; \; \forall \; k \neq l \quad (5)$$

The quantity $\kappa = \frac{\epsilon\sqrt{P}}{\delta}$ is a measure of the difficulty of the clustering problem. The recovery of clusters when κ is small is expected to be easier.

2.3. Theoretical Guarantees

We study the problem of clustering $\{\mathbf{x}_i\}$ in the presence of entries missing uniformly at random. We arrange the points $\{\mathbf{x}_i\}$ as columns of a matrix $\mathbf{X}$. We assume that each entry of $\mathbf{X}$ is observed with probability p_0. The entries measured in the i^{th} column are denoted by:

$$\mathbf{y}_i = \mathbf{S}_i \, \mathbf{x}_i, \;\; i = 1, .., KM \quad (6)$$

where $\mathbf{S}_i$ is the sampling matrix, formed by selecting rows of the identity matrix. We consider solving the following optimization problem to obtain the cluster memberships from data with missing entries:

$$\begin{aligned} \{\mathbf{u}_i^*\} = &\min_{\{\mathbf{u}_i\}} \sum_{i=1}^{KM} \sum_{j=1}^{KM} \|\mathbf{u}_i - \mathbf{u}_j\|_{2,0} \\ &\text{s.t } \|\mathbf{S}_i \, (\mathbf{x}_i - \mathbf{u}_i)\|_\infty \leq \frac{\epsilon}{2}, i \in \{1 \ldots KM\} \end{aligned} \quad (7)$$

We claim that the above algorithm can successfully recover the clusters with high probability when the clusters are well separated (low κ), the sampling probability p_0 is sufficiently high and the coherence μ_0 is small. We state our theoretical guarantees after defining the following quantities:

- Upper bound for probability that two points have $< \frac{p_0^2 P}{2}$ commonly observed locations: $\gamma_0 := (\frac{e}{2})^{-\frac{p_0^2 P}{2}}$
- Given that two points from different clusters have $> \frac{p_0^2 P}{2}$ commonly observed locations, upper bound for probability that they can yield the same $\mathbf{u}$ without violating the constraints in (7): $\delta_0 := e^{-\frac{p_0^2 P(1-\kappa^2)^2}{\mu_0^2}}$
- Upper bound for probability that two points from different clusters can yield the same $\mathbf{u}$ without violating the constraints in (7): $\beta_0 := 1 - (1 - \delta_0)(1 - \gamma_0)$

- Upper bound for failure probability of (7): $\eta_0 := \sum_{\{m_j\}\in\mathcal{S}} \left[\beta_0^{\frac{1}{2}(M^2-\sum_j m_j^2)} \prod_j \binom{M}{m_j}\right]$ where $\mathcal{S}$ is the set of all sets of positive integers $\{m_j\}$ such that: $2 \leq \mathcal{U}(\{m_j\}) \leq K$ and $\sum_j m_j = M$. Here, the function $\mathcal{U}$ counts the number of non-zero elements in a set. For example, if $K = 2$ then $\eta_0 = \sum_{i=1}^{M-1} \left[\beta_0^{i(M-i)} \binom{M}{i}^2\right]$.
- For $K = 2$ and $\log \beta_0 \leq \frac{1}{M-1} + \frac{2}{M-2} \log \frac{1}{M-1}$, we have $\eta_0 \leq M^3 \beta_0^{M-1} := \eta_{0,\text{approx}}$.

Lemma 2.1. *Consider any two points* $\mathbf{x}_1$ *and* $\mathbf{x}_2$ *from the same cluster. A solution* $\mathbf{u}$ *exists for the following equations:*

$$\|\mathbf{S}_i(\mathbf{x}_i - \mathbf{u})\|_\infty \quad \leq \quad \frac{\epsilon}{2};\ \ i = 1, 2 \tag{8}$$

with probability 1.

Lemma 2.2. *Consider any two points* $\mathbf{x}_1$ *and* $\mathbf{x}_2$ *from different clusters, and assume that* $\kappa < 1$*. A solution* $\mathbf{u}$ *exists for the following equations:*

$$\|\mathbf{S}_i(\mathbf{x}_i - \mathbf{u})\|_\infty \quad \leq \quad \frac{\epsilon}{2};\ \ i = 1, 2 \tag{9}$$

with probability less than β_0*.*

The above lemmas indicate that two points from the same cluster can always be assigned the same centre $\mathbf{u}^*$. However, for a pair of points from different clusters, this can happen with a probability $< \beta_0$. We note that β_0 decreases with a decrease in κ. Using lemmas 2.1 and 2.2, we get the following result for a large number of points from multiple clusters:

Lemma 2.3. *Assume that* $\{\mathbf{x}_i : i \in \mathcal{I}, |\mathcal{I}| = M\}$ *is a set of points chosen randomly from multiple clusters (not all are from the same cluster). If* $\kappa < 1$*, a solution* $\mathbf{u}$ *does not exist for the following equations:*

$$\|\mathbf{S}_i(\mathbf{x}_i - \mathbf{u})\|_\infty \leq \frac{\epsilon}{2};\ \ \forall i \in \mathcal{I} \tag{10}$$

with probability exceeding $1 - \eta_0$*.*

We note here, that for a low value of β_0 and a high value of M, we will arrive at a very low value of η_0. Lemma 2.3 can be used to arrive at our main result:

Theorem 2.4. *If* $\kappa < 1$*, the solution to the optimization problem* (7) *is identical to the ground-truth clustering with probability exceeding* $1 - \eta_0$*.*

The reasoning follows from the fact that all solutions with cluster sizes smaller than M are associated with a higher cost than the ground-truth solution. In the special case where there are no missing entries, the constraints of optimization problem (7) reduce to: $\|\mathbf{x}_i - \mathbf{u}_i\|_\infty \leq \frac{\epsilon}{2}$. We have the following theorem guaranteeing successful recovery for the clusters:

Theorem 2.5. *If* $\kappa < 1$*, the solution to the optimization problem* (7) *is identical to the ground-truth clustering in the absence of missing entries.*

3. RELAXATION OF THE ℓ_0 PENALTY

We propose to solve the following relaxation of the optimization problem (7), which is more computationally feasible:

$$\begin{aligned}\{\mathbf{u}_i^*\} = \arg\min_{\{\mathbf{u}_i\}} \sum_{i=1}^{KM} \|\mathbf{S}_i(\mathbf{u}_i - \mathbf{x}_i)\|_2^2 \\ + \lambda \sum_{i=1}^{KM}\sum_{j=1}^{KM} \phi(\|\mathbf{u}_i - \mathbf{u}_j\|_2)\end{aligned} \tag{11}$$

Here ϕ is a function approximating the ℓ_0 norm, such as:

- ℓ_p norm: $\phi(x) = |x|^p$, for some $0 < p < 1$.
- H_1 penalty: $\phi(x) = 1 - e^{-\frac{x^2}{2\sigma^2}}$.

Similar to [14, 15], we reformulate the problem by majorizing the penalty ϕ using a quadratic surrogate functional: $\phi(x) \leq w(x)x^2 + d$, where $w(x) = \frac{\phi'(x)}{2x}$, and d is a constant. We now state the majorize-minimize formulation for problem (11) as:

$$\begin{aligned}\{\mathbf{u}_i^*, w_{ij}^*\} = \arg\min_{\{\mathbf{u}_i, w_{ij}\}} \sum_{i=1}^{KM} \|\mathbf{S}_i(\mathbf{u}_i - \mathbf{x}_i)\|_2^2 \\ + \lambda \sum_{i=1}^{KM}\sum_{j=1}^{KM} w_{ij}\|\mathbf{u}_i - \mathbf{u}_j\|_2^2\end{aligned} \tag{12}$$

We solve problem (12) by alternating between minimization with respect to $\{\mathbf{u}_i\}$ and $\{w_{ij}\}$ till convergence.

4. RESULTS

4.1. Study of Theoretical Guarantees

We observe the behaviour of $\gamma_0, \delta_0, \beta_0$ and η_0 as a function of p_0, P, κ and M. In Fig 2 (a), the change in γ_0 is shown as a function of p_0 for different values of P. In subsequent plots, we fix $P = 50$ and $\mu_0 = 1.5$. In Fig 2 (b), the change in δ_0 is shown as a function of p_0 for different values of κ. In Fig 2 (c), the behaviour of β_0 is shown. We consider $K = 2$ for subsequent plots. $(1 - \eta_0)$ is plotted in (d) as a function of p_0 for different values of κ and M. As expected, the probability of success of the clustering algorithm increases with decrease in κ and increase in p_0 and M.

4.2. Clustering of Simulated Data

We simulated datasets with $K = 2$ disjoint clusters in $\mathbb{R}^{50}$ with a varying number of points per cluster. The points in each cluster follow a uniform random distribution. We study the probability of success of the H_1 penalty based clustering algorithm as a function of κ, M and p_0. For a particular set of parameters the experiment was conducted 20 times. Fig 3 (a) shows the result for datasets with $\kappa = 0.39$ and $\mu_0 = 2.3$. The theoretical guarantees for successfully clustering the dataset

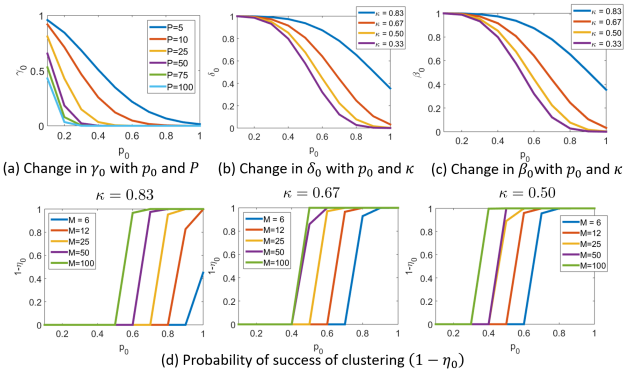

Fig. 2: Study of Theoretical Guarantees. Quantities γ_0, δ_0 and β_0 defined in Section 2.3 are studied in (a), (b) and (c). In (b), (c) and (d), $P = 50$ and $\mu_0 = 1.5$. As expected, β_0 decreases with increase in p_0 and decrease in κ. Considering $K = 2$ clusters, a lower bound for the probability of successful clustering $(1 - \eta_0)$ is shown in (d) for different κ.

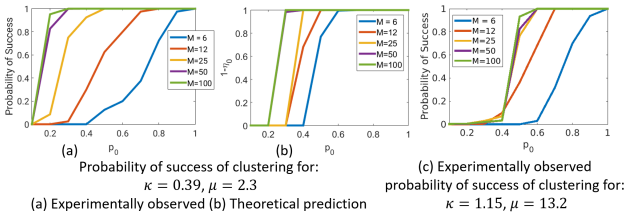

Fig. 3: Experimental results for probability of success. Guarantees are shown for a simulated dataset with $K = 2$ clusters. For (a) and (b), $\kappa = 0.39$ and $\mu_0 = 2.3$. (a) and (b) show the experimental and theoretical values for the probability of success respectively. (c) shows the experimentally obtained probability of success for a more challenging dataset with $\kappa = 1.15$ and $\mu_0 = 13.2$. We do not have theoretical guarantees for this case, since our analysis assumes $\kappa < 1$.

are shown in (b). Our theoretical guarantees hold for $\kappa < 1$. However, we demonstrate in (c) that even with $\kappa = 1.15$ and $\mu_0 = 13.2$, our clustering algorithm is successful.

Clustering results with $K = 3$ simulated clusters are shown in Fig 4. We simulated Dataset-1 with $K = 3$ disjoint clusters in $\mathbb{R}^{50}$ and $M = 200$ points in each cluster. For each of these 3 cluster centres, 200 noisy instances were generated by adding zero-mean white Gaussian noise of variance 0.1. The dataset was sub-sampled with varying fractions of missing entries ($p_0 = 1, 0.9, 0.8, \ldots, 0.3, 0.2$). We also generate Dataset-2 by halving the distance between the cluster centres, while keeping the intra-cluster variance fixed. We test the proposed algorithm on these datasets using the H_1 penalty. Since the points lie in $\mathbb{R}^{50}$, we take a PCA of the points and their estimated centres and plot the 2 most significant components. The 3 colours distinguish the points according

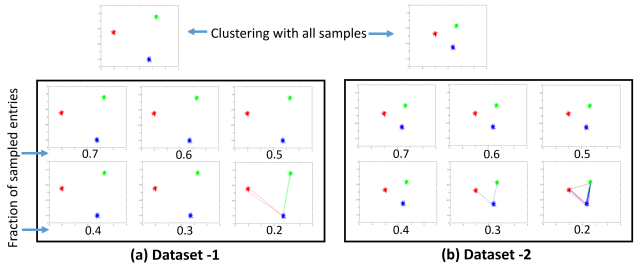

Fig. 4: Clustering results in simulated datasets. The H_1 penalty is used to cluster two datasets with varying fractions of missing entries. We show here the 2 most significant principal components of the solutions. The original points $\{\mathbf{x}_i\}$ are connected to their cluster centre estimates $\{\mathbf{u}_i^*\}$ by lines.

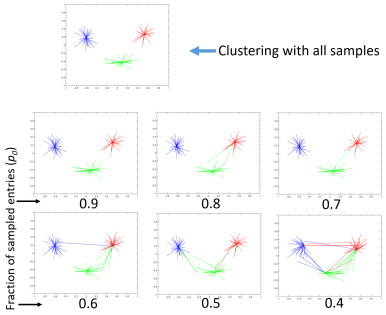

Fig. 5: Clustering the Wine dataset. The H_1 penalty is used for clustering with varying fractions of missing entries.

to their ground-truth clusters. Each point $\mathbf{x}_i$ is joined to its centre estimate $\mathbf{u}_i^*$ by a line. We observe that the clustering algorithm is more stable with fewer missing entries.

4.3. Clustering of Wine Dataset

We apply the clustering algorithm to the Wine dataset [16]. Each data point has $P = 13$ features. We created a dataset without outliers by retaining only $M = 40$ points per cluster, resulting in 120 points. The results are displayed in Fig 5 using the PCA technique as explained in the previous subsection. It is seen that the clustering is quite stable and degrades gradually with increasing fractions of missing entries.

5. CONCLUSION

We propose a clustering technique that can handle the presence of missing feature values. We derive theoretical guarantees for the successful recovery of the clusters using the proposed optimization problem. We also propose an algorithm to efficiently solve a relaxation of the above problem. This algorithm is demonstrated on simulated and real datasets. It is observed that the proposed scheme can perform clustering even in the presence of a large fraction of missing entries.

6. REFERENCES

[1] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin, "A review of clustering techniques and developments," *Neurocomputing*, 2017.

[2] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[3] M. C. De Souto, P. A. Jaskowiak, and I. G. Costa, "Impact of missing data imputation methods on gene expression clustering and classification," *BMC bioinformatics*, vol. 16, no. 1, p. 64, 2015.

[4] R. M. Bell, Y. Koren, and C. Volinsky, "The bellkor 2008 solution to the netflix prize," *Statistics Research Department at AT&T Research*, 2008.

[5] 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.

[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," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 10, pp. 617–621, 1979.

[8] T. D. Hocking, A. Joulin, F. Bach, and J.-P. Vert, "Clusterpath an algorithm for clustering using convex fusion penalties," in *28th international conference on machine learning*, 2011, p. 1.

[9] G. K. Chen, E. C. Chi, J. M. O. Ranola, and K. Lange, "Convex clustering: An attractive alternative to hierarchical clustering," *PLoS Comput Biol*, vol. 11, no. 5, p. e1004228, 2015.

[10] J. T. Chi, E. C. Chi, and R. G. Baraniuk, "k-pod: A method for k-means clustering of missing data," *The American Statistician*, vol. 70, no. 1, pp. 91–99, 2016.

[11] C. Zhu, H. Xu, C. Leng, and S. Yan, "Convex optimization procedure for clustering: Theoretical revisit," in *Advances in Neural Information Processing Systems*, 2014, pp. 1619–1627.

[12] S. Poddar and M. Jacob, "Dynamic mri using smoothness regularization on manifolds (storm)," *IEEE Tran. Medical Imaging*, vol. 35, no. 4, pp. 1106–1115, April 2016.

[13] S. Poddar, S. G. Lingala, and M. Jacob, "Joint recovery of under sampled signals on a manifold: Application to free breathing cardiac mri," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6904–6908.

[14] Y. Q. Mohsin, G. Ongie, and M. Jacob, "Iterative shrinkage algorithm for patch-smoothness regularized medical image recovery," *IEEE transactions on medical imaging*, vol. 34, no. 12, pp. 2417–2428, 2015.

[15] Z. Yang and M. Jacob, "Nonlocal regularization of inverse problems: A unified variational framework," *IEEE Transactions on Image Processing*, vol. 22, no. 8, pp. 3192–3203, Aug 2013.

[16] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml