Effects of Matrix Completion on the Classification of Undersampled Human Activity Data Streams

Sofia Savvaki*†, Grigorios Tsagkatakis†, Athanasia Panousopoulou†, and Panagiotis Tsakalides*†
*Department of Computer Science, University of Crete
†Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH)
Heraklion, 70013, Greece

***Abstract*—Classification of activities of daily living is of paramount importance in modern healthcare applications. However, hardware monitoring constraints lead frequently to missing raw values, dramatically affecting the performance of machine learning algorithms. In this work, we study the problem of efficient estimation of missing linear acceleration and angular velocity measurements, experimenting on a public Human Activity Recognition (HAR) dataset. We exploit the data correlation to formulate the problem as an instance of low-rank Matrix Completion (MC) within a general classification framework. We consider the effects of our proposed reconstruction method on the classification accuracy as related to the size of the training and test sets, and the single versus collective recovery. Additionally, we compare the performance of our approach with popular imputation and expectation maximization algorithms for treating missing measurements, in conjunction with several state-of-the-art classifiers. The results highlight that robust and efficient classification is feasible even with a substantially reduced amount of measurements.**

I. Introduction

Human Activity Recognition (HAR) has received considerable attention over the last decade due to its numerous context-aware applications, particularly within the fields of health, well-being, and entertainment [1]. Sensor-based activity recognition integrates the emerging area of sensor networks with novel data mining and machine learning techniques to model a wide range of human activities [2]. Nonetheless, the evolution of activity recognition has led to an increasing number of challenges regarding sensor-based recognition systems.

The lack of a sufficient volume of data is a ubiquitous problem in many signal analysis areas, especially those depending on observational data, such as Human Activity Recognition. High data rates, energy limitations, memory constraints, and sensor failures, constitute only a subset of the existing factors leading to undersampled datasets. From the perspective of high level applications, these constraints contribute to a lack of sufficient data samples for efficient activity recognition. In this work, we explore the potential of accurate classification from a reduced amount of acquired data. To this end, we consider Matrix Completion (MC), a novel approach for estimating low rank matrices from a limited number of randomly selected entries [3], [4].

Formally, our objective is to assess the efficiency of MC in conjunction with supervised machine learning algorithms on HAR data streams. We aim to answer these key questions: Is efficient classification feasible from MC-reconstructed measurement matrices, as opposed to fully-populated ones? And if so, which could be the lower boundary of unobserved data?

We depart from current state-of-the-art by evaluating the applicability of inexact Augmented Lagrange Multipliers (ALM) based MC recovery [5], [6] on 3-axial linear acceleration and angular velocity data. We introduce a complete framework illustrated in Fig. 1 for data structuring, reconstruction, classification, and assessment of the overall recognition process in the presence of missing values. Without loss of generality, evaluation relies on a publicly available HAR dataset [7], [8], extensively used in literature [9]–[11]. The presented results provide useful insights on applying the MC framework in HAR data, and highlight the efficacy of our concept.

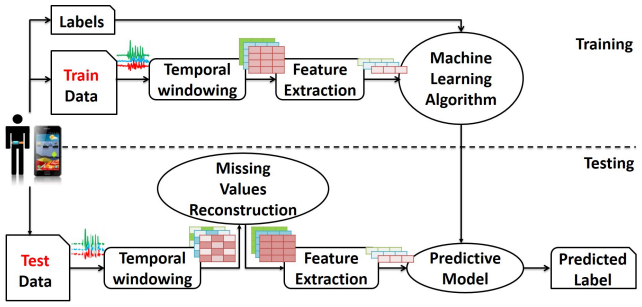


Fig. 1. The proposed framework. MC reconstruction incorporated on test phase. Subsequently, feature extraction and classification using the predictive models formed on training phase.

II. Matrix Completion

Consider a partially observed $n_1 \times n_2$ measurement matrix $\mathbf{M}$. In general, the recovery of the complete set of entries in a matrix using only $K << n_1 \times n_2$ entries is an underdetermined problem. However, it has been recently shown that such a recovery is possible, when imposing constraints on the number of missing entries and the rank of $\mathbf{M}$ [12]–[14]. Although one could seek an approximate matrix $\mathbf{X}$ by minimizing the rank [15], rank minimization is an NP-hard problem. Still, a relaxation of this problem produces accurate estimations by replacing the rank with the computationally tractable nuclear norm [12]. The relationship is manifested by the Singular Value Decomposition (SVD) of the $n_1 \times n_2$ measurements matrix. According to the spectral theorem associated with the SVD, the MC recovery problem can be expressed as:

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

$$\begin{aligned} &\underset{\mathbf{X}}{\text{minimize}} \quad \|\mathbf{X}\|_* \\ &\text{subject to} \quad \mathcal{A}(\mathbf{X}_{ij}) = \mathcal{A}(\mathbf{M}_{ij}), \quad (i,j) \in S \end{aligned} \tag{1}$$

where the nuclear norm is defined as the sum of the singular values of $\mathbf{X}$. The linear map $\mathcal{A}$, is defined as a random sampling operator recording a small number of entries from matrix $\mathbf{M}$, that is $\mathcal{A}(M_{ij}) = \{1 \text{ if } (i,j) \in S \mid 0 \text{ otherwise}\}$, where S is the sampling set. Recovery of the matrix is possible, provided it satisfies an incoherence property. In that case, the solution of Eq. (1) will converge to the solution of rank minimization with probability $1 - cq^{-3}$, once $K \geq Cq^{6/5} r log(q)$ random matrix entries are obtained, where $q = max(n_1, n_2)$, C and c are appropriate constants, and r is the matrix rank [12]. MC has been applied in various scenarios, including the reconstruction of water-treatment [4] and vital signs [16] data.

In this work, each undersampled time-series vector of length n is transformed into a partially observed $n_1 \times n_2$ measurement matrix $\mathbf{M}$ through a Hankelization process $\mathcal{H}$, where n_2 represents the window size. Consequently, we employ the Augmented Lagrange Multipliers (ALM) based MC to solve the nuclear norm minimization problem, due to its recovery performance and computational complexity [5], [6]. We assess data recovery for different cases of missing data, defined by the *fill ratio* $f = {}^{\#non\_zero\,elements}/_{n_1 \times n_2}$.

III. Methodology

Activity recognition is formulated as a classification problem (Fig. 1). On classification, there is a training phase for the classifiers and a test phase to evaluate the performance of the respectively produced predictive models.

A. Training phase

Temporal Windowing and Feature Extraction: 3-axial sensor streams are structured in Hankel matrices of $[n_1]$ consecutive lagged temporal windows of $[n_2]$ samples. Subsequently, feature extraction is applied to each window to obtain a vector of 22 statistical features, namely mean, standard deviation, min, max, 1st component of principal component analysis, interquartile range, variance, kurtosis, skewness, median, zero crossing rate, and an histogram of 10 bins.

Machine Learning Algorithms: For each data channel, the previously extracted feature vectors are utilized to train each classifier. We evaluate the performance of 3 state-of-the-art classifiers for HAR: a decision tree of 8 maximum splits and *Gini's diversity index* as a split criterion, K-Nearest Neighbours (KNN) with $k = 10$ and *Cosine / Euclidean* distance metrics, and Support Vector Machines (SVM) using *Gaussian / Quadratic* kernels, respectively. The aforementioned parameters have been fine-tuned through experimentation and provide the best predictive model for each classifier.

B. Testing phase

To evaluate the performance of the system, missing values are artificially introduced in the test streams. Without loss of generality, we consider zero as a missing value. In order to simulate realistic scenarios, we apply random zero-placement at the same instances for each *sensing modality*, i.e., for each set of 3-axial sensor streams. Consequently, data are segmented and structured in Hankel matrices which are naturally undersampled and need to be reconstructed before feature extraction is applied.

During the testing phase, we assess the reconstruction performance when considering data from a single or multiple sensors as shown in Fig. 2. Specifically, Scenario 1 defined as the *single sensor recovery* case, assumes that data recovery takes place on each Hankel matrix locally, by employing only data from a single modality (Fig. 2a). Scenario 2, defined as the *collective recovery* case, incorporates a central processing unit and applies the proposed reconstruction method to the collective measurements matrices that correspond to the vertical concatenation of the individual matrices per sensing modality (Fig. 2b). In Scenario 3, *collective recovery* is also employed, however at this point vertical concatenation involves all Hankel matrices grouped into a single one containing all data streams (Fig. 2c). Subsequently, for the two latter cases, data fusion is applied to form the initial, yet reconstructed Hankel matrices.

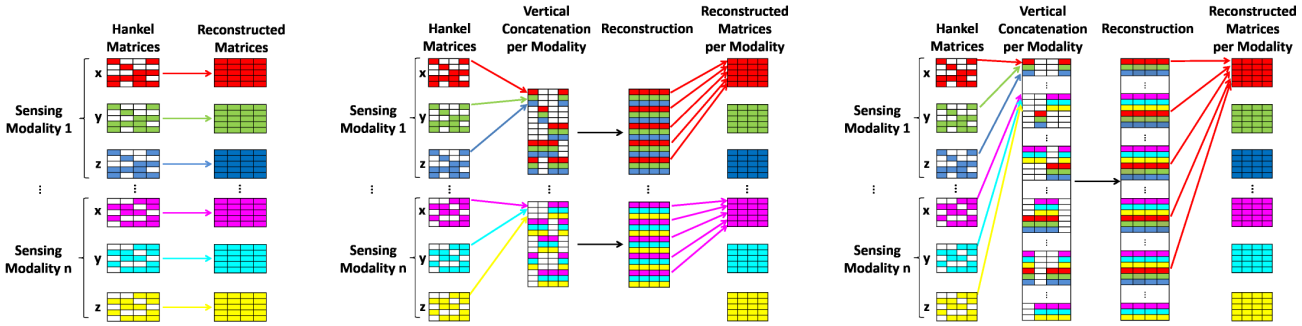
For each scenario, we employ the proposed MC recovery and compare its performance with two state-of-the-art data recovery methods, namely the k-Nearest Neighbour (k-NN) with k=1 [18], and the Regularized Expectation Minimization (RegEM) [17]. The reconstruction quality is evaluated using the Normalized Mean Square Error (NMSE) metric, defined as the mean squared error between the fully-populated and the reconstructed measurements matrix, normalized with respect to the l_2 norm.

IV. Experimental Evaluation

In this study, we consider a popular HAR database [7] created from the sensor recordings of 30 subjects performing 5 activities of daily living (Walking, Climbing Stairs, Sitting, Standing, Laying). A smart phone (Samsung Galaxy S II) attached on the waist was utilized as the recording device. 3-axial linear acceleration and angular velocity constitute the available sensing modalities, providing us with 6 data streams in total. Data were captured at 50Hz and pre-processed by applying noise filters, while the experiments were labelled manually through video-recordings. The obtained dataset was randomly partitioned into non-overlapping training and testing sets of varying sizes, structured in Hankel matrices of consecutive temporal windows of $[n_2 = 128]$ samples (2.56 seconds) with a 50% overlap [19].

A. Effects of training set size on classification performance

The objective of this experiment is to assess how the performance of the classifiers is associated to the size of the training set. The system is trained with numerous sizes of randomly selected data, corresponding to up to 21 subjects, i.e. 70% of the dataset. Subsequently, the classifiers are tested on the resulted predictive models and evaluated with respect to the classification accuracy on predicting the activities, namely labels, of the test set. For each training set, the test set



(a) Scenario 1: Reconstruction of each individual data matrix.

(b) Scenario 2: Reconstruction of vertically concatenated data matrices per sensing modality and fusion to form the individual, yet reconstructed, matrices.

(c) Scenario 3: Reconstruction of all vertically concatenated data and fusion to form the individual, yet reconstructed, matrices.

Fig. 2. Employed scenarios for missing values reconstruction.

considers the data of one randomly selected user. Note that in this experiment the test Hankel matrices, constructed from the data streams are fully-populated, i.e. $f = 1$, therefore providing a ground truth for our later experiments.

Fig. 3 illustrates the performance of each classifier measured by the classification accuracy as a function of the number of trained users. As expected, increasing the number of users in the training phase has a positive effect on the system's learning. All considered classifiers present stable performance when trained with at least 14 distinct users. This observation provides a useful insight regarding the amount of time necessary for the training phase. Since, our data streams are captured at a constant rate of 50 Hz, 7.25 minutes are needed to capture the data of an average user (340 windows), and therefore we can conclude that $7.25 \times 14 \simeq 100$ minutes of non-recurring train data, set a sufficient training period for our system. Moreover, we can observe that SVMs achieve the highest performance among all employed classifiers outperforming them by $3-4\%$, and managing over 90% accuracy.

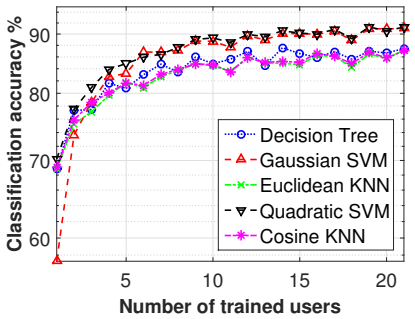


Fig. 3. Classification accuracy w.r.t. number of trained users for all classifiers with $f = 1$.

B. *Effects of test set size and fill ratio on reconstruction error*

In this set of experiments, we investigate the MC's recovery abilities with respect to the value of f and the size of the measurements matrix. The objective is to assess how the ALM-based NMSE is associated to the size of the data, by evaluating the recovery as a function of different sizes of

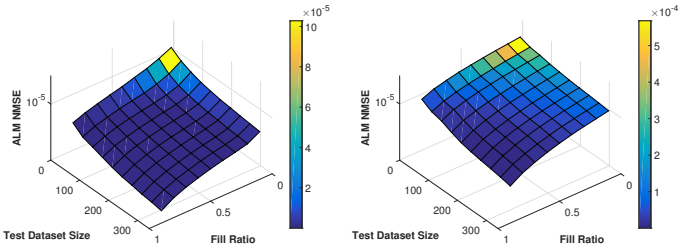


Fig. 4. NMSE w.r.t. f and test data size for x-axis accelerometer (left) and gyroscope (right) considering Scenario 1.

measurements matrices, ranging from $n_1 = 34$ up to 340 consecutive windows of $n_2 = 128$ samples and fill ratios from $f = 0.1$ to 0.9.

Fig. 4 illustrates the recovery performance measured by the NMSE as a function of f and the size of test data for Scenario 1. It is straightforward to observe, that higher fill ratios lead to more accurate measurements reconstruction as expected, in accordance to the theoretical models. Moreover, the size of the measurements matrix also plays a crucial role to the recovery performance, since larger matrices are clearly shown to present lower reconstruction error. This is a reasonable expectation, considering that larger data matrices contain a greater number of observed measurements, which can be exploited by the MC method for more accurate reconstruction of the unobserved ones. However, one cannot fail to notice that there is an important trade-off concerning the computational complexity, as the matrices grow to higher dimensions.

We were also interested in comparing our proposed MC-based reconstruction with the aforementioned popular recovery techniques. Fig. 5 depicts comparative plots of the NMSE and the corresponding processing times of all applied reconstruction methods as a function of f, on x-axis channels of both modalities available in the dataset at hand, namely accelerometer and gyroscope data of size $[n_1 = 340] \times [n_2 = 128]$.

It is observed that, when moving on to higher fill ratios, ALM outperforms all other employed imputation algorithms in terms of NMSE. However, for low fill ratios RegEM achieves better reconstruction quality, at the expense of its tremendous computational complexity, which is translated into

a dramatic increase in its running time. Consequently, RegEM can be considered inefficient for large scale data processing on commodity hardware, e.g. CPU. K-NN performs poorly in comparison with the other employed schemes for all fill ratios. Another key observation is related to the superior reconstruction quality of the accelerometer data as opposed to that of the gyroscope data. It is obvious that all reconstruction algorithms considered behave significantly better for the case of accelerometer data, due to higher linear correlation manifested by the smaller number of dominating singular values.

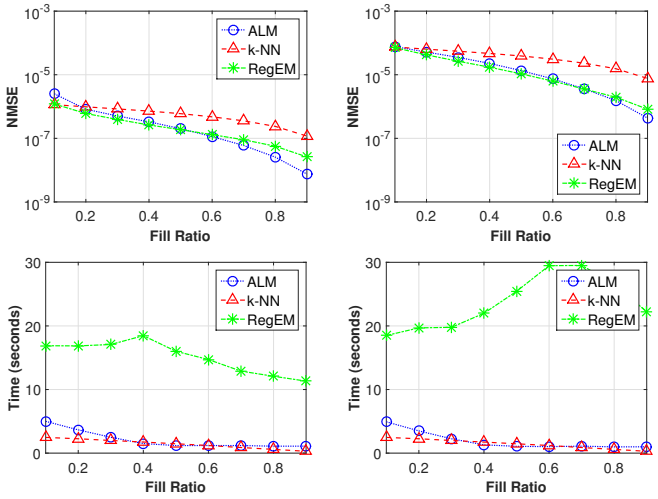


Fig. 5. NMSE (top) and corresponding times (bottom) w.r.t. f of all applied reconstruction methods for x-axis accelerometer (left) and gyroscope (right) data (Scenario 1).

C. Single vs Collective Recovery

In this experiment, we were interested in comparing the reconstruction performance between scenarios 1, 2, and 3, depicted in Fig. 2. ALM-recovery performance of each scenario with respect to f, is shown in Fig. 6, for x-axis accelerometer and x-axis gyroscope data respectively.

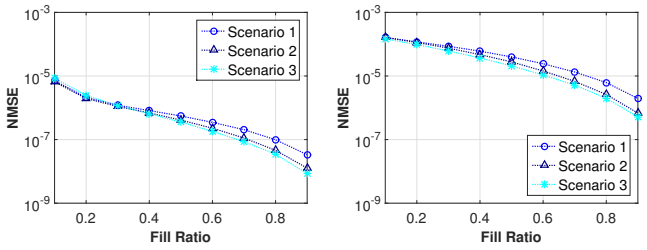


Fig. 6. ALM NMSE as a function of f w.r.t. all employed scenarios for x-axis accelerometer (left) and gyroscope (right).

It is remarkable that, collective recovery (Scenarios 2 and 3) significantly outperforms the single sensor case (Scenario 1) in terms of reconstruction, over all different values of f. Moreover, Scenario 3, involving both sensing modalities, even the less correlated gyroscopes, presents the most promising results. This outcome suggests that collective MC recovery can fully utilize the correlation that exists among sensors, even if such correlations are not explicitly encoded into the recovery process, thus highlighting the generalization ability of the proposed schemes.

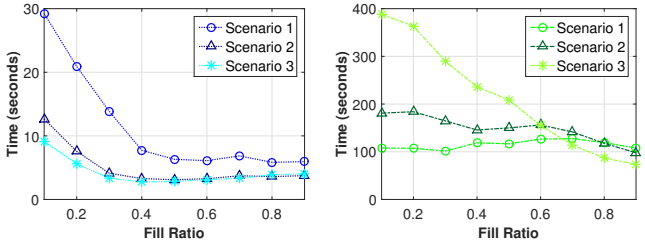


Fig. 7. ALM (left) and RegEM (right) reconstruction time w.r.t. f, considering all 3-axial data streams for Scenarios 1,2, and 3.

Furthermore, one would expect ALM to present higher running times for collective recovery, since it is dealing with measurements matrices of greater sizes. However, this is not the case. In fact, as highlighted in Fig. 7, Scenario 3 for ALM is the most efficient, also in terms of computational complexity. Specifically, instead of performing separate ALM reconstructions individually for all 6 relatively "small-scale" data streams (Scenario 1), we perform only one reconstruction call of higher computational complexity, which is cumulatively more cost-effective. Finally, RegEM algorithm presents an exponential increase in its running time for Scenario 3, as illustrated in Fig. 7.

D. Effects of reconstruction on classification accuracy

In the final set of experimental results, we incorporate the concept of missing measurements structuring and reconstruction into the overall classification process, and attempt to quantify the association of the reconstruction error with the resulting classification accuracy.

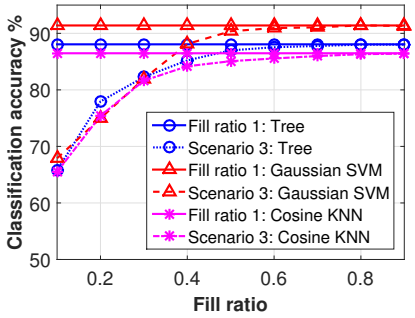


Fig. 8. Classification accuracy for ALM Scenario 3 recovery w.r.t. to f.

Fig. 8 illustrates the classification performance of an indicative subset of the employed classifiers, compared to the one achieved from a fully-populated matrix for ALM Scenario 3 reconstruction, with respect to f. A significant observation regarding the overall performance of our proposed framework is that, for all employed classifiers efficient classification accuracy can be achieved by extremely undersampled matrices, lacking half of their observations. More specifically, let us consider as ground truth the classification accuracy feasible

by the classifiers for fill ratio 1. We notice that, at fill ratio 0.5, all classifiers manage accuracy of only $1-2\%$ lower than the ground truth, whereas for $f = 0.6$ they nearly achieve optimal performance.

Fig. 9 outlines the classification performance of KNN with a Cosine distance metric for ALM Scenario 1 reconstruction, as a function of f and the size of test data. It presents the same behaviour with the corresponding Fig. 4 of MC as expected, and explicitly shows the direct relationship between NMSE metric and the resulting classification accuracy of the system.

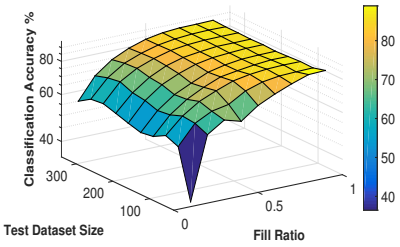


Fig. 9. Classification accuracy of Cosine KNN for ALM Scenario 1 reconstruction, w.r.t. f and test data size.

Finally, Fig. 10 represents the classification accuracy of SVM with a Gaussian kernel with respect to f, for ALM reconstruction considering all employed scenarios. It furtherly confirms our above drawn conclusions regarding the effectiveness of collective recovery, and gives a fairly good intuition, as far as the efficiency of ALM matrix completion is concerned, on the performance of the proposed scheme in truly lost or unavailable measurements.

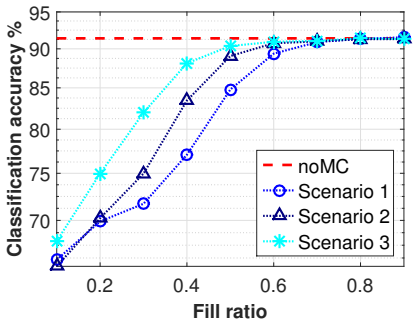


Fig. 10. Classification accuracy of Gaussian SVM for ALM Scenario 1,2,3 reconstruction, w.r.t. f.

V. Conclusion

In this work, we have investigated the effects of missing measurements reconstruction on the classification performance of human activity data streams. We have focused on Matrix Completion on artificially introduced missing data. Based on our experimental findings, we can conclude that effective classification accuracy is feasible even with only $50-60\%$ data observations. Furthermore, Matrix Completion shown to be a more efficient recovery method compared to other popular Imputation and Expectation Maximization algorithms, in terms of time as well as reconstruction performance. Additionally, collective recovery has been proved to achieve better reconstruction than single sensor recovery as it better exploits the correlations among the data. Future work could include the investigation of the effects of missing measurements reconstruction from multiple sensing modalities, as well as classification from heterogeneous sources.

Acknowledgment

This work was funded by the DEDALE project, contract no. 665044, within the H2020 Framework Program of the European Commission.

References

[1] O.D. Lara, and M.A. Labrador, *A survey on human activity recognition using wearable sensors*, Communications Surveys & Tutorials, IEEE, 15(3), pp. 1192-1209, 2013.

[2] M. Shoaib, S. Bosch, O.D. Incel, H. Scholten, and P.J. Havinga, *A survey of online activity recognition using mobile phones*, Sensors, 15(1), pp. 2059-2085, 2015.

[3] J. Cheng, Q. Ye, H. Jiang, D. Wang, and C. Wang, *STCDG: An efficient data gathering algorithm based on matrix completion for wireless sensor networks*, Wireless Commun., IEEE Trans. on, 12(2), pp. 850-861, 2013.

[4] S. Savvaki, G. Tsagkatakis, A. Panousopoulou, and P. Tsakalides, *Application of Matrix Completion on Water Treatment Data*, Proc. of CySWater, p. 3, 2015.

[5] Z. Lin, M. Chen, and Y. Ma, *The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices*, arXiv preprint arXiv:1009.5055, 2010.

[6] C. Chen, B. He, and X. Yuan, *Matrix completion via an alternating direction method*, IMA Journal of Numerical Analysis, 32(1), pp. 227-245, 2012.

[7] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J.L. Reyes-Ortiz, *A Public Domain Dataset for Human Activity Recognition using Smartphones*, ESANN, 2013.

[8] HAR dataset, https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones.

[9] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J.L. Reyes-Ortiz, *Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine*, Ambient assisted living and home care, Springer Berlin Heidelberg, pp. 216-223, 2012.

[10] J.W. Lockhart, and M.W. Gary, *Limitations with activity recognition methodology & data sets*, Proc. of UbiComp: Adjunct Publication, ACM, pp. 747-756, 2014.

[11] J.L. Reyes-Ortiz, A. Ghio, X. Parra, D. Anguita, J. Cabestany, and A. Catal, *Human Activity and Motion Disorder Recognition: towards smarter Interactive Cognitive Environments*, ESANN, 2013.

[12] E.J. Candes, and B. Recht, *Exact matrix completion via convex optimization*, Foundations of Computational mathematics, 9(6), pp. 717-772, 2009.

[13] E.J. Candes, and T. Tao, *The power of convex relaxation: Near-optimal matrix completion*, Information Theory, IEEE Trans., 56(5), pp. 2053-2080, 2010.

[14] E.J. Candes, and Y. Plan, *Matrix completion with noise*, Proc. of the IEEE, 98(6), pp. 925-936, 2010.

[15] B. Recht, M. Fazel, and P.A. Parrilo, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, SIAM review, 52(3), pp. 471-501, 2010.

[16] S. Yang, K. Kalpakis, C.F. Mackenzie, L.G. Stansbury, D.M. Stein, T.M. Scalea, and P.F. Hu, 2012, *Online recovery of missing values in vital signs data streams using low-rank matrix completion*, ICMLA, Vol. 1, pp. 281-287, IEEE, 2012.

[17] T. Schneider, *Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values*, Journal of Climate, 14(5), pp. 853-871, 2001.

[18] E. Acuna, and C. Rodriguez, *The treatment of missing values and its effect on classifier accuracy*, Classification, clustering, and data mining applications, Springer Berlin Heidelberg, pp. 639-647, 2004.

[19] L. Bao, and S.S. Intille, *Activity recognition from user-annotated acceleration data*, Pervasive computing, Springer Berlin Heidelberg, pp. 1-17, 2004.