Information in Missing Patterns: Enhancing Prediction Accuracy in Weighted Linear Regression with Missing Data Using Soft Clustering

Mohammad Amin Fakharian, Ashkan Esmaeili, and Yasaman Amiri Abyaneh
esmaeili.ashkan@alumni.stanford.edu

Abstract—**The linear system with missing information is investigated in this paper. New methods are introduced to improve the Mean Squared Error (MSE) on the test set in comparison to state-of-the-art methods, through appropriate tuning of Bias-Variance trade-off. The concept is to cluster the data and adapt the learning model to each cluster. Hence, we set forth a controlled bias into the problem and positively utilize it to enhance learning capability on the instances considered in some specific neighborhood. To deal with missing infrormation, we propose a novel algorithm "Missing-SCOP" based on SCOP-KMEANS algorithm introduced by Wagstaff, et al., utilizing the missing pattern of the dataset for construction of a soft-constraint matrix and clustering in missing scenario. It is shown that controlled over-fitting suggested by our algorithm improves prediction accuracy in various cases. Numerical experiments approve the efficacy of our proposed algorithm in enhancing the prediction accuracy.**

Index Terms. Missing Information, Soft-Impute, Linear Regression, Prediction, Soft-Clustering, Matrix Completion.

I. INTRODUCTION

RECENTLY, there has been a growing interest in enhancing prediction accuracy in machine learning. Although previous studies indicate that clustering may improve accuracy [23], training set shrinkage and data ignorance would be the penalties since it assigns hard weights to the subjects (i.e. each member has a weight parameter $w \in \{0, 1\}$).
In this paper, a novel weighted ensemble learning method of classification is presented based on weighted ensemble learning [16]. We call this method Soft Weighted Prediction (SWP), which weighs each cluster [1] obtained from training set (possibly each training example if they form a cluster themselves) based on its Euclidean distance from each test set subject.

Missing information has been gaining importance quite recently due to wide vision of applications it accompanies in practice as recommendation systems [6], [17], quantized rating systems and quantized data analysis [11], predictive sparse models with missing information [7], [10], semi-supervised learning with missing information [9]. Several clustering methods are developed in the literature to enhance prediction and regression accuracy. Several studies have been constructed on constrained clustering recently [15]. Hard and soft constrained clustering algorithms are aimed to modify the K-means algorithm to consider the side information regarding the connectivity graph of instances. Soft constrained clustering (SCC) concept, introduced by Kiri Wagstaff [25] known as KSCOP accounts for the baseline of our work. In this paper, we aim to extend the concept of SCC to prediction scenarios with missing information.

Data loss or idleness could be considered as a practical paradigm of inducing missing parameters in the structure of medical prediction problem. Obviously, in such cases missing values are not randomly distributed, e.g. patients suffering from the same disease, are more likely to be recorded with the same blood factors and symptoms. Thus, patients with similar missing factors, tend to be clustered together and have tendency to be reported with correlated medical diagnosis [2], [18]. This lack of similar recorded parameters (jointly missing parameters for subjects) is assumed as a constraint parameter in soft clustering. Prediction for medical data with missing information can be found at [20].

II. MODEL ASSUMPTIONS

In matrix representation, linear models are represented as follows:

$$Y = X\beta + \varepsilon, \tag{1}$$

where

$$\varepsilon \sim N(0, \sigma^2\mathbf{I})$$

$\mathbf{X}$ is the oracle instance-feature matrix. However, in practice, $\mathbf{X}$ is partially observed. Mathematically speaking, the observed matrix is obtained by applying a random mask on the original data matrix. The mask contains zeros on the entries which are missing or lost, i.e. we have access to a data matrix $\tilde{\mathbf{X}} = \mathbf{X} \bigodot \mathbf{M}$, where $\mathbf{M}$ is the oracle mask, and $\bigodot$ denotes the Hadamard product. Y is the observed measurement vector. β is the parameters (weights) coefficients.

A. Mathematical Approaches in Extracting the True Model (Imputing Coefficients)

Coefficients vector β could be estimated knowing $\mathbf{X}$ and Y as b. There are several regularization methods based on assumed constraints on vector β such as sparsity, to find the estimator b as it is not unique in many cases. However, our main concern is superior prediction of vector Y, not the coefficient. As Lasso constrains desired over-fitting, the Least-Square (LS) solution is used for each cluster in controlled bias setting.

1) Lasso Solution: Assuming β as a sparse vector, desired b will be obtained from optimization 2.

$$\min_b \frac{1}{2}||Y - Xb||_2^2 + \lambda||b||_1, \tag{2}$$

where parameter λ controls sparsity rate of coefficient β which is equivalent to balancing the trade-off.
Letting $\lambda = 0$, P2 turns into the ordinary least square problem. As λ approaches zero, this solution will have less bias and more variance errors. Thus, such is a data-dependent (training set) solution. As a result, test and train variation will lead to an inferior estimation and larger MSE. Further, as λ approaches infinity, b will be constrained to be sparse. Thus, training set variation effect decreases and estimator data-dependency will be omitted.

The least square solution is a particular case of $LASSO$ ($\lambda = 0$) which can be obtained from the normal equations are as follows:

$$(X^TX)b = X^TY$$

Solving for b,

$$b = (X^TX)^{-1}X^TY \tag{3}$$

Let $Y = X\beta$, adding noise $\varepsilon \sim N(0, \sigma^2\mathbf{I})$, the solution of the problem is:

$$b = (X^TX)^{-1}X^TY + (X^TX)^{-1}X^T\varepsilon \tag{4}$$

$$b = \beta + (X^TX)^{-1}X^T\varepsilon \tag{5}$$

Taking expectation yields to:

$$E[b] = E[\beta] + (X^TX)^{-1}X^TE[\varepsilon] \tag{6}$$

Knowing $E[\varepsilon] = 0$,

$$E[b] = \beta \tag{7}$$

Thus, unlike Lasso, least square solution is an unbiased estimation.

B. Controlled Overfitting

Overfitting occurs in test and training set variation cases. This error could be controlled by constraining the training set based on its similarity to each test example. This constraining could be done by either soft or hard weighting methods. In hard weighting algorithms training set would be shrunk to the most similar members to test example, such as clustering. On the other hand, Soft Weighting method prevents such data losses by applying a weighting mask based on similarities. Although **SWP** methods may cause accuracy reduction for estimator b specifically in sparse cases, more accurate Y estimation will be obtained. Specific estimator b is calculated for each test member based on its distance from X, which is not necessarily a good estimation of β, but more accurate prediction for Y. We can also refrain from separate estimation of β for each test sample by assigning each test sample to one cluster comparing its distance to different centroids determined by each cluster. As overfitting is controlled (by similarity) and satisfying in such scenarios, the introduced clustering algorithm, segments X and allocates each test set example, a cluster based on its Euclidean distance from its centroid. Thus, estimator b is trained by specific members, which results in increase of variance and reduction in bias term of predicted Y error. By increasing the number of clusters, overfitting and increase in variance term error will be seen. K-mapping [19] is one of the methods trying to optimize Bias-Variance trade-off [12]. The error expression is:

$$E[(y - \hat{f}(x))^2] = (f(x) - \frac{1}{k}\sum_{i=1}^{k} f(N_i(x)))^2 + \frac{\sigma^2}{k} + \sigma^2 \tag{8}$$

Supposing k nearest neighbors are chosen from the training set. Bias, which is the first term, has a monotonous rise as k increases, on the other hand, variance, the second term, drops off at the same time.
Although variance minimization leads to worse interpolation of training set, depending to its answer Y, it removes data dependency. Bias minimization has the reverse effect, i.e. although estimator b leads to the best Y calculation dependent to the specific training set X, vector b itself has larger MSE to the real coefficient coefficient β. Obviously in such cases if test data does not fit in any of the clusters, the estimated Y will face a larger error (large variance and small bias).

III. PROPOSED ALGORITHM

Clustering as a so-called method of tuning variance-bias trade-off has been studied and discussed in the literature recently as in [23]. Although simulations depicted enhancement of prediction responses in some cases, hard clustering results in uncontrolled overfitting and data loss.

As K-means Algorithm with squared Euclidean distance parameter is used for k-mapping, minimum distance of test set samples to centroid of clusters, leads to the appropriate assignment of test samples to each cluster. Following the least square solution, the predicted b is found. Multiplying test and estimator b, results in predicted Y matrix. As the number of clusters (k) increases, members of each cluster will decrease. Although this will lead to lower bias, variance term of error will increase. If test varies from training set, Estimated Y accuracy will be greatly depressed. Proposed solution to the problem is comprised of assigning each training set subject, specific weight based on its similarity to test sample. This filter is set to be an exponential function of distance. $\mathbf{W}$ is an $m \times 1$ matrix (filter) containing normalized distance between test and each training set subject. Parameter w controls the strength of filtering. As it approaches infinity, filter approaches one (no filtering).

The SWP algorithm is provided in Alg. 1. Obviously, all sub-figures of Fig. 1 in V-B1 depicts Bias-Variance tradeoff.

IV. TREATING WITH MISSING VALUES

Introduced methods are dependent on data matrix (training set). Considering missing values, clustering would not be possible (by k-means). Therefore, **SWP** algorithm requires a new definition of similarity to address the missing values.

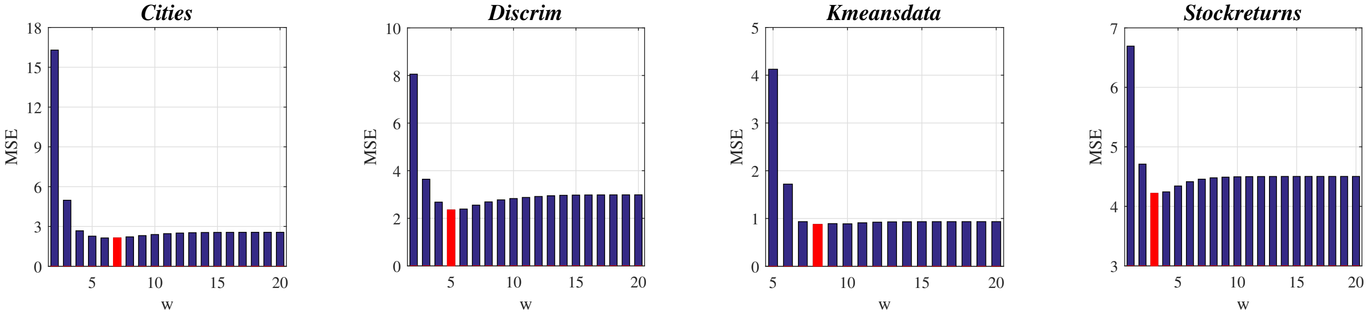

Fig. 1: MSE as a function of weight tuning parameter w.

Algorithm 1 SWP

Input: Training set X_{train}, Response vector Y_{train}, Test set X_{test}, Weight tuning parameter w

Output: Test set response vector Y_{test}

```
function SWP($X_{train}, Y_{train}, X_{test}, w$)
    for all $data_{new} = X_{test}(i,:)$ do
        $diff(j) := ||data_{new} - X_{train}(j,:)||_2^2$
        $diff \leftarrow \frac{diff}{min(diff)}$
        $W := diag(e^{\frac{-diff}{2^w}})$
        $b \leftarrow (X_{train}^T W X_{train})^{-1} X_{train}^T W Y_{train}$
        $Y_{test}(i,:) \leftarrow data_{new} \times b$
    end for
    return $Y_{test}$
end function
```

Algorithm 2 Missing-SCOP

Input: Training set X, Number of Clusters k, Proportional Tuning Parameter w

Output: Index vector idx, Centroids matrix C

```
function MISSING_SCOP($X, k, w$)
    $mask$ := not($X$==0)
    for all i,j do
        if i==j then
            continue
        end if
        $D_{miss}(i,j) := ||mask(x_i) - mask(x_j)||_2^2$
        $co\_mask(i,j) := mask(i,:) \odot mask(j,:)$
        $D_{dist}(i,j) := ||x_i - x_j||_2^2 \odot co\_mask(i,j)$
        $D(i,j) := w \times D_{miss}(i,j) + (1-w) \times D_{dist}(i,j)$
    end for
    $S(i,j) = 1 - 2\sqrt{\frac{D(i,j)}{\max D(:)}}$
    $[idx, C] \leftarrow$ SCOP_KMEANS [25] $(X, k, S)$
end function
```

If the missing is block-wise meaning that there are certain feature sets and a patient for example has either records for one feature set or not, then the clustering can be carried out based on the patient profiles. similarity in each profile can be addressed easily as the profiles are consistent among patients yielding to similar missing patterns. However, if the missing data is not block-wise, the non-missing pattern would differ among patients. Consequently, there is no similar profile based on which one can categorize the patients. Rather, we must infer from the data missing pattern how the patients may be similar. There are two approaches in dealing with non-block-wise missing data. The first is to impute the missing data followed by **SWP**. Therefore, we discuss a couple of off-the-shelf matrix efficient matrix completion and imputation methods next. A long list, however, can be found in [5], [4], [3].

A. *Imputation Methods*

1) Soft Impute[13]: In this method, Z is considered as a low-rank matrix. As $rank(Z)$ is a non-convex function, relaxation could be carried out by minimizing equivalent nuclear norm of Z. Finding matrix Z which satisfies 9, is desired.

$$||X - Z||_2^2 \; subject\ to \; ||Z||_* \leq \tau \tag{9}$$

The Lagrangian is given as:

$$\min_{Z} \frac{1}{2}||X - Z||_F^2 + \lambda||Z||_*, \tag{10}$$

The solution is given by Singular vlue thresholding (SVT) as follows:

$$S_\lambda := U(S - \lambda I)_+ V^T$$

Where $(S - \lambda I)_+$ is either positive or zero, otherwise.

To optimize the algorithm time complexity the suggested idea is to start Z from mean-estimation which makes iterative code converge faster.

2) MCPAT [8]: MCPAT is an efficient and adaptive matrix completion method which functions properly for highly missing scenarios which yields high SNRs in retrieving information.

B. Non-Impute Method

Soft-Impute, an Imputation method, applies low-rank restriction on the recovered dataset. Data loss is an inevitable consequence of the solution, as linearly dependent features could be ignored in clustering.
Many recent studies have focused on clustering datasets containing missing informations. Most common suggested solutions offer modifications to clustering algorithms such as KMEANS and FCM illustrated in [24] and [14], respectively. Although the main concern in such solutions are similarity of observed elements, it is worth noting that the same missing features represent a kind of resemblance in such scenarios. Balancing n-dimensional distance of observed data and missing features similarity by a weight tuning parameter leads to desired clustering.

1) Missing-SCOP: We have chosen SCOP-KMEANS Algorithm [25] as a baseline for the development of missing values clustering. As the real model dictates, missing pattern contains information and is profitable in clustering as a factor of similarity, i.e. we leverage the missing mask similarity of each pair in training set as a constraint in soft constrained clustering. Let matrix S be an $m \times m$ matrix, which assigns each pair $(x_i, x_j) \in \mathbf{X} \times \mathbf{X}$ a constraint $s \in [-1, 1]$. s is assigned based on mask similarities and jointly observed features Euclidean distance using a proportional tuning parameter w. As s approaches -1, the constraint forces separation. On the other hand, when s approaches 1, the two members of the pair must be clustered in the same group. Replicative Kmeans algorithm is employed in centroid initialization due to local minimum trap prevention.

2) SWP via Missing-SCOP: **SWP** algorithm consists of splitting the training set to one member clusters, and specifying each cluster a weight based on its distance to each individual. Another solution to the problem is soft clustering algorithms [21] utilization to find the probability matrix $\mathbf{U}$ for the test example. Thus, weight matrix is a diagonal matrix in which members of same clusters have the same weights.
As the problem contains missing values, introduced Missing-SCOP algorithm is used to obtain more precise clustering in comparison to imputation methods.
Let $\mathbf{X}$ be the dataset matrix, divided to $m \times n$ train set $\mathbf{X_{train}}$ and $p \times n$ test set $\mathbf{X_{test}}$. Assuming $\mathbf{X_{train}}$ is clustered into k sub-matrices by centroid matrix $\mathbf{C}$ and index vector idx, probability matrix $\mathbf{U}$ is defined in 11.

$$U = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1k} \\ u_{21} & u_{22} & \cdots & u_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pk} \end{bmatrix} \tag{11}$$

, where for each $i \in [1, p],\ j \in [1, k]$

$$u_{ij} := \frac{min\{u_{i1}, u_{i2}, ..., u_{ik}\}}{||X_{test}(i,:) - C(j,:)||_2^2} \tag{12}$$

Weight matrix W in **SWP** algorithm would be obtained by matrix $\mathbf{U}$, consequently. As u_{ij} is a normalized factor of similarity between i^{th} test set example and j^{th} cluster centroid, vector $W_{clusters}$ and matrix $\mathbf{W}$ are defined for each $\mathbf{X_{test}}$ example in 13 and 14 respectively.

$$W_{clusters} := e^{\frac{-(\mathbf{U}(i,:)^{-1})}{2^w}}, \tag{13}$$

which is calculated for i^{th} $\mathbf{X_{test}}$ example.

$$\mathbf{W} := diag\Big(W_{clusters}(j) \times (idx == j)\Big), \tag{14}$$

where $j = [1 : k]$.

Weighted least square solution in the algorithm requires matrix completion which could be obtained by MCPAT [8] algorithm.

V. SIMULATION RESULTS

A. Datasets

1) Simulated Data: As the real problems dictate, training set and test set are random processes which consist of normally distributed random sequences (features). Let X be an $m \times n$ random process consists of random variables $X = \{X_1, X_2, ..., X_n\}$ where $X_1, X_2, ..., X_n$ are normally distributed with uniformly random parameters i.e. $X_i \sim N(\mu, \sigma)$. As Law of Large Numbers (LLN) states, the average of the results obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed. Due to data-dependency of the simulation results, our reported MSEs are averaged on 20 generated random data.

2) Sample Data: Algorithms are also tested on following MATLAB sample datasets:

$$cities,\ discrim,\ kmeansdata,\ stockreturns$$

3) Missing Mask: Real cases depict significant and meaningful similarities in missing patterns of similar elements. Suggested missing mask consists of same missing pattern for each cluster in $Dataset$ matrix. A Gaussian logic mask is added to this mask as expected in real world. Considering $m \times n$ dataset $\mathbf{X}$ clustered into k sub-matrices each consists of $n_1, n_2, ..., n_k$ members by index vector idx. Explained $m \times n$ logic mask is generated as described in 15.

$$mask(idx == i, :) := ones(n_i, 1) \times \Big(r \geq (r_{max} \times m_{rate})\Big) \tag{15}$$

, where $i = [1 : k]$, $r_{max} = max(r(:))$, m_{rate} is the missing rate and $r_{1 \times n} \sim \mathbf{unif}$.

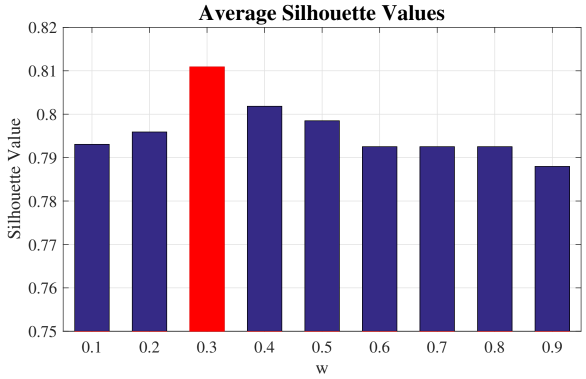

Fig. 2: Averaged Silhouette Values as a function of weight tuning parameter w tested on $kmeansdata$.

TABLE I: Silhouette Values of each solution.

Algorithm Dataset	Impute	non-Impute	no-Missing
$Cities$	0.3802	0.3829	0.4221
$Discrim$	0.4589	0.4716	0.6173
$Kmeansdata$	0.7958	0.8109	0.8606
$Stockreturns$	0.8111	00.8352	0.9585

B. No Missing Scenario

1) SWP: Algorithm is tested on datasets described in V-A. Results are respectively depicted in Fig. 1. Although optimal tuning parameter w varies from case to case, general behavior of the figures are the same.

C. Missing Scenario

Introduced methods dealing with missing elements of training set, are tested on mentioned datasets.

1) Clustering: Our main concern of dealing with missing cases is clustering. Impute and non-impute methods, introduced in Section IV are tested on datasets explained in V-A, which masked by the mentioned method.
Silhouettes [22] as a well-known method of clustering accuracy assessment is utilized. Simulation results are depicted in TABLE I to compare and find the efficiency of each clustering algorithm.
Silhouette values of $kmeansdata$ as an appropriate dataset for clustering are depicted in fig. 2. This figure illustrates a trade-off between missing mask similarity and observed values correlation tuned by parameter w described in algorithm 2. Notable improvement of clustering accuracy is observed in this case.

VI. CONCLUSION

An innovative method of prediction enhancement is introduced and explained on linear models. **SWP** algorithm as a developed weighted least square solution is suggested and surpassed many state-of-the-art methods such as clustering in simulation results. Datasets containing missing informations have been studied; adjusted SWP is developed for such scenarios, too. Clustering as a fundamental part of this adjustment is discussed and Missing-SCOP algorithm is introduced as a mean of handling missing values in clustering. Mentioned algorithm considers missing mask similarity of each example as a constraint of clustering by weight tuning parameter w. Comparing mean silhouette values as a factor of clustering precision, simulation results depicted that Missing-SCOP algorithm, a non-impute clustering method of cases with missing values, outperformed imputation methods like soft-impute.

REFERENCES

[1] Margareta Ackerman, Shai Ben-David, Simina Brânzei, and David Loker. Weighted clustering. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[2] Douglas G Altman and J Martin Bland. Missing data. *Bmj*, 334(7590):424–424, 2007.

[3] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982, 2010.

[4] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[5] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

[6] Michael D Ekstrand, John T Riedl, Joseph A Konstan, et al. Collaborative filtering recommender systems. *Foundations and Trends® in Human–Computer Interaction*, 4(2):81–173, 2011.

[7] Ashkan Esmaeili, Arash Amini, and Farokh Marvasti. Fast methods for recovering sparse parameters in linear low rank models. In *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1403–1407. IEEE, 2016.

[8] Ashkan Esmaeili, Ehsan Asadi, and Farokh Marvasti. Iterative null-space projection method with adaptive thresholding in sparse signal recovery and matrix completion. *arXiv preprint arXiv:1610.00287*, 2016.

[9] Ashkan Esmaeili, Kayhan Behdin, Mohammad Amin Fakharian, and Farokh Marvasti. Transductive multi-label learning from missing data using smoothed rank function. *Pattern Analysis and Applications*, 2020.

[10] Ashkan Esmaeili and Farokh Marvasti. Comparison of several sparse recovery methods for low rank matrices with random samples. In *2016 8th International Symposium on Telecommunications (IST)*, pages 191–195. IEEE, 2016.

[11] Ashkan Esmaeili and Farokh Marvasti. A novel approach to quantized matrix completion using huber loss measure. *IEEE Signal Processing Letters*, 26(2):337–341, 2019.

[12] Scott Fortmann-Roe. Understanding the bias-variance tradeoff. *I: URl: http://scott. fortmann-roe. com/docs/BiasVariance. html (hämtad 2017-09-29)*, 2012.

[13] T. Hastie, R. Mazumder, J. Lee, and R. Zadeh. Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares. *ArXiv e-prints*, October 2014.

[14] R. J. Hathaway and J. C. Bezdek. Fuzzy c-means clustering of incomplete data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 31(5):735–744, Oct 2001.

[15] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[16] Daniel Jiménez. Dynamically weighted ensemble neural networks for classification. In *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98CH36227)*, volume 1, pages 753–756. IEEE, 1998.

[17] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998, 2010.

[18] Colleen M Norris, William A Ghali, Merril L Knudtson, C David Naylor, and L Duncan Saunders. Dealing with missing data in observational health care outcome analyses. *Journal of clinical epidemiology*, 53(4):377–383, 2000.

[19] Leif E Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

[20] Archana Purwar and Sandeep Kumar Singh. Hybrid prediction model with missing value imputation for medical data. *Expert Systems with Applications*, 42(13):5621–5631, 2015.

[21] PS Raja and K Thangavel. Soft clustering based missing value imputation. In *Annual Convention of the Computer Society of India*, pages 119–133. Springer, 2016.

[22] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987.

[23] Shubhendu Trivedi, Zachary A. Pardos, and Neil T. Heffernan. The Utility of Clustering in Prediction Tasks. *CoRR*, abs/1509.06163, 2015.

[24] Kiri Wagstaff. *Clustering with Missing Values: No Imputation Required*, pages 649–658. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[25] Kiri Lou Wagstaff. *Intelligent Clustering with Instance-level Constraints*. PhD thesis, Ithaca, NY, USA, 2002. AAI3059148.