Coping with missing data in an unobtrusive monitoring system for office workers

Stavros Nousias [1], Giwrgos Papoulias [1], Otilia Kocsis [1], Miriam Cabrita [3],
Aris S. Lalos [1,2] and Konstantinos Moustakas [1]

[1] Department of Electrical & Computer Engineering, University of Patras, Greece
[2] Industrial Systems Institute, Athena Research Center
[3] University of Twente, Enschede, the Netherlands

***Abstract*—Current trend of population ageing at global level is accompanied by increased prevalence of chronic diseases and higher rates of early retirement and labor market exit. In particular, the lifestyle of office workers is characterized by prolonged sitting and overall sedentary life, which alone is a high risk factor for cardiometabolic diseases, obesity and other related chronic diseases. The SmartWork unobtrusive monitoring system allows for continuous monitoring of various lifestyle, health, behavioural and work related parameters of office workers targeting to empower work ability sustainability. The large amounts of collected data in such systems are often characterized by the presence of missing entries. This work is an exploratory study on the potential of a Laplacian matrix completion variant for data imputation on the multi-channel time-series data collected with wearable or work devices in the SmartWork system.**

***Index Terms*—missing data, heart rate, activity tracker, office worker, data completion**

I. Introduction

Unobtrusive monitoring of daily activities and physiological signs has started to be deployed at large especially for older adults, being empowered by technology availability on one side (e.g. decreasing prices of wearable devices) and the society needs with respect to prolonged active and healthy life years of citizens. The societal needs are driven by the global trend of ageing of the population, which is severely challenging the health and pension systems, especially in Europe. As people get older, health chronic condition prevalence increases, resulting in high rates of early retirement and labour market exit, especially for people aged 55-64 [1]. A wide range of ambient assistive technologies have been tested and are currently piloted at large scale in smart homes, supporting home health monitoring of elderly people (over 65 years old) for their independent living, but in most cases such solutions are either too focused (e.g. disease specific) or not available/transferable (e.g. home embedded sensing technologies) at the workplace or on the move to support professionally active ageing.

On the technology side, the interconnected systems forming body area networks (e.g. activity tracker along with smart phone applications) in connection with increased bandwidth of wireless communication (e.g. 5G mobile network) and the exponential expansion of the Internet of Things (IoT) makes it possible to easily transfer and store large volumes of data.

Accounting for both, societal needs and technological advances, SmartWork project aims to build a Worker-Centric AI System for work ability sustainability, by integrating unobtrusive sensing and modelling of the worker state with a suite of innovative services for context and worker-aware adaptive work support [2]. Work ability is directly linked to the functional abilities and cognitive capacity of the worker, which are continuously assessed by unobtrusively and pervasively monitoring the health, behavioral, cognitive and emotional status of the office worker. The lifestyle of office workers is characterized by prolonged sitting and overall sedentary life, which alone is a high risk factor for cardiometabolic diseases, obesity and other related chronic diseases (e.g. diabetes) especially at older ages, as low physical activity is associated with increased morbidity and premature mortality [3]. The exploitation of the large volumes of person-generated health data (PGHD) in SmartWork, collected by personal wearable devices (e.g. Fitbit activity tracker) and work devices (e.g. smart mouse), together with advanced processing tools to interpret the data and implement decision support systems on various personal and work life dimensions (e.g. health, lifestyle, work flexibility) is a prerequisite in order to prolong the healthy and active life years of office workers.//

When it comes to big data, it is important to ensure reliable collection and transmission, along with making the processes more efficient to facilitate real-time analysis and optimized storage [4]. Collection of dense time-series (e.g. heart rate data acquired using an activity tracker) is often characterized by the presence of missing entries, which may impact on the value of data for the application and end users. Furthermore, efficient transmission and storage may even require lossy data compression (irreversible) on the edge (e.g. sensing or mobile device) or before the long-term storage, followed by recovery of intentionally removed data on the server side or whenever the complete time-series data is needed for further analysis. Sparse modelling and optimization tools, such as low-rank matrix completion, facilitate the sparse non-uniform sampling of monitored processes, allowing the recovery of corrupted, missing or intentionally removed entries [5], [6].

This work is an exploratory study on the potential of a Laplacian matrix completion variant for data imputation

978-1-7281-4754-3/19/$31.00 ©2019 IEEE / BIA-2019, ID 48344X

on various time-series data collected with wearable or work devices in the SmartWork unobtrusive monitoring system [2]. The usage of the proposed algorithm for big data is investigated by considering variable matrix size and by comparing the results to other imputation approaches. The remaining of the paper is organized as follows: section II presents the proposed Laplacian matrix completion method for multidimensional data; section III presents the experimental study performed; and section IV discusses the results and concludes this study.

TABLE I: Summary of Notations

$a, \mathbf{a}$ and $\mathbf{A}$	Scalar, vector and matrix variables
$[\mathbf{A}]_{ij}$	Matrix element at the i-th row and j-th column
$\mathbf{I}_N$	$N \times N$ identity matrix
$\mathbf{0}_{N\times K}$	$N \times K$ matrix with zeros
c_{p_k}	p_{th} estimation of the k_{th} channel
Ω	Set containing matrix positions of observed entries
$\|\cdot\|_*$, $\|\cdot\|_F$	Nuclear and Frobenius norms of matrix
$\circ$	Element-wise (Hadamard) matrix product
SVT	Singular Value Thresholding with threshold t
$\{x\}$	Fractional part of x
$\lfloor x \rfloor$	Integer part of x

II. Laplacian matrix completion for multidimensional data

This section presents an efficient Laplacian matrix completion adaptation scheme for multi-channel data modelling. We assume that a sparse matrix $\mathbf{M} \in \mathbb{R}^{n\times m}$ represents the available observations from a multi-channel sampled process S. For a given number of estimations of the process P we denote the p_{th} estimation as $c_{p_k}, p \in [1, 2, \cdots, w]$, $k \in [1, 2, \cdots, k_c]$ is the channel index, k_c is the number of channels. To take advantage of the aforementioned properties we formulate square or near square matrices by partitioning and stacking process into matrix $\mathbf{S} \in \mathbb{R}^{n\times m}$ where $w = n \cdot m$. For a two channel sampled process we would have the following matrix.

$$\mathbf{S} = \begin{bmatrix} c_{1_1} & c_{(n+1)_1} & \cdots & c_{(m(n-1)+1)_1} \\ c_{1_2} & c_{(n+1)_2} & \cdots & c_{(m(n-1)+1)_2} \\ c_{2_1} & c_{(n+2)_1} & \cdots & c_{(m(n-1)+2)_1} \\ c_{2_2} & c_{(n+2)_2} & \cdots & c_{(m(n-1)+2)_2} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n_1} & c_{(2n)_1} & \cdots & c_{(m\cdot n)_1} \\ c_{n_2} & c_{(2n)_2} & \cdots & c_{(m\cdot n)_2} \end{bmatrix} \tag{1}$$

We also denote $\mathbf{S}_{s_i s_j}$ the element $\mathbf{S}$ located at s_i row and s_j column. The equivalent vector $\mathbf{r}$ can be expressed as

$$\mathbf{r} = [c_{1_1}\, c_{1_2}\, c_{2_1}\, c_{2_2} \cdots c_{(m\cdot n)_1}\, c_{(m\cdot n)_2}] \tag{2}$$

where the r_{th} observation can be expressed as

$$r = (p-1)n_c + k \tag{3}$$

p is the p_{th} observation of P, n_c the number of channels and k the k_{th} channel. Then Ω is the subset of indices of the sampled process that are assumed to be known. The reshaping dictates the translation of the known entries of timeseries so that $\mathbf{S}_{s_i s_j}$, $(s_i, s_j) \in E_\Omega$ is the set of known entries of the reshaped matrix $\mathbf{S}$.

$$\mathbf{U} = \begin{cases} 1 & (s_i, s_j) \in E_\Omega \\ 0 & otherwise \end{cases} \tag{4}$$

Any data recovery approach aims to efficiently recover missing entries generating an estimation very close to the ground-truth.

Candes et al. [7] established a low-rank sparse matrix $\mathbf{S}$ can be perfectly recovered solving the nuclear norm optimization problem

$$\textit{minimize}\ \tau\|\mathbf{X}\|_* \quad \textit{s.t.}\ [\mathbf{X}]_{ij} = [\mathbf{S}]_{ij} \tag{5}$$

where $(i, j) \in E_\Omega$, $\mathbf{X}_*$ is the nuclear norm of the optimization variable $\mathbf{X}$ and τ is a weighting parameter depending on the matrix rank. In order to minimize (5) singular value thresholding D_τ [8] can be employed.

In our multi-channel setting we require to establish constraints taking advantage of the local coherence [9]. Two entries r_i and r_j of $\mathbf{r}$ are generated from the same channel k if

$$p_i + \frac{k}{n_c} = \frac{r_i + n_c}{n_c} \quad \text{and} \quad p_j + \frac{k}{n_c} = \frac{r_j + n_c}{n_c} \tag{6}$$

Subsequently, we formulate the generalized adjacency matrix $\mathbf{A} \in \mathbb{R}^{m\times m}$ so that the estimation of a missing entry from channel k is inversely proportional to the temporal distance between the estimated value and known values of the same channel.

$$\mathbf{A}_\Omega = \begin{cases} 1 & i \in \Omega, j = i \\ |i-j| & i \notin \Omega, \left\{\frac{i+n_c}{n_c}\right\} = \left\{\frac{j+n_c}{n_c}\right\}, j \in N_k(i) : \Omega \\ 0 & otherwise \end{cases} \tag{7}$$

The term

$$\left\{\frac{i+n_c}{n_c}\right\} = \left\{\frac{j+n_c}{n_c}\right\} \tag{8}$$

ensures that i and j are indices of $\mathbf{r}$ generated from the same channel and $\{x\}$ denotes the fractional part of x. We also denote $N_k(i) : \Omega$ as set of k nearest neighbors of i in Ω. and $\mathbf{D} \in \mathbb{R}^{w,w}$ as the diagonal matrix $[\mathbf{D}]_{ii} = |N(i)|$ where $N(i)$ is the row-wise sum of the i-th row of $\mathbf{A}_\Omega$. Then the Laplacian matrix $\mathbf{L}$ is defined as

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A} \tag{9}$$

To establish the temporal smoothness constraint near the missing component $x_i, i \notin \Omega$ of the time-series x and the weighted mean of the available nearest neighbours of the same channel we employ the following constraint:

$$\|\mathbf{L}vec(\mathbf{X})\|_2^2 = \sum_i \left\| x_i - \sum_{j\in N_k(i):\Omega} \frac{\mathbf{D}_{i,i} - \mathbf{A}_{ij}}{\mathbf{D}_{i,i}} x_j \right\|_2^2 \tag{10}$$

Finally, the optimization problem can be formulated as:

$$\min_{\mathbf{X}} \frac{1}{2} \|\mathbf{U} \circ (\mathbf{X} - \mathbf{S})\| + \tau \|\mathbf{X}\|_* + \mu \|\mathbf{L}vec(\mathbf{X})\|_2^2 \tag{11}$$

where the first term minimizes the error between the known values and the estimated, the second term imposes low rank

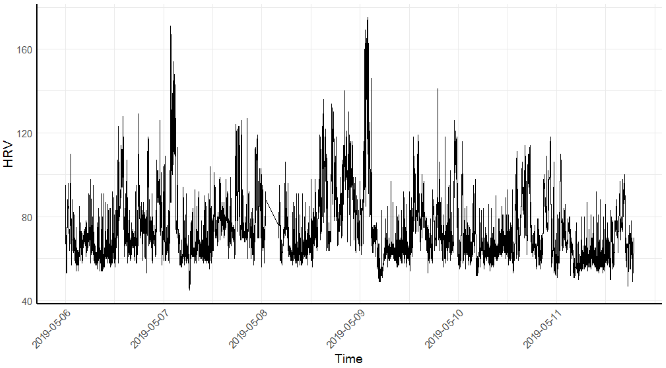


Fig. 1: Heart rate measurements from 2019-05-06 to 2019-05-11

to the recovered matrix and the last term moves the estimated value "near" to the weighted average of the k-nearest known entries neighbours. The parameter μ is the regularization parameter of the Laplacian and $\|.\|_F$ is the Frobenius norm. Equation (11) can be efficiently minimized using alternating direction method of multipliers (ADMM) on the splitting version of the equivalent augmented Lagrangian [10], [11] also presented in [6], [9]

III. Experimental Study

A. Dataset description and simulation setup

In total 57114 heart rate measurements and equal number of step measurements were collected with a FitBit activity tracker for a period of five consecutive days leading to the formulation of a 334×342 near-square matrix adopting the multi-channel setup described in (1). In order for the validation process to be realized in a robust manner for providing reliable comparisons across different methods, a Missing Completely At Random (MCAR) approach was employed. Thus, missing entries were generated so that their percentage over the total samples ranges from 5% to 50%. For each distinct level on the percentage scale of missing values, 20 permutations were generated for ensuring the construct validity of our employed methods. Figure 2 depicts the distribution of singular values for each singular component. The number of $k = 100$ singular values were observed to sum up at least 95% of the nuclear norm ($\mathbf{X}_*$) of the examined matrix. In particular, two setups are examined: a) a single channel setup using as input only HR data and b) a multichannel setup, using both HR measurements and steps aligned for the same time interval.

B. Multidimensional Data Completion Results

1) Single channel case: We compared Laplacian matrix completion (LMC) with classical matrix completion (MC), k-nearest neighbours imputation (KNN) [12] and missForest (MF) [13]. Figure 4 presents the reconstruction error for different levels of missing entries as a mean value of 20 permutations for each level demonstrating that LMC appears to demonstrate the lowest error using the Normalized Root Mean Square Error (NRMSE) [14]. Figure 5 presents the distribution of reconstruction error across all permutations revealing that all methods demonstrate similarly small deviation.

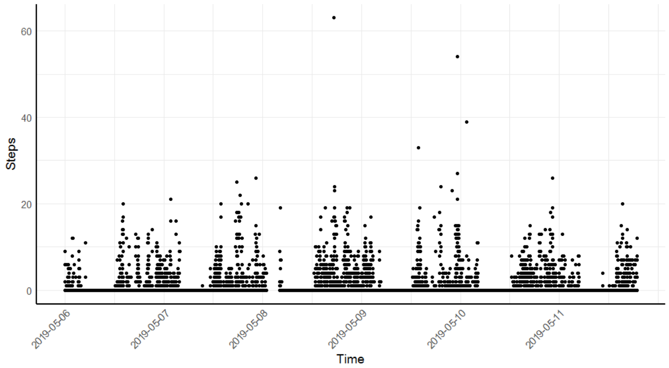
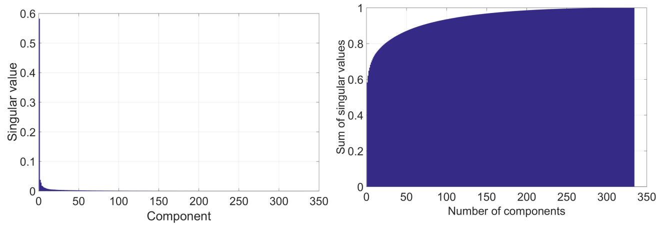


Fig. 2: Steps taken from 2019-05-06 to 2019-05-11



(a) (b)

Fig. 3: a) Singular values as a function of singular component index b) Sum of singular values as a function of singular component index

2) Multi channel case: Likewise compared Laplacian matrix completion (LMC) with classical matrix completion (MC), k-nearest neighbours imputation (KNN) [12] and missForest (MF) [13]. Figure 6 presents the reconstruction error for different levels of missing entries as a mean value of 20 permutations for each level demonstrating that LMC appears to demonstrate the lowest error. Figure 7 presents the distribution of reconstruction error across all permutations. It is important to highlight what KNN yields very low reconstruction accuracy due to the matrix format containing both channels in nearby positions. Evaluation of KNN accuracy can be better revealed in single channel case where the reconstruction results of KNN are worse but yet close to MF and MC.

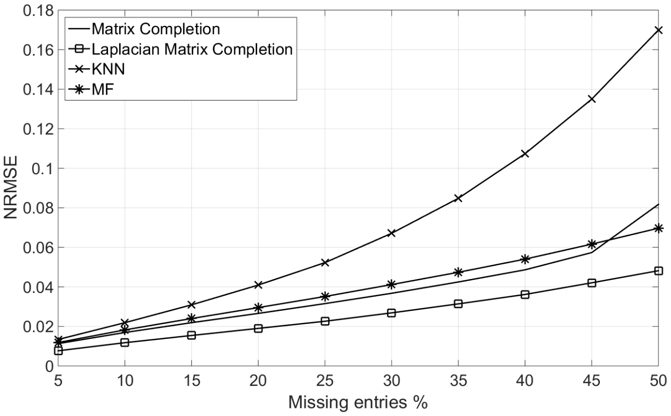


Fig. 4: Reconstruction error as a function of level of missing entries for the single channel case

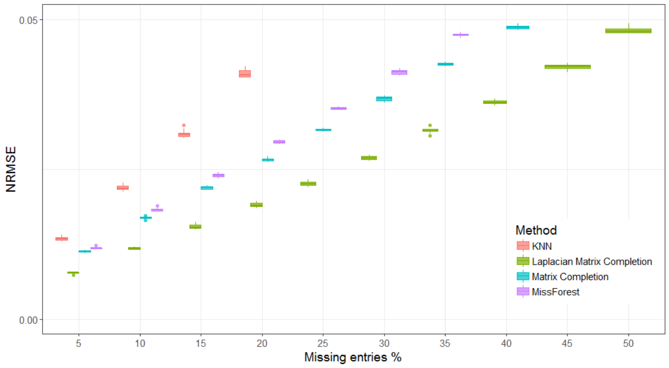


Fig. 5: Distribution of reconstruction error as a function of level of missing entries for the single channel case

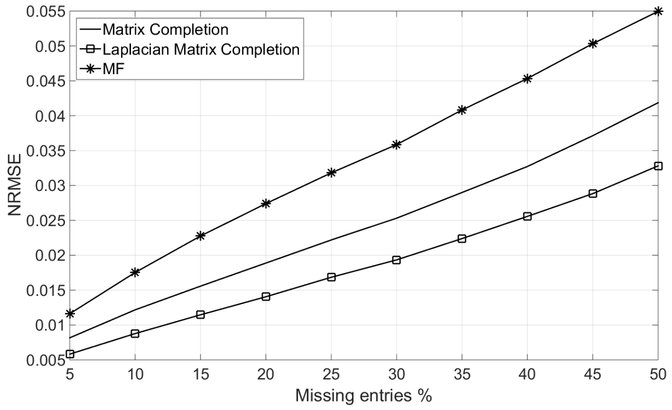


Fig. 6: Reconstruction error as a function of level of missing entries for the multi-channel case

IV. Discussion and Conclusion

The current study only focused on the MCAR problem, although it has been observed that the collection of heterogeneous data sets (multi-channel) in SmartWork is also by block missing data. Taking into account the nature of data (e.g. health related), in such cases only alternative channel data is used, when available, to complete the parameter specific time-series (e.g. HR is collected both by the activity tracker and by the smart mouse). This work is an exploratory study on the potential of a Laplacian matrix completion variant for data imputation on multi-channel time-series data collected with an activity tracker in the SmartWork unobtrusive monitoring system. We investigated the scenario where the formed matrices exhibit low-rank properties with missing entries ranging from 5% to 50%. The proposed approach was compared with classical matrix completion, KNN and missForest imputation methods, demonstrating the lowest error.

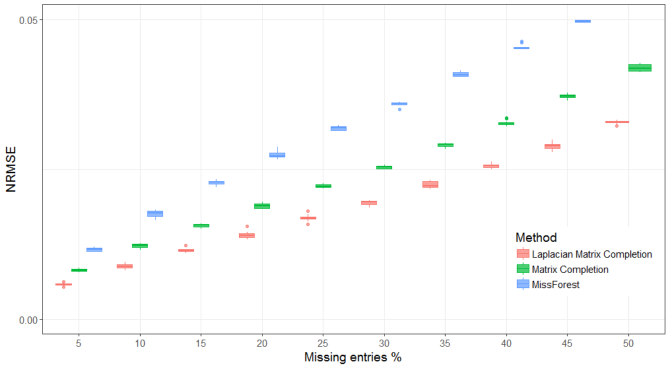


Fig. 7: Distribution of reconstruction error as a function of level of missing entries for the multi-channel case

Acknowledgment

This work has been partially supported by the SmartWork project (GA 826343), EU H2020, SC1-DTH-03-2018 - Adaptive smart working and living environments supporting active and healthy ageing.

References

[1] E. Commission. (2017) Eurostat: Employment and unemployment statistics. [Online]. Available: http://ec.europa.eu/eurostat/web/lfs/data/main-tables

[2] O. Kocsis, K. Moustakas, N. Fakotakis, C. Vassiliou, A. Toska, G. C. Vanderheiden, A. Stergiou, D. Amaxilatis, A. Pardal, J. Quintas *et al.*, “Smartwork: designing a smart age-friendly living and working environment for office workers,” in *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. ACM, 2019, pp. 435–441.

[3] S. Parry and L. Straker, “The contribution of office work to sedentary behaviour associated risk,” *BMC public health*, vol. 13, no. 1, p. 296, 2013.

[4] S. Nousias, A. S. Lalos, A. Kalogeras, C. Alexakos, C. Koulamas, and K. Moustakas, “Sparse modeling and optimization tools for energy efficient and reliable IoT,” in *Societal Automation*, Krakow,Poland, 2019.

[5] Z. Alansari, N. B. Anuar, A. Kamsin, S. Soomro, M. R. Belgaum, M. H. Miraz, and J. Alshaer, “Challenges of internet of things and big data integration,” in *International Conference for Emerging Technologies in Computing*. Springer, 2018, pp. 47–55.

[6] S. Nousias, C. Tselios, O. Orfila, S. Jamson, P. Mejuto, D. Amaxilatis, O. Akrivopoulos, I. Chatzigiannakis, A. S. Lalos, K. Moustakas *et al.*, “Managing nonuniformities and uncertainties in vehicle-oriented sensor data over next generation networks,” in *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 2018, pp. 272–277.

[7] E. J. Candès and B. Recht, “Exact matrix completion via convex optimization,” *Foundations of Computational mathematics*, vol. 9, no. 6, p. 717, 2009.

[8] J.-F. Cai, E. J. Candès, and Z. Shen, “A singular value thresholding algorithm for matrix completion,” *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[9] S. Nousias, C. Tselios, D. Bitzas, A. S. Lalos, K. Moustakas, and I. Chatzigiannakis, “Uncertainty management for wearable iot wristband sensors using laplacian-based matrix completion,” in *2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*. IEEE, 2018, pp. 1–6.

[10] V. Kalofolias, X. Bresson, M. Bronstein, and P. Vandergheynst, “Matrix completion on graphs,” *arXiv preprint arXiv:1408.1717*, 2014.

[11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[12] L. Beretta and A. Santaniello, “Nearest neighbor imputation algorithms: a critical evaluation,” *BMC medical informatics and decision making*, vol. 16, no. 3, p. 74, 2016.

[13] D. J. Stekhoven and P. Bühlmann, “Missforest—non-parametric missing value imputation for mixed-type data,” *Bioinformatics*, vol. 28, no. 1, pp. 112–118, 2011.

[14] S. Oba, M.-a. Sato, I. Takemasa, M. Monden, K.-i. Matsubara, and S. Ishii, “A bayesian missing value estimation method for gene expression profile data,” *Bioinformatics*, vol. 19, no. 16, pp. 2088–2096, 2003.