Novel Techniques for the Estimation of Multi-modal Missing Data in Wireless Sensor Networks

Sofia Savvaki

Thesis submitted in partial fulfilment of the requirements for the

Masters' of Science degree in Computer Science

University of Crete
School of Sciences and Engineering
Computer Science Department
University Campus, Voutes, Heraklion, GR-70013, Greece

Thesis Advisor: Professor *Panagiotis Tsakalides*

Heraklion, November 2016

This work has been performed at the **Foundation for Research and Technology–Hellas, Institute of Computer Science (FORTH–ICS), N. Plastira 100 Vassilika Vouton, GR-700 13 Heraklion, Crete, Greece.**

This work was partially supported by the **DEDALE project, within the H2020 Framework Program of the European Commission** under contract number **665044.**

University of Crete
Computer Science Department

Novel Techniques for the Estimation of Multi-modal Missing Data in Wireless Sensor Networks

Thesis submitted by
Sofia Savvaki
in partial fulfilment of the requirements for the
Masters' of Science degree in Computer Science

THESIS APPROVAL

Author: _____
Sofia Savvaki

Committee approvals: _____
Panagiotis Tsakalides
Professor, Thesis Supervisor

_____
Athanasios Mouchtaris
Associate Professor, Committee Member

_____
Maria Papadopouli
Associate Professor, Committee Member

Departmental approval: _____
Antonis Argyros
Professor, Director of Graduate Studies

Heraklion, November 2016

Abstract

Over the last decades Wireless Sensor Networks (WSNs) have attracted great attention, as they constitute a key enabling technology for implementing sophisticated services in numerous application domains, including area and environmental sensing, health care monitoring, and industrial control systems. Despite their wide applicability, WSNs suffer from network and energy imperfections, which inevitably often lead to missing measurements. The resulting low volume of available data dramatically affects subsequent processing and learning tasks, such as detection of unusual events, clustering, and classification.

In this thesis, we address the problem of missing WSN data by employing two non-conventional techniques, which are capable of recovering measurements in a reliable fashion, namely: a) Matrix Completion (MC), and b) Tensor Completion (TC). The key theoretical principle adopted is that a complex signal can be recovered from a small number of random measurements, by exploiting the underlying redundancies of the sensing data. However, this assumption is not satisfied in real-life, and often, noisy datasets, which tend to be full rank. We tackle this limitation by introducing the concept of appropriately forming the available data streams into low-rank 2D and 3D structures, thereby enabling the utilization of MC and TC in the WSN domain.

To test the efficacy of our approach, we experiment on two prominent fields, namely WSN-based Smart Water Management (SWM) and Human Activity Recognition (HAR). We synthesize their respective processing and classification frameworks, which encapsulate our proposed modules for data sampling, structuring, and recovery. These frameworks are evaluated against numerous aspects, related to the quality of reconstruction on different volumes of missing data, the accuracy of subsequent analysis (e.g. classification), and the impact of sub-sampling on the network's lifetime. Our analysis highlights the interaction of different recovery scenarios in terms of data structuring and origin, with several state-of-the-art classifiers. The results demonstrate that high reconstruction accuracy can be achieved through the developed modules, even for the case of extremely under-sampled, multi-modal streams of data, lacking up to 80% of their measurements.

Περίληψη

Κατά τη διάρκεια των τελευταίων δεκαετιών, τα ασύρματα δίκτυα αισθητήρων έχουν προσελκύσει μεγάλο ενδιαφέρον καθώς αποτελούν τη βασική τεχνολογία που επιτρέπει την υλοποίηση εξελιγμένων υπηρεσιών σε πολλαπλούς τομείς εφαρμογών, οι οποίοι περιλαμβάνουν την παρακολούθηση του περιβάλλοντος, ενός βιομηχανικού συστήματος και της ατομικής υγείας. Όμως, παρά την ευρεία εφαρμογή τους, τα ασύρματα δίκτυα αισθητήρων πάσχουν από δικτυακές και ενεργειακές ατέλειες, που αναπόφευκτα οδηγούν σε απώλεια μετρήσεων. Ως αποτέλεσμα, ο χαμηλός όγκος από διαθέσιμα δεδομένα επηρεάζει δραματικά τις επακόλουθες εργασίες επεξεργασίας και μάθησης, όπως η ανίχνευση ασυνήθιστων γεγονότων, η ομαδοποίηση και η ταξινόμηση.

Στην παρούσα διπλωματική εργασία αντιμετωπίζουμε το πρόβλημα αυτό, προτείνοντας τις μη συμβατικές μεθόδους της Συμπλήρωσης α) Πινάκων (ΣΠ) και β) Τανυστών (ΣΤ), οι οποίες είναι σε θέση να ανακτούν τις απούσες μετρήσεις με έναν αξιόπιστο τρόπο. Η βασική θεωρητική αρχή που υιοθετείται σε αυτές τις τεχνικές, είναι ότι ένα σύνθετο σήμα μπορεί να ανακτηθεί από ένα μικρό αριθμό τυχαίων μετρήσεων, αξιοποιώντας τις υποκείμενες συσχετίσεις ανάμεσα στα δεδομένα. Ωστόσο, η υπόθεση αυτή δεν ικανοποιείται σε πραγματικά, και συχνά θορυβώδη, σύνολα δεδομένων, τα οποία τείνουν να είναι μεγάλης τάξεως. Αντιμετωπίζουμε αυτόν τον περιορισμό, εισάγοντας την ιδέα της κατάλληλης διαμόρφωσης των διαθέσιμων ροών δεδομένων σε δισδιάστατες και τρισδιάστατες δομές χαμηλής τάξης, η οποία επιτρέπει την χρήση των ΣΠ και ΣΤ στον τομέα των ασύρματων δικτύων αισθητήρων.

Για να δοκιμάσουμε την αποτελεσματικότητα της προσέγγισής μας, πειραματιζόμαστε σε δύο εξέχοντες τομείς και συγκεκριμένα στην έξυπνη διαχείριση του νερού μέσω των ασύρματων δικτύων αισθητήρων και στην αναγνώριση της ανθρώπινης δραστηριότητας. Συνθέτουμε τα αντίστοιχα συστήματα επεξεργασίας και ταξινόμησης, τα οποία ενσωματώνουν τις προτεινόμενες μεθόδους μας για δειγματοληψία, δόμηση και ανάκτηση των δεδομένων. Τα εν λόγω συστήματα αξιολογούνται κατά διάφορες πτυχές, που σχετίζονται με την ποιότητα της ανακατασκευής για ποικίλα ποσοστά ελλειπουσών τιμών, την ακρίβεια της μετέπειτα ανάλυσης (π.χ. ταξινόμηση) και την επίδραση της υπό-δειγματοληψίας στη διάρκεια ζωής του δικτύου. Η ανάλυσή μας τονίζει την αλληλεπίδραση μεταξύ διαφόρων σεναρίων ανάκαμψης, από την οπτική της δόμησης και της προέλευσης των δεδομένων, και πολλαπλών σύγχρονων ταξινομητών. Τα αποτελέσματα επιδεικνύουν ότι, οι μέθοδοι που έχουν αναπτυχθεί επιτυγχάνουν υψηλή ακρίβεια ανακατασκευής, ακόμη και για περιπτώσεις εξαιρετικής υπο-δειγματοληψίας πολυτροπικών ροών δεδομένων, στις οποίες έχει χαθεί μέχρι και το 80% των μετρήσεων.

Acknowledgements

First of all, I would like express me sincere gratitude to my supervisor, Professor Panagiotis Tsakalides, for his continuous support and for showing belief in me from my senior under-graduate years and throughout my MSc studies. Prof thank you very much for giving me the opportunity to work with you and for your valuable guidance, advising and motivating up until today. You are my respected role-model and simultaneously the brightest and coolest person I know.

I am also very grateful to my co-advisors, Dr. Grigorios Tsagkatakis and Dr. Athanasia Panousopoulou, for their constant encouragement, their constructive ideas and the feedback they provided to me in all our productive meetings. This work would not have been completed without their valuable help, patience and immense knowledge. Greg, Nansy, I was very lucky to work with you.

Special thanks also go to the members of my dissertation committee, Associate Professors Athanasios Mouchtaris and Maria Papadopouli for their insightful comments and questions during my MSc studies.

I would like to acknowledge the Institute of Computer Science (FORTH-ICS) for providing financial support and all the necessary equipment during this work.

I would also like to thank all my colleagues and friends at the Telecommunications and Network Lab and the Signal Processing Lab for their friendship and support during these years. My warmest thanks to Konstantina, Katerina, Maria, and Tasos for all the helpful discussions, the encouragement during good and bad times, and the nice atmosphere. Guys I had a great time having all of you around!

A dedicated thank you, accompanied by a huge hug, to my dearest friends Niki and Claire for always being there, listening, smiling and encouraging.

Special thanks to the closest person in my life, Charis, for his continuous support, patience and advice.

Last, but definitely not least, I would like to thank my family for supporting me by all means throughout my whole life. Βασίλη, μαμά, μπαμπά σας ευχαριστώ και σας αγαπώ πολύ.

To my family

Στην οικογένειά μου

Contents

Abstract **iii**

List of tables **xv**

List of figures **xvii**

List of abbreviations **xxi**

1 Introduction **1**

1.1 The emergence of Wireless Sensor Networks 1

1.1.1 WSNs for Smart Water Management (SWM) 1

1.1.2 WSNs for Human Activity Recognition (HAR) 2

1.2 The challenge of missing measurements in WSNs 3

1.2.1 High temporal sampling rates 3

1.2.2 Low temporal sampling rates 4

1.3 Motivation and Objectives 5

1.4 Contribution 5

1.5 Related publications 6

1.6 Roadmap 6

2 Matrix & Tensor Completion **9**

2.1 Notation and preliminaries 9

2.2 Problem formulation for Matrix Completion - MC 11

2.2.1 State-of-the-art in Matrix Completion algorithms and our employed approach 12

2.2.1.1 Overview of the method of Augmented Lagrange Multipliers . . . 14

2.3 Problem formulation for Tensor Completion 14

2.3.1 State-of-the-art in Tensor Completion algorithms and our employed approach 15

3 A novel modular system for missing data completion in WSN-based applications **17**

3.1 The sampling module 17

3.1.1 Approach A: Introducing missing values via sub-sampling 17

3.1.2 Approach B: Introducing missing values via temporal super-resolution . . . 18
3.2 The Hankelization module . . . 19
3.2.1 From data streams to matrices . . . 20
3.2.2 Generalization: From matrices to tensors . . . 20
3.3 The recovery module . . . 21

4 Case study 1: Matrix Completion for the recovery of water treatment data colected by a WSN 25
4.1 The proposed framework for SWM . . . 25
4.1.1 The data collection module . . . 26
4.1.2 The sampling module . . . 26
4.1.3 The recovery module . . . 27
4.1.4 The evaluation module . . . 28
4.2 Experimental evaluation of the proposed framework . . . 28
4.2.1 Dataset . . . 28
4.2.2 Effects of measurement matrix dimensions . . . 28
4.2.3 MC for temporal super-resolution . . . 30
4.2.4 Impact of the number of devices . . . 32

5 Case Study 2: Matrix and tensor completion for recovering multi-modal HAR data 35
5.1 The proposed framework for HAR . . . 35
5.1.1 The partitioning module . . . 36
5.1.2 The Hankelization module . . . 36
5.1.3 The recovery module . . . 37
5.1.4 The feature extraction & classification modules . . . 37
5.1.4.1 Utilized classifiers . . . 38
5.2 Experimental evaluation . . . 38
5.2.1 Datasets . . . 39
5.2.2 Effects of measurement matrix dimensions on MC-based NMSE . . . 40
5.2.3 Sufficient training time per dataset . . . 41
5.2.4 Comparing Different Recovery Techniques . . . 42
5.2.5 Running Time per Recovery Scenario . . . 43
5.2.6 Impact of MC & TC on Classification Accuracy . . . 45
5.2.7 Different Sampling per Device . . . 50
5.2.8 Effects of Sub-sampling on Battery Consumption of Shimmer Sensing Platforms . . . 51

6 Conclusions and future work **53**

6.1 Concluding remarks 53

6.2 Future directions 54

A Analysis on setting the optimum parameters for the Hankelization module **55**

List of Tables

4.1 Case study 1 - Dimensionality and fill ratio for different sampling rates. 32

5.1 Comparison of the employed datasets for Human Activity Recognition. 39
5.2 Case study 2 - Comparison of classification accuracy of the proposed classification framework for non-recovered vs. MC-recovered (Scenario 3) data. 49
5.3 Case study 2 - Classification accuracy of Gaussian SVM considering different sampling per device for MC/TC-recovered data. 51
5.4 Case study 2 - Shimmer3 battery status after 9 hrs of use. 52

List of Figures

1.1 Example of missing values due to temporal super-resolution in a WSN. At the top there is the fully-populated measurements matrix, whereas at the bottom we have the undersampled super-resolved matrix, lacking 50% of its measurements. 4

2.1 Fibers (top) and slices (bottom) of a third-order tensor. 9

3.1 Example of artificial sub-sampling in a network of I_1 sensor nodes and n 3-axial modalities of I_2 samples, where zero-placement is applied at the same temporal instances per modality. 18
3.2 Example of temporal super-resolution via doubling the sampling rate in a network of I_1 sensor nodes and n 3-axial modalities acquired at I_2 time intervals. 19
3.3 The Hankelization process. The available fully-populated data streams are structured into low-rank sub-sampled Hankel matrices. 20
3.4 Extension of the Hankelization process from matrices to third-order tensors. 21
3.5 Instance of the 2D recovery module for Scenario 2, regarding the 3-axial modality 1 in a test-bed of 2 sensing devices. 22
3.6 Instance of the 3D recovery module for Scenario 2, regarding the 3-axial modality 1 in a test-bed of 2 sensing devices. 23

4.1 Case study 1 - The proposed framework for water treatment data. 26
4.2 Case study 1 - The desalination plant (left) and the WSN module used for the collection of water desalination data (right). 26
4.3 Case study 1 - The proposed structuring of data (general case). 27
4.4 Case study 1 - The proposed structuring of data, based upon the water dataset at hand. 28
4.5 Case study 1 - Normalized MSE as a function of f with respect to 4 different measurements matrix sizes. 29
4.6 Case study 1 - (a) The original fully-sampled measurements matrix and the MC-recovery results: (b) reconstructed matrix from 10% of the measurements, (c) reconstructed matrix from 50% of the measurements, (d) reconstructed matrix from 90% of the measurements. 30

4.7 Case study 1 - The initial (left) and reconstructed (right) when the dimensions of the measurements matrix are (a) [50] × [72], (b) [50] × [144], (c) [50] × [288], (d) [50] × [432]. 31
4.8 Case study 1 - NMSE w.r.t. f for single-device vs. collective recovery regrading the devices 1 and 2. 33
4.9 Case study 1 - NMSE w.r.t. f for single-device vs. collective recovery regrading the devices 1 and 5. 33

5.1 Case study 2 - The proposed HAR framework. Reconstruction is incorporated in the testing phase. Subsequently, feature extraction and classification is performed using the predictive models formed on training phase. 36
5.2 Case study 2 - NMSE as a function of f and test data size for the x-axis accelerometer (left) and the x-axis gyroscope (right) of the HAR dataset, considering Scenario 1. 40
5.3 Case study 2 - The magnitude of the singular values of x-axis accelerometer (left) and x-axis gyroscope (right) channels of the HAR dataset. 41
5.4 Case study 2 - HAR (left) and MHEALTH (right) classification accuracy w.r.t. training time for all classifiers on fully-populated test matrices. (Mean of 10 runs.) 42
5.5 Case study 2 - Cumulative NMSE (top) and corresponding running times (bottom) w.r.t. f of all applied reconstruction methods for HAR (left) and MHEALTH (right). (Scenario 1 - Mean of 10 runs.) 43
5.6 Case study 2 - TC (top), MC (middle) and RegEM (bottom) running time for HAR (left) and mHealth (right) w.r.t. f, for Scenario 1,2,3. (Mean of 10 runs.) . . 44
5.7 Case study 2 - Classification accuracy of Decision Trees on HAR (top) and MHEALTH (bottom) w.r.t. to f, considering ground truth, and MC (left) or TC (right) reconstruction for all scenarios. (Mean of 10 runs.) 46
5.8 Case study 2 - Classification accuracy of Euclidean K-NN on HAR (top) and MHEALTH (bottom) w.r.t. to f, considering ground truth, missing (non-recovered) data, and MC (left) or TC (right) reconstruction for all scenarios. (Mean of 10 runs.) 46
5.9 Case study 2 - Classification accuracy of Cosine KNN on HAR (top) and MHEALTH (bottom) w.r.t. to f, considering ground truth, and MC (left) or TC (right) reconstruction for all scenarios. (Mean of 10 runs.) 47
5.10 Case study 2 - Classification accuracy of Gaussian SVM on HAR (top) and MHEALTH (bottom) w.r.t. to f, considering ground truth, and MC (left) or TC (right) reconstruction for all scenarios. (Mean of 10 runs.) 47
5.11 Case study 2 - Classification accuracy of Quadratic SVM on HAR (top) and MHEALTH (bottom) w.r.t. to f, considering ground truth, and MC (left) or TC (right) reconstruction for all scenarios. (Mean of 10 runs.) 48

5.12 Case study 2 - Classification accuracy for MC (left) and TC (right) reconstruction w.r.t. to f, considering Scenario 3 for HAR (top) and MHEALTH (bottom). (Mean of 10 runs.) 50
5.13 Case study 2 - Classification accuracy of Cosine K-NN for TC-Scenario 3 reconstruction w.r.t. the remaining battery capacity of a Shimmer3 platform after 9 hours of use. 52

A.1 NMSE w.r.t. f for window size set to 128 samples and various overlaps, regarding the x(left)/y(middle)/z(right)-axis accelerometers (top) and gyroscopes (bottom) in HAR dataset. 56
A.2 NMSE w.r.t. f for overlap set to 50% of the window size and various sizes of windows, regarding the x(left)/y(middle)/z(right)-axis accelerometers (top) and gyroscopes (bottom) in HAR dataset. 56

List of abbreviations

WSN . . . . . Wireless Sensor Network

IoT . . . . . Internet of Things

SWM . . . . . Smart Water Management

HAR . . . . . Human Activity Recognition

SWN . . . . . Smart Water Network

CPS . . . . . Cyber-Physical System

MC . . . . . Matrix Completion

TC . . . . . Tensor Completion

SVD . . . . . Singular Value Decomposition

ALM . . . . . Augmented Lagrange Multipliers

NMSE . . . . . Normalized Mean Square Error

RegEM . . . . . Regularized Expectation Maximization

K-NN . . . . . K-Nearest Neighbours

SVM . . . . . Support Vector Machine

Chapter 1

Introduction

1.1 The emergence of Wireless Sensor Networks

Wireless Sensor Networks (WSNs) constitute an emerging technology that bridges the physical and the digital information worlds under the ever-growing vision of the Internet of Things (IoT). Intelligent data monitoring and management can be achieved through the use of networked embedded devices, called sensor nodes, transmitting useful measurements and control instructions via distributed WSNs [1–3].

A WSN is composed by a large number of nodes, each one equipped with sensors that monitor physical or environmental conditions, in one or multiple sensing modalities, i.e. types of measurements, such as temperature, humidity, sound, motion, and so on. The network comprises homogeneous or heterogeneous sensor nodes, sampling the observed field at various rates depending on their respective modalities, the configuration of the network, and the remaining lifetime of each node.

The first developments of WSNs were motivated by military applications, notably surveillance in conflict zones. Today, the variety of possible applications of WSNs to the real world is practically unlimited. Application fields include industrial infrastructure monitoring, home automation, healthcare, and traffic control. In this thesis, we focus on two prominent WSN-related fields, namely Smart Water Management (SWM) and Human Activity Recognition (HAR). In the following subsections, we provide a brief discussion related to the two application domains considered herein along with our motivation for engaging with the particular research areas.

1.1.1 WSNs for Smart Water Management (SWM)

Smart Water Networks (SWNs) [4] are on a fast rise over the last decade, since they constitute an emerging engineering field which addresses the interconnection between data technologies and water infrastructures, with the objective to deliver sustainable solutions related to water resource utilization. Driven by the application demands of modernizing water quality monitoring, acting upon alerting events, and improving our awareness of water allocation and consumption, Cyber-

Physical Systems (CPS) and Wireless Sensor Networks (WSNs) encapsulate the key-enabling technologies for the next generation of SWN systems [5, 6].

Existing solutions in the arena of SWN are considered extremely useful both for responding to alarming situations [7, 8], as well as for engaging citizens in becoming part of water sustainable policies [9, 10]. In their majority, they are deployed close to urban areas thereby yielding feedback on the actual quality of the water that flows within the pipes towards the plumbing facilities of the individuals [11–14]. Despite their significance, such solutions cannot address water purification for human consumption, which is becoming a critical aspect of water management, especially as the natural resources of fresh water become scarcer and the population of urban areas grows. Consequently, the necessity intensifies for expanding the engineering focus towards monitoring and control of industrial treatment plants, which are responsible for water purification for human consumption.

Water purification involves the combination of slowly varying physical and chemical processes, for making untreated water suitable for human consumption. Such industrial processes are responsible for monitoring and controlling critical microbiological parameters, based on sparse and sporadic samples, and they rely on off-line testing procedures and the involvement of experienced personnel in the control loop. It has been recognized that employing WSNs in water treatment plants would improve their autonomous character, by introducing novel paradigms of data acquisition and processing.

1.1.2 WSNs for Human Activity Recognition (HAR)

Sensor-based Human Activity Recognition integrates the emerging area of wireless sensing with novel data mining and machine learning techniques to model a wide range of human activities [15]. HAR is a rapidly expanding research area that attracts more and more attention in recent years, since it holds a vital role in the monitoring and enhancement of human health and well-being status [16], [17]. Extracting contextual information from sensor-acquired physical data is therefore encountered in numerous health-related applications, such as elder care support, rehabilitation assistance, chronic conditions management, and fitness coaching, just to name a few [18–20]. For instance, patients with diabetes, obesity, or heart disease are often required to follow a well defined exercise routine as part of their treatments. Thus, recognizing activities such as walking, running, or cycling becomes quite useful for providing feedback to the caregiver about the patient's behaviour. Likewise, patients with dementia and other mental pathologies could be monitored for the detection of abnormal activities, and, thereby, the prevention of undesirable consequences.

Currently there are various sensor-based approaches employed for monitoring the human activity. Recent technological advances in smartphones and their ever-growing daily use by the general public, has transformed them into an ideal first-hand tool for non-intrusive sensing. Latest devices come with embedded built-in sensors such as microphones, dual cameras,

accelerometers, gyroscopes, etc. This large set of sensing capabilities combined with their computational competence and their mobile nature, consolidates smart-phones as a promising solution for HAR [16, 21–23].

Another broadly employed approach for collecting activity information is the utilization of wearable platforms, for instance SHIMMER sensing devices [24], directly attached to the body [25]. Such platforms have demonstrated their ability to monitor a variety of multiple attributes, related to the user's movement (e.g. using accelerometers or gyroscopes), environmental variables (e.g. temperature and humidity), or physiological signals (e.g. heart rate or electrocardiogram). A platform approach, where the sensing capability can be modified via physical and software device configuration, addresses the need for heterogeneous sensing capabilities, while minimizing the complexity of the hardware and software development, validation, and support.

It is often the case that the measurements recorded using the above mentioned approaches are merged to those obtained by a number of other "smart" devices including sensors, such as smart watches or even smart shoes. Combining sensor readings increases the overall quality of information beyond the sum of the parts [24]. Hence, a typical activity recognition infrastructure can be comprised of noumerous heterogeneous sensig devices, attached to several body parts, acquiring different sensing modalities at various sampling rates, depending on the application demands [26] [27]. Subsequently, the collected data are naturally indexed over the time dimension, and processed by supervised machine learning algorithms for the detection of the underlying activity over each window of time.

1.2 The challenge of missing measurements in WSNs

The evolution of Wireless Sensor Networks and the ever-increasing demands of their applications have led to a number of limitations regarding the acquisition of sensor-based measurement datasets. Limited lifetime, communication failures, memory and energy constraints, sensor desynchronization and portability, constitute only a subset of the existing factors leading to unobserved or lost measurements in a typical WSN [28–30]. From a pragmatic point of view, such under-sampled datasets hinder the efficient extraction of knowledge from the available data, thus highlighting the need for the reconstruction of the unobserved data in the development of any efficient WSN-based application [31, 32]. In the following, we briefly review these critical aspects, which are strongly related to the temporal sampling rate used for data acquisition on the underlying WSN.

1.2.1 High temporal sampling rates

For WSNs operating in high data sampling rate regimes, although frequent sampling offers a high-quality monitoring of the underlying processes, it may also have a dramatic effect on the lifetime of the network. This effect is attributed to the close relationship between measurement

acquisition and energy consumption. Waking-up a node, acquiring a measurement, performing quantization, and storing the data in local memory, are all extremely energy-demanding tasks. If communications with other nodes is also necessary due to, *e.g.*, storage requirements, then the impact on network lifetime is even more pronounced.

Another case that entails a large number of missing values in high data rate WSNs, concerns measurements encoded in packets lost due to communications failures. This scenario occurs typically in industrial environments, where heavy machinery has a detrimental impact on link quality. Furthermore, in multi-hop networks, congestion and duty cycling can also lead to dropped packets and, thus, lost measurements.

1.2.2 Low temporal sampling rates

A third scenario, is related to a low-frequency temporal sampling WSN. Either by design, or due to clock de-synchronization, each sensor may end up sampling the underlying field at a different time instance. As an example, consider the paradigm illustrated in Figure 3.1.2. Let us assume that we have a network monitoring a field, where the entire collection of measurements can be organized into a measurements matrix, where rows correspond to sensors and columns to sampling instances. A fully synchronized network of 2 sensors, configured to inquire and record the field every 10 minutes, would produce a [2 (sensors)] $\times$ [6 (measurements per sensor)] matrix, (cf. Figure 3.1.2, top), with columns corresponding to measurements acquired at time intervals 00:10, 10:20, 20:30,... 50:60. One sensor could sample at 00:03, 00:13,..., 00:53 while the other could sample at 00:07, 00:17, ..., 00:57. However, the overlying application could demand data from the field, at a sampling rate beyond the temporal sampling capabilities of the network, namely every 5 minutes, and specifically at timestamps 00:00, 00:05, 00:10, 00:15, ..., 00:55. This scenario, would result in a 2×12 matrix, whose columns correspond to the timestamps required by the application. Such a matrix will naturally miss 50% of its measurements, due to the requested temporal super-resolution, (cf. Figure 3.1.2, bottom).

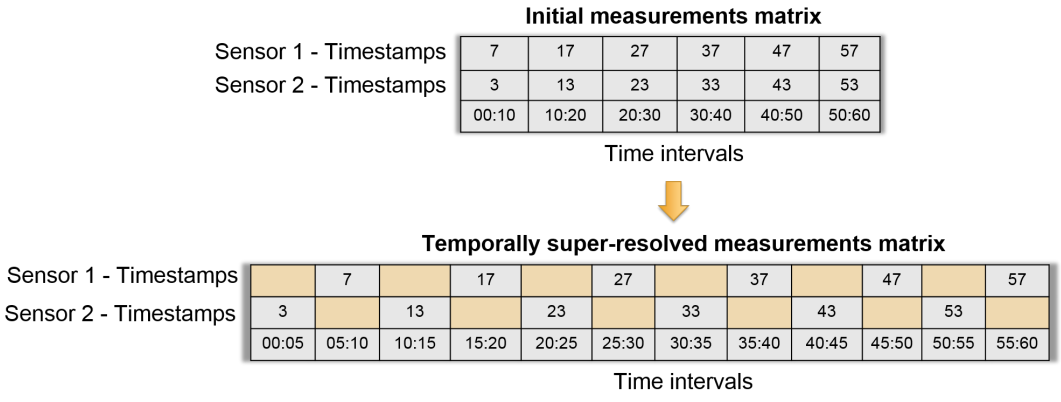

Figure 1.1: Example of missing values due to temporal super-resolution in a WSN. At the top there is the fully-populated measurements matrix, whereas at the bottom we have the undersampled super-resolved matrix, lacking 50% of its measurements.

All of the above discussed scenarios, result in a significant number of missing data that can have a dramatic impact on subsequent tasks, such as detection of unusual events, clustering of the measurements, or data classification. For instance, from the perspective of Smart Water Management, these constraints are translated to the lack of sufficient data samples for characterising different aspects of the water purification process. On the other hand, considering Human Activity Recognition, such under-sampled datasets hinder the performance of the underlying machine learning algorithms.

1.3 Motivation and Objectives

The previously described scenarios regarding the occurrence of unobserved measurements and their negative effects on high-level applications, highlight the fact that data recovery is an indispensable operation in WSNs. These considerations serve as our motivation for the introduction of efficient and robust mechanisms for data recovery in a reliable fashion.

In this thesis, we focus on two recovery techniques, namely Matrix Completion (MC) and Tensor Completion (TC), which have been recognized as two promising novel approaches for addressing the problem of missing values from a signal processing perspective. Their underlying mathematical concept is that a complex signal can be recovered from a small number of random measurements, far below the traditional Nyquist-Shannon limit. The key assumption herein, is that the signal is sparse and that randomly sub-sampled matrix or tensor measurement data are available, therefore making MC and TC appealing for WSN applications.

Our goal is to examine the efficacy of the MC and TC-based approaches on the aforementioned WSN-based domains. In order to do so, we propose novel modules for realistic data sampling, 2D and 3D structuring and recovery, and we encapsulate them in two application-specific frameworks. The evaluation process relies on real-valued datasets.

1.4 Contribution

This thesis focuses on the accurate estimation of missing measurements on two highly distinct fields of WSNs and it provides useful insights regarding the efficiency of our proposed methods for both applications under scope.

More specifically, concerning the Smart Water Management study, we recommend the formulation of the problem of the unobserved water treatment data as an instance of low rank Matrix Completion and we propose a novel framework for the evaluation of MC theory. We consider various approaches for the assessment of the system, which are related to the volume of the available data, the number of the sensors providing measurements for recovery, and ultimately the realistic temporal super-resolution aspect that expresses the relationship between the sampling rate and the netwiork operational characteristics. For all of the aforementioned considerations, we have implemented their respective modules and embedded them to our proposed framework.

Concerning the Human Activity Recognition study, we formulate it as a classification problem and we propose a modular and scalable classification framework for the assessment of the overall recognition process in the presence of artificially introduced and subsequently recovered missing values. Moreover, we provide a direct comparison of the proposed techniques for Matrix and Tensor Completion to another sophisticated method, that of the regularized expectation maximization. We review the interaction of recovery with a variety of evaluated classifiers, belonging to major families of Machine Learning algorithms, namely Decision Trees, the instance-based method of K-Nearest Neighbours, and kernel-based SVM classifiers. Furthermore, we introduce a novel Hankelization process for constructing low-rank matrices and tensors from data streams and three realistic scenarios for collective data recovery. We also reach useful conclusions on the performance of the proposed framework under the pragmatic conditions of non-uniform occurrence of missing values per device and examine the impact of sub-sampling on the lifetime of the underlying sensor network.

1.5 Related publications

Our proposed methodology, the associated evaluation studies, and the experimental findings with real data have been summarized in the following three original publications, which have resulted form this thesis.

1. S. Savvaki, G. Tsagkatakis, A. Panousopoulou, and P. Tsakalides, "*Recovering Multimodal Physical Data: Matrix & Tensor Completion on a Classification Framework*", Journal of Biomedical and Health Informatics (J-BHI), Submitted.

2. S. Savvaki, G. Tsagkatakis, A. Panousopoulou, and P. Tsakalides, "*Effects of Matrix Completion on the Classification of Undersampled Human Activity Data Streams*", in Proc. 24th European Signal Processing Conference (EUSIPCO 2016), Budapest, Hungary, August 29 - September 2, 2016.

3. S. Savvaki, G. Tsagkatakis, A. Panousopoulou, and P. Tsakalides "*Application of Matrix Completion on Water Treatment Data*", in Proc. 1th International Workshop on Cyber-Physical Systems for Smart Water Networks (CySWater 2015), CPS Week 2015, Seattle, WA, USA, April 13-16, 2015.

1.6 Roadmap

The remainder of this thesis is organized as follows: In the next chapter, we present the necessary preliminaries on the underlying theory of Matrix and Tensor Completion and we introduce our approach for solving the corresponding problems. In Chapter 3, we describe our methodology regarding the modules of data sampling, structuring, and recovery within a WSN framework,

while in Chapters 4 and 5 we validate the efficiency of our concepts by applying them on two highly distinctive WSN domains. More specifically, in Chapter 4, we incorporate our modules within a proposed Smart Water Management framework for the reconstruction of WSN-based water desalination data. In Chapter 5, we extend our studies in the Human Activity Recognition field, experimenting on physical-kinetic human data and we extensively discuss our conducted experiments and derived results. Final remarks are presented in Chapter 6 along with directions for future work.

Chapter 2

Matrix & Tensor Completion

In this chapter, our objective is to present the fundamentals on the underlying mathematical theory of the two main techniques utilized in this thesis for missing data recovery, namely Matrix and Tensor Completion (MC and TC, respectively). First, we focus on basic tensor operations and then we proceed with the mathematical formulation of the MC and TC problems. Alongside, we briefly review some state-of-the-art solvers for completing matricew and tensors, and lastly we describe the specific algorithms that we have employed for MC and TC.

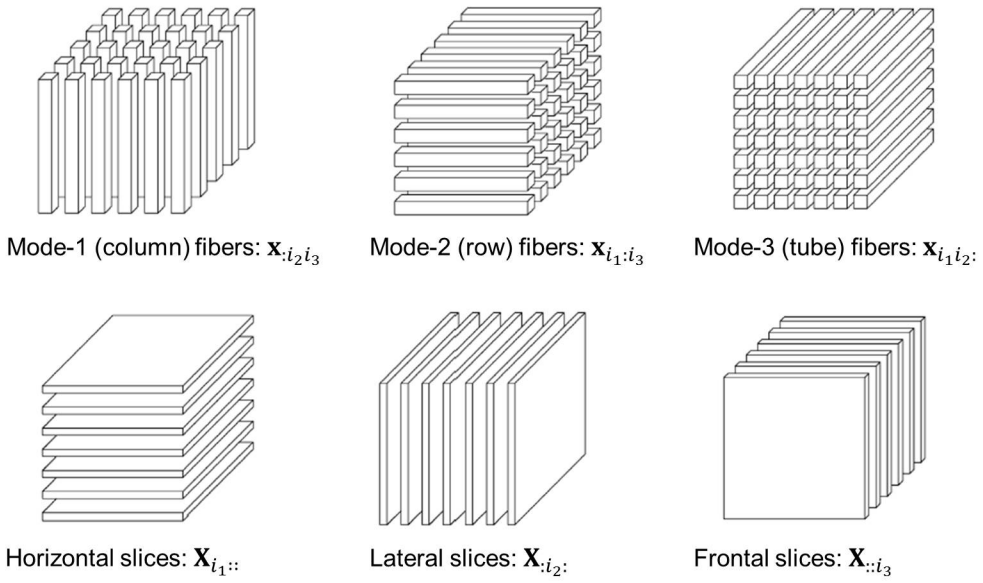

Figure 2.1: Fibers (top) and slices (bottom) of a third-order tensor.

2.1 Notation and preliminaries

In this work, following [33], we use bold lower-case letters $\mathbf{x}$, $\mathbf{y}$, ... for vectors, bold upper-case letters $\mathbf{X}$, $\mathbf{Y}$, ... for matrices, and bold calligraphic letters $\boldsymbol{\mathcal{X}}$, $\boldsymbol{\mathcal{Y}}$, ... for tensors.

A tensor is a generalization of a vector and a matrix. A vector is a first-order (also called one-way or one-mode) tensor and a matrix is a second-order tensor. An N-order tensor is defined as $\boldsymbol{\mathcal{X}} \in \Re^{I_1 \times I_2 \times \cdots \times I_N}$ and its $(i_1, i_2, \ldots, i_N)$-th component is denoted as $x_{i_1,i_2,\ldots,i_N}$, where $1 \leq i_k \leq I_k$ and $1 \leq k \leq N$.

A *fiber* of $\boldsymbol{\mathcal{X}}$ is a vector $\mathbf{x}$ obtained by fixing all indices of $\boldsymbol{\mathcal{X}}$ except one, while a *slice* of $\boldsymbol{\mathcal{X}}$ is a matrix $\mathbf{X}$ acquired by fixing all indices of $\boldsymbol{\mathcal{X}}$ except two. This is illustrated in Figure 2.1, which shows the fibers (at the top) and the slices (at the bottom) of a third-order tensor $\boldsymbol{\mathcal{X}} \in \Re^{I_1 \times I_2 \times I_3}$.

It is often very convenient to represent a tensor as a matrix. *Unfolding*, also known as matricization or flattening, is a process of reordering the elements of an N-order tensor into a matrix. The unfolding operation along the n-th mode on a tensor $\boldsymbol{\mathcal{X}} \in \Re^{I_1 \times I_2 \times \cdots \times I_N}$ is denoted as $\mathbf{X}_{(n)} \in \Re^{I_n \times \Pi_{j \neq n} I_j}$, which is a matrix whose columns are the mode-n fibers of $\boldsymbol{\mathcal{X}}$. Notice that, the choice of the ordering of the columns of $\mathbf{X}_{(n)}$ does not matter for practical purposes. It is enough that one sticks to the same rule to arrange the n-mode vectors as fibers of the nth mode unfolding. As a simple example, the modal unfoldings for $\boldsymbol{\mathcal{X}} \in \Re^{3 \times 4 \times 2}$ are shown below, where $\mathbf{X}(:,:,1)$ and $\mathbf{X}(:,:,2)$ are the frontal slices of $\boldsymbol{\mathcal{X}}$.

$$\mathbf{X}(:,:,1) = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{bmatrix}, \qquad \mathbf{X}(:,:,1) = \begin{bmatrix} 13 & 14 & 15 & 16 \\ 17 & 18 & 19 & 20 \\ 21 & 22 & 23 & 24 \end{bmatrix}$$

$$\mathbf{X}_{(1)} = \begin{bmatrix} 1 & 2 & 3 & 4 & 13 & 14 & 15 & 16 \\ 5 & 6 & 7 & 8 & 17 & 18 & 19 & 20 \\ 9 & 10 & 11 & 12 & 21 & 22 & 23 & 24 \end{bmatrix}$$

$$\mathbf{X}_{(2)} = \begin{bmatrix} 1 & 5 & 9 & 13 & 17 & 21 \\ 2 & 6 & 10 & 14 & 18 & 22 \\ 3 & 7 & 11 & 15 & 19 & 23 \\ 4 & 8 & 12 & 16 & 20 & 24 \end{bmatrix}$$

$$\mathbf{X}_{(3)} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 & 17 & 18 & 19 & 20 & 21 & 22 & 23 & 24 \end{bmatrix}$$

The n-rank of a N-way tensor $\boldsymbol{\mathcal{X}}$, denoted as $\text{rank}_n(\boldsymbol{\mathcal{X}})$, is a generalization of the matrix rank. The rank is an indicator of the correlations existing among the underlying data, meaning that the lower the rank, the higher the data correlations. The rank of a matrix $\mathbf{X} \in \Re^{I_1 \times I_2}$ is defined as the size of the largest collection of its linearly independent columns (the column rank), or its linearly independent rows (the row rank). It is a non-negative integer, upper bounded by either I_1 or I_2, thus rank $(\mathbf{X}) \leq min(I_1, I_2)$. A matrix having a rank that is as large as possible is said to be full-rank. Similarly, the rank of a tensor $\boldsymbol{\mathcal{X}}$ is defined as an array: (rank($\mathbf{X}_{(1)}$), ...,

rank($\mathbf{X}_{(N)}$)), containing the ranks of all node unfoldings $\mathbf{X}_{(n)}$ of the tensor. We say that $\boldsymbol{\mathcal{X}}$ is approximately low-rank if $\mathbf{X}_{(n)}$ is approximately low-rank for all n.

2.2 Problem formulation for Matrix Completion - MC

Let us consider a sensor network consisting of I_1 nodes. Each node samples at a fixed rate and forwards the data to a sink through a multi-hop way. Thus, at the sink we should get an $[I_1] \times [I_2]$ matrix, where I_2 is the number of samples obtained in each node, i.e., $\mathbf{X} \in \Re^{I_1 \times I_2}$. However, due to lossy transmissions, the failure of sensor nodes, and other factors previously described in Section 1.2, some data are missing and the matrix is incomplete. Hence, we end up with a partially observed $[I_1] \times [I_2]$ matrix $\mathbf{M}$ whose missing entries we wish to recover using only the available $k << I_1 \times I_2$ measurements. In general, this is an ill-posed problem, unless some additional constraints are imposed on $\mathbf{M}$. Specifically, it was recently proved [34], that exact recovery is feasible from most sets of k sampled entries, even of surprisingly small cardinality, given that $\mathbf{M}$ is low-rank.

Formally, let Ω be the set of known indices (i_1, i_2) corresponding to the available measurements. The linear map $\mathcal{A}$ is defined as a projection operator setting all unknown indices to zero, that is

$$\mathcal{A}(\mathbf{M}_{i_1 i_2}) = \begin{cases} 1, & \text{if } (i_1 i_2) \in \Omega \\ 0, & \text{otherwise} \end{cases}$$

A natural way to fill in the missing values is to estimate the lowest-rank matrix $\mathbf{X}$ which agrees with the given data in $\mathbf{M}$ [35], by solving:

$$\begin{aligned} &\underset{\mathbf{X}}{\text{minimize}}\ rank(\mathbf{X}) \\ &\text{subject to}\ \ \mathcal{A}(\mathbf{X}_{i_1 i_2}) = \mathcal{A}(\mathbf{M}_{i_1 i_2}), \quad \forall (i_1 i_2) \in \Omega\ . \end{aligned} \tag{2.1}$$

Although one could seek an approximate matrix $\mathbf{X}$ by minimizing the rank [35], rank minimization is an NP-hard problem in general [36]. However, it was recently shown that exact matrix recovery is possible through convex optimization [34], [37]. The relaxation of the above problem that was shown to produce accurate approximations, is based on the replacement of the rank function with the more computationally tractable nuclear norm, which represents the convex envelope of the rank. Singular Value Decomposition (SVD) decomposes the $[I_1] \times [I_2]$ measurements matrix, into a product of an orthonormal matrix $\mathbf{U}$, a diagonal matrix $\mathbf{S}$, and another orthonormal matrix $\mathbf{V}$, such that:

$$\mathbf{M} = \mathbf{U}\mathbf{S}\mathbf{V}^T. \tag{2.2}$$

According to the spectral theorem associated to the SVD, the number of singular values, *i.e.*, the

diagonal entries of $\mathbf{S}$, reveals the rank of the matrix. Low rank matrices, such as the ones produced by spatio-temporally correlated processes, are characterized by a small number of singular values. Furthermore, the rank of a measurement matrix might be artificially increased, due to noise that typically follows an independent distribution. Hence, considering a lower-rank approximation of the matrix results in an implicit de-noising of the sampled data. Based on the SVD analysis of a matrix, the minimization in (2.1) can be reformulated as:

$$\begin{aligned} &\underset{\mathbf{X}}{\text{minimize}} \quad \|\mathbf{X}\|_* \\ &\text{subject to} \quad \mathcal{A}(\mathbf{X}_{i_1 i_2}) = \mathcal{A}(\mathbf{M}_{i_1 i_2}), \quad \forall (i_1 i_2) \in \Omega, \end{aligned} \tag{2.3}$$

where the nuclear norm $\|\mathbf{X}\|_*$ of a matrix $\mathbf{X}$ is defined as the sum of its singular values,

$$\|\mathbf{X}\|_* = \sum \sigma_i(X)$$

and it reveals its rank. Equation (2.3) constitutes a semi-definite, computationally tractable problem [38]. Recovery of the matrix is possible, provided that Ω is sampled uniformly and matrix $\mathbf{M}$ obeys a low coherence condition. Then, with probability $1 - n^{-3}$, the solution of (2.3) will converge to the solution of (2.1), provided that the number of obtained samples obeys

$$k \geq Cn^{6/5} r log(n),$$

where $n = max(I_1, I_2)$, C is an appropriate constant, and r is the matrix rank.

For the noisy case, an approximate version of (2.3) can be solved [39], by replacing the equality constraint with an inequality constraint given by $\|\mathcal{A}(\mathbf{X}_{i_1 i_2}) - \mathcal{A}(\mathbf{M}_{i_1 i_2})\|_F^2 \leq \epsilon$, where

$$\|\mathbf{X}\|_F^2 = \sum \lambda_i^2$$

denotes the Frobenius norm and ϵ is the approximation error. The optimization is therefore formulated as:

$$\begin{aligned} &\underset{\mathbf{X}}{\text{minimize}} \quad \|\mathbf{X}\|_* \\ &\text{subject to} \quad \|\mathcal{A}(\mathbf{X}_{i_1 i_2}) - \mathcal{A}(\mathbf{M}_{i_1 i_2})\|_F^2 \leq \epsilon, \quad \forall (i_1 i_2) \in \Omega. \end{aligned} \tag{2.4}$$

2.2.1 State-of-the-art in Matrix Completion algorithms and our employed approach

Computing the SVD in order to design a standard nuclear norm solver unsurprisingly plays a critical computational role for large matrices. The efficient and accurate solution to this problem has attracted much research attention in recent years. Some of the existing methods for doing so, include but are not limited to:

- The singular value thresholding (SVT) algorithm proposed in [40], that is essentially a gradient method for solving the dual of a regularized approximation of Equation (2.3).
- The Fixed-Point Continuation with Approximate (FPCA) Singular Value Decomposition method in [41], solving a least squares (regularized with the nuclear norm) Lagrangian version of Equation (2.3).
- The Accelerated Proximal Gradient Lagrangian (APGL) method in [42] that solves another Lagrangian version of Equation (2.3).
- The Proximal Point Algorithm (PPA) in [43] that solves the general nuclear norm minimization problem with linear equality and second-order cone constraints.
- Interior-point methods in [35], [34], and [44] for solving the semi-definite programming reformulation of Equation (2.3).

However, most of the aforementioned related efforts involve applying a soft-thresholding operator on the singular values of an iterate, which requires repeated calls to an SVD or truncated SVD solver. Thus, such approaches are not scalable to large-scale problems [40], [41], [42] that typically occur within a WSN.

In the following, we present some previous work on applying the MC theory in WSNs. The authors in [45] suggest a method to recover the lost data in internet traffic matrices by utilizing low-rankness and spatio-temporal correlation. Moreover, [29] proposes an algorithm using the low-rank structure, time stability, space similarity, and multi-attribute correlation to estimate the missing data in highly incomplete data matrices. Authors in [46] present an algorithm that utilizes the low-rankness and short term stability features to reduce data traffic in WSNs.

We depart from these approaches by addressing the issue of high computational complexity presented by state-of-the-art solvers of the MC problem. Hence, in this thesis, we employ the Augmented Lagrange Multipliers (ALM) based MC algorithm proposed in [47], to solve a reformulation of the nuclear norm minimization problem, that is:

$$\begin{aligned} &\underset{\mathbf{X},\mathbf{E}}{\text{minimize}} \quad \|\mathbf{X}\|_* \\ &\text{subject to} \quad \boldsymbol{X} + \boldsymbol{E} = \boldsymbol{M}, \quad \mathcal{A}(\boldsymbol{E}_{i_1 i_2}) = 0, \quad \forall\, (i_1 i_2) \in \Omega, \end{aligned} \tag{2.5}$$

Equation (2.5) is strongly connected to the formulation of the Robust Rrincipal Component Analysis (RPCA) problem, extensively described in [47], which can be solved very efficiently. According to literature, the considered ALM algorithm exhibits high recovery performance and quick convergence [47], [48].

Algorithm 1 General Method of Augmented Lagrange Multiplier

Output: $\mathbf{X}_k$

1: $\rho \geq 1$
2: **while** not converged **do**
3: Solve $\mathbf{X}_{k+1} = arg\ min_{\mathbf{X}}\ \mathcal{L}(\mathbf{X}, \mathbf{Y}_k, \mu_k)$.
4: $\mathbf{Y}_{k+1} = \mathbf{Y}_k + \mu_k h(\mathbf{X}_{k+1})$;
5: $\mathbf{Y}_{k+1} = \mathbf{Y}_k + \mu_k (\mathbf{X}_{k+1})$;
6: Update μ_k to μ_{k+1}.
7: **end while**

2.2.1.1 Overview of the method of Augmented Lagrange Multipliers

In [49], the general method of augmented Lagrange multipliers is introduced for solving constrained optimization problems of the kind:

$$\begin{aligned} &\underset{\mathbf{X}}{\text{minimize}} \quad f(\mathbf{X}) \\ &\text{subject to} \quad h(\mathbf{X}) = 0, \end{aligned} \tag{2.6}$$

where $f : \Re^{I_1} \Rightarrow \Re$ and $h : \Re^{I_1} \Rightarrow \Re^{I_2}$. One may define the augmented Lagrangian function:

$$\mathcal{L}(\mathbf{X}, \mathbf{Y}, \mu) = f(\mathbf{X}) + \langle \mathbf{Y}, h(\mathbf{X}) \rangle + \frac{\mu}{2} \|h(\mathbf{X})\|_F^2,$$

where μ is a positive scalar. Then, the optimization problem can be solved via the method of augmented Lagradge multipliers, outlined in Algorithm 1.

Under some rather general conditions, when μ_k is an increasing sequence and both f and h are continuously differentiable functions, it has been proven in [49] that the Lagrange Multipliers $\mathbf{Y}_k$ produced by Algorithm 1 converges Q-linearly to the optimal solution when μ_k is bounded, and super-Q-linearly when μ_k is unbounded. This superior convergence property of ALM makes it very attractive. Another merit of ALM is that the optimal step size to update $\mathbf{Y}_k$ is proven to be the chosen penalty parameter μ_k, making the parameter tuning much easier than the iterative thresholding algorithm. A third merit of ALM is that the algorithm converges to the exact optimal solution, even without requiring μ_k to approach infinity [49]. In contrast, strictly speaking both the iterative thresholding and APG approaches mentioned earlier can only find approximate solutions to the problem. Finally, the analysis of convergence and the implementation of the ALM-based algorithms is relatively simple.

2.3 Problem formulation for Tensor Completion

Since tensors constitute a generalization of matrices, the theory of Tensor Completion is an extension of the theory of Matrix Completion. Thus, in direct analogy to the formulation of the MC problem, let us assume that the measurements in our considered sensor network are now forwarded at a sink implemented by a third-order tensor structure. At this case, the aforemen-

tioned factors leading to missing measurements within a WSN would result in an under-sampled $[I_1] \times [I_2] \times [I_3]$ tensor $\boldsymbol{\mathcal{T}}$, which we wish to recover from a fraction k of its entries being available.

Equation (2.3) for the matrix case (i.e., the two-order tensor) is extended to higher-order tensors by solving the following optimization problem to estimate the lowest-rank tensor $\boldsymbol{\mathcal{X}}$ which agrees with the given data:

$$\begin{aligned} &\underset{\boldsymbol{\mathcal{X}}}{\text{minimize}} \quad \|\boldsymbol{\mathcal{X}}\|_* \\ &\text{subject to} \quad \mathcal{A}(\boldsymbol{\mathcal{X}}_{i_1 i_2 i_3}) = \mathcal{A}(\boldsymbol{\mathcal{T}}_{i_1 i_2 i_3}), \quad \forall (i_1 i_2 i_3) \in \Omega, \end{aligned} \tag{2.7}$$

where Ω is the index set (i_1, i_2, i_3) of observed entries and the linear map $\mathcal{A}$ is defined as a random projection operator keeping the entries in Ω and zeroing out others; that is

$$\mathcal{A}(\boldsymbol{\mathcal{T}}_{i_1 i_2 i_3}) = \begin{cases} 1, & \text{if } (i_1 i_2 i_3) \in \Omega \\ 0, & \text{otherwise} \end{cases}$$

Nonetheless, the tensor nuclear norm is not defined as the convex envelope of the tensor rank, as in the matrix case. Unlike matrices, computing the rank of a general tensor (mode number $>$ 2) is an NP hard problem [50]. Therefore, there is no explicit expression for the convex envelope of the tensor rank to the best of our knowledge.

However, [51] proposes a convex formulation of Equation (2.7), by defining the tensor nuclear norm as:

$$\|\boldsymbol{\mathcal{X}}\|_* = \sum_{i=1}^{n} \alpha_i \|\boldsymbol{\mathcal{X}}_{(i)}\|_*$$

where α_i's are constants satisfying $\alpha_i \geq 0$ and $\sum_{i=1}^{n} \alpha_i = 1$. Thus, the nuclear norm for a general tensor case is defined in [51] as the convex combination of the nuclear norms of all matrices unfolded along each of its modes. Under this definition, (2.7) can be written as:

$$\begin{aligned} &\underset{\boldsymbol{\mathcal{X}}}{\text{minimize}} \quad \sum_{i=1}^{n} \alpha_i \|\boldsymbol{\mathcal{X}}_{(i)}\|_* \\ &\text{subject to} \quad \mathcal{A}(\boldsymbol{\mathcal{X}}_{i_1 i_2 i_3}) = \mathcal{A}(\boldsymbol{\mathcal{T}}_{i_1 i_2 i_3}), \quad \forall (i_1 i_2 i_3) \in \Omega \end{aligned} \tag{2.8}$$

2.3.1 State-of-the-art in Tensor Completion algorithms and our employed approach

According to literature [52] [51], state-of-the-art methods for low-rank tensor completion involve unfolding the tensor into a matrix and the succeeding application of a matrix nuclear-norm minimization algorithm using Singular Value Decomposition (SVD), such as FPCA [41], APGL [42] and many others. However, as previously reported, this approach can be very slow or not applicable for large-scale problems. Moreover, such methods that treat tensors as matrices

utilize only one mode low-rankness of the underlying tensor and do not exploit all the available correlations that exist in the structure.

To address these issues, we employ the recently proposed approach of Low-rank *Tensor Completion* using Parallel Matrix Factorization [33]. According to this technique, each mode of the tensor is unfolded to a set of matrix factors, which are being updated alternatively by dynamically adjusting their rank estimates, a computationally more efficient practice than SVD.

Formally, tensor $\boldsymbol{\mathcal{T}}$ is unfolded to a set of matrix factors $\mathbf{X}_n$ $\mathbf{Y}_n$, such that $\mathbf{T}_n \approx \mathbf{X}_n\mathbf{Y}_n$, for $n = 1, 2, 3$ denoting the number of dimensions. Introducing one common variable $\mathbf{Z}$ to relate these matrix factorizations, we solve the following problem to recover the low-rank tensor $\boldsymbol{\mathcal{X}}$ that agrees with the given data:

$$\begin{aligned} &\underset{\mathbf{X},\mathbf{Y},\mathbf{Z}}{\text{minimize}} \quad \sum_{n=1}^{3} \frac{\alpha_n}{2} \|\mathbf{X}_n\mathbf{Y}_n - \mathbf{Z}_n\|_F^2 \\ &\text{subject to} \quad \mathcal{A}(\boldsymbol{\mathcal{X}}_{i_1 i_2 i_3}) = \mathcal{A}(\boldsymbol{\mathcal{T}}_{i_1 i_2 i_3}), \quad (i_1, i_2, i_3) \in \Omega \ , \end{aligned} \tag{2.9}$$

where $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}$ and $\mathbf{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3\}$. In the model, α_n are weights that satisfy $\sum_n \alpha_n = 1$ and $\|\mathcal{X}\|_F^2 = \sqrt{\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}} \rangle}$ denotes the Frobenius norm of $\boldsymbol{\mathcal{X}}$.

The drawback of this approach is that all mode ranks r_1, r_2, r_3 must be specified in the algorithm, yet the knowledge of their true values is not assumed. To tackle this difficulty, we adopt the rank increasing scheme introduced in [33]. We dynamically update the rank estimates starting from underestimated ranks for the factor matrices, i.e. $r = 1$, in each mode. The ranks are then gradually increased if the algorithm detects slow progress in the updates of the singular values of their corresponding factor matrices. Thus, (2.9) is solved by cyclically updating $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$. The procedure is performed for all modes in parallel, making this a rather fast recovery process, as well as effective since it exploits the low-rankness of all tensor modes. Though this TC method is non-convex, it has already delivered promising results on both synthetic and real-world MRI and hyper-spectral data [33] [53].

Chapter 3

A novel modular system for missing data completion in WSN-based applications

In this chapter, we describe our proposed modular approach for the effective recovery of missing WSN measurements within the matrix and tensor completion frameworks. We consider three key aspects: How do missing values occur in a typical WSN? Which is the optimal way for structuring the under-sampled data? And should the data be recovered using a centralized or a distributed strategy? We attack these issues by implementing three respective modules, which will be thereafter embedded into a complete processing system for real data experimentation on two specific application domains, in order to evaluate their performance.

3.1 The sampling module

The goal of this module is to implement the realistic conditions described in Section 1.2 that lead to unobserved measurements in WSNs, by considering two distinct approaches. Without loss of generality, in this thesis we consider zeros in the place of missing values we are trying to recover. The percentage of missing data is indicated by the *fill ratio* metric, defined as the ratio of the amount of non-zero elements over the number of all the entries in a measurement matrix of dimensions $[I_1] \times [I_2]$:

$$f = \frac{\#non - zero\, elements}{[I_1] \times [I_2]}.$$

3.1.1 Approach A: Introducing missing values via sub-sampling

In order to reproduce the sub-sampling conditions of WSNs operating in high data rates, such as the ones described in Section 1.2.1, we artificially introduce missing values, namely zeros, in the datasets at hand.

Considering the fact that WSN nodes are able to simultaneously acquire measurements corresponding to multiple sensing modalities, each one often comprising 3-axially mapped data encoded to packets, we implement the sampling module by applying random zero-placement at the same temporal instances for all the data channels related to a specific modality, according to a fixed fill ratio. As a realistic example, consider the failure of the battery of a sensor node leading in missing packets, which are lost at different moments for each one of the node's sensing modalities.

A visual representation of this sampling scheme is illustrated in Figure 3.1, showing a dataset created from the recordings of I_1 sensor nodes, where each node captures n 3-axial modalities of I_2 samples. The white-coloured cells indicate the missing values, i.e. zeros, at the same cells per modality.

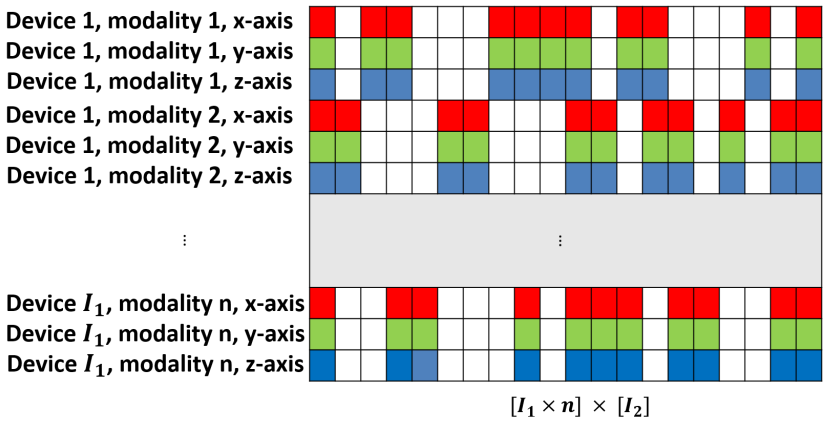

Figure 3.1: Example of artificial sub-sampling in a network of I_1 sensor nodes and n 3-axial modalities of I_2 samples, where zero-placement is applied at the same temporal instances per modality.

3.1.2 Approach B: Introducing missing values via temporal super-resolution

Herein, we attempt to approach the problem of missing WSN measurements from a different perspective, in order to cover WSNs operating at low data rates, as described in Section 1.2.2. Unlike the previous sampling scheme, in this case the zero-valued entries are not randomly introduced according to a specified fill ratio. Instead, they naturally arise as we increase the temporal sampling rate beyond the operating characteristics of the network.

More specifically, let us consider an $[I_1 \times n] \times [I_2]$ fully-populated measurements matrix, depicted in Figure 3.2 (left), resulting from the measurements acquired by I_1 devices, where each node captures n 3-axial modalities at time intervals I_2. The data therein, were recorded at a sampling rate s, which is the maximum sampling rate of our WSN. Increasing the value of s to, *e.g.*, $2s$, leads to an increase in the dimensionality of the measurements matrix, such that the number of columns, i.e. the size of each time interval, is doubled, while the number of rows remains constant, namely $[I_1 \times n] \times [I_2 \times 2]$. Since we are operating beyond the temporal sampling

capabilities of the network, there are not enough measurements to fill the expanded measurements matrix, according to the timestamps on which the measurements were obtained. Subsequently, zero-valued measurements are introduced at cells, as presented in Figure 3.2 (right). Moreover, the value of the filling ratio, f, in this sampling approach is inversely proportional to the increase in the sampling rate. Thus, for the illustrated case, f is set to 0.5.

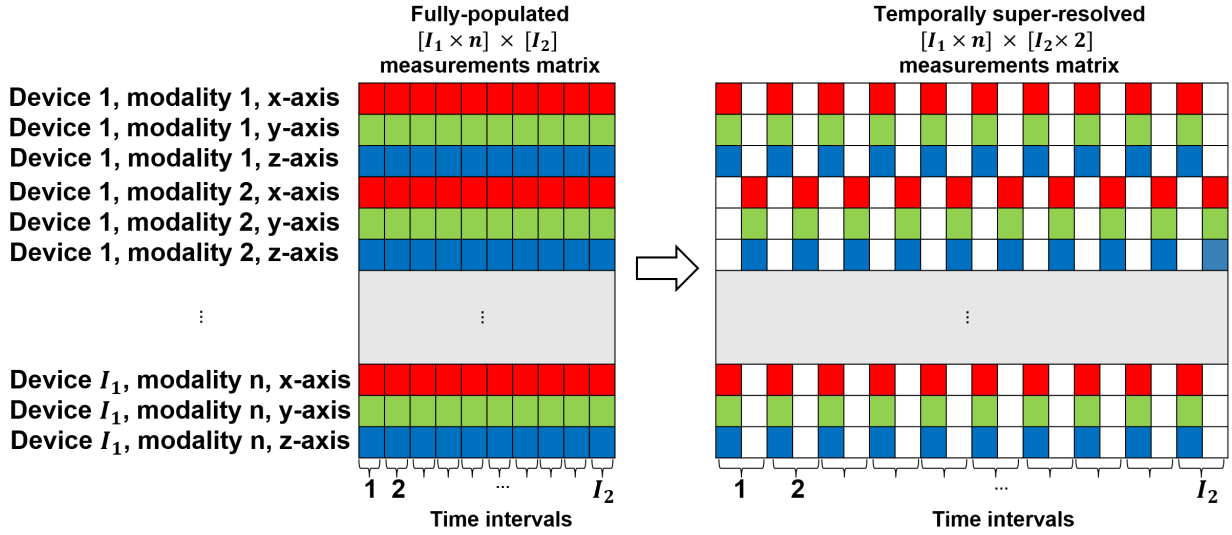

Figure 3.2: Example of temporal super-resolution via doubling the sampling rate in a network of I_1 sensor nodes and n 3-axial modalities acquired at I_2 time intervals.

3.2 The Hankelization module

The purpose of the Hankelization module is to address the fact that real WNS-obtained observations are often contaminated by noise, resulting in sub-sampled matrices that tend to be of full rank. Hence, matrix and tensor completion methods, which assume low-rankness of the underlying data, may be inefficient for high-rank noisy matrices. Herein, we tackle this problem by appropriately organizing the available streams of data into low-rank Hankel matrices, thereby enabling the utilization of MC and TC in the WSN domain.

A Hankel matrix is a square matrix in which each ascending skew-diagonal from left to right is constant, e.g.:

$$\begin{bmatrix} a & b & c & d & e \\ b & c & d & e & f \\ c & d & e & f & g \\ d & e & f & g & h \\ e & f & g & h & i & g \end{bmatrix}$$

Hankel matrices have recently been employed in numerous applications including system iden-

tification [54], recognition of actions in video [55], and the reconstruction of vital signs [56], due to their ability to capture the essence of the temporal evolution of the data in a compact way. We found that this property also stands for the WSN-acquired data, since few of their singular values capture most of their nuclear norm, facilitating the application of our recovery methods. Therefore, we propose the Hankelization process described in the next section.

3.2.1 From data streams to matrices

Assume we have a data time series $X = \{x_1, x_2, \ldots, x_I\}$ of length I. The first step of the Hankelization process involves the *segmentation* of the data stream into consecutive windows of size I_2 with an overlap l. According to these parameters, the time series X is mapped into l lagged windows, $X_i = \{x_i, \ldots, x_{i+I_2-1}\}$ for $i = 1, \ldots, I_1$, where $I_1 = I - I_2 + 1$. The resulting trajectory matrix $\mathbf{X}$ of dimensions $I_1 \times I_2$ is a Hankel matrix having one main property; cross-diagonal elements of $\mathbf{X}$ are equal: $x_{j+i-1} = x_{i+j-1}$. Thus, $\mathbf{X}$ is written:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{I_1} \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 & \ldots & x_{I_2} \\ x_2 & x_3 & x_4 & \ldots & x_{I_2+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{I_1} & x_{I_1+1} & x_{I_1+2} & \ldots & x_I \end{bmatrix}$$

Nevertheless, the random introduction of missing measurements on a Hankel matrix, using the approaches described in the previous section, would violate its structure. Therefore, we take one step back and sample the available fully-populated data streams, in order to obtain Hankel matrices, which contain missing values. The overall Hankelization process is illustrated in Figure 3.3.

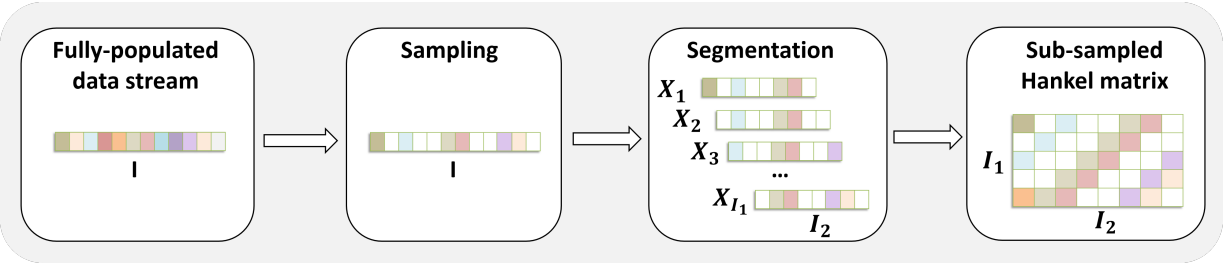

Figure 3.3: The Hankelization process. The available fully-populated data streams are structured into low-rank sub-sampled Hankel matrices.

3.2.2 Generalization: From matrices to tensors

In this section, we scale up our proposed Hankelization process to higher dimensional data, by forming third-order matrices, namely tensors. For an $I_1 \times I_2 \times I_3$ tensor $\boldsymbol{\mathcal{X}}$ to be created, more than one available data streams are required. Each data stream of length I is structured into

a separate "small scale" Hankel matrix $\mathbf{X}$ of dimensions $I_1 \times I_2$ and the tensor $\boldsymbol{\mathcal{X}}$ is formed by vertically concatenating these Hankel-slices along the third dimension, as displayed in Figure 3.4.

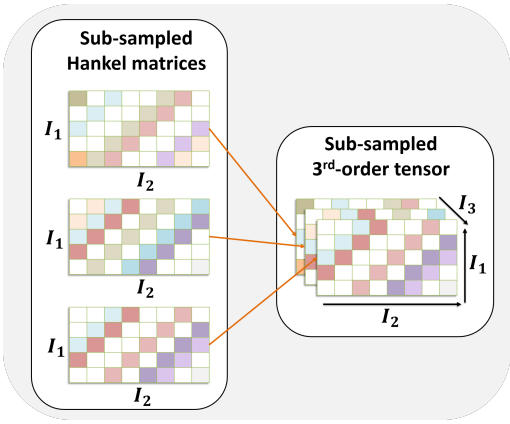

Figure 3.4: Extension of the Hankelization process from matrices to third-order tensors.

3.3 The recovery module

This is the most crucial module proposed in this thesis, where our employed recovery techniques are applied on the constructed low-rank, sub-sampled 2D and 3D Hankel structures, respectively. A key question we aim to answer herein, is related to the origin and structuring of the data involved in each instance of the reconstruction module. Our goal was to explore the ability of MC and TC to exploit the correlations that may exist on data originating from a single or multiple sensing modalities and devices. To this end, we consider three different scenarios as follows:

- *Scenario 1 - Single-device recovery*: In this scenario, we follow a distributed WSN mentality for the reconstruction. More specifically, we assume that each device in the network has access only to its stored sub-sampled structures, that contain the grouped data of all the modalities captured by this single sensing device. Thus, in Scenario 1, the MC-based or TC-based recovery takes place on each device locally, and its instances are as many as the devices within the WSN.

- *Scenario 2 - Collective recovery per modality*: This scenario implements a centralized WSN scheme, where the available measurements are sent to a central processing station, concatenated per modality, and thereafter recovered. More specifically, each call of the reconstruction module herein, involves all the data in the network, corresponding to a specific sensing modality captured by all the devices. Hence, the instances of Scenario 2 are as many as the available modalities.

- *Scenario 3 - Collective recovery*: In this scenario, a centralized WSN is also considered, however, in this case the reconstruction involves the totality of the available data streams, originating from all the sensing modalities and devices in the network, hence there is only one reconstruction call.

Subsequently, recovery is applied on:

- the vertically concatenated collective $[I_1 \times \#$ of involved data streams$] \times [I_2]$ matrices per scenario for the case of MC, and
- the $[I_1] \times [I_2] \times [I_3]$ tensors per scenario for TC.

Finally, the data are reorganized to form the initial, yet reconstructed matrix per data stream. Figure 3.5 illustrates an instance of the 2D recovery module for Scenario 2, regarding the 3-axial modality 1 in a test-bed of 2 sensing devices. The respective instance of the 3D recovery module is depicted in Figure 3.6. Instances of Scenario 1 and Scenario 3 are formed in a similar manner, according to their correspondingly involved data streams.

The modules described above are rather general and can be easily instantiated within numerous WSN-based application domains, challenged by unobserved or lost measurements. In the following two chapters, we examine the efficiency of our proposed methods, by utilizing them for the composition of a processing framework for Smart Water Management (SWM) and a classification framework for Human Activity Recognition (HAR), respectively.

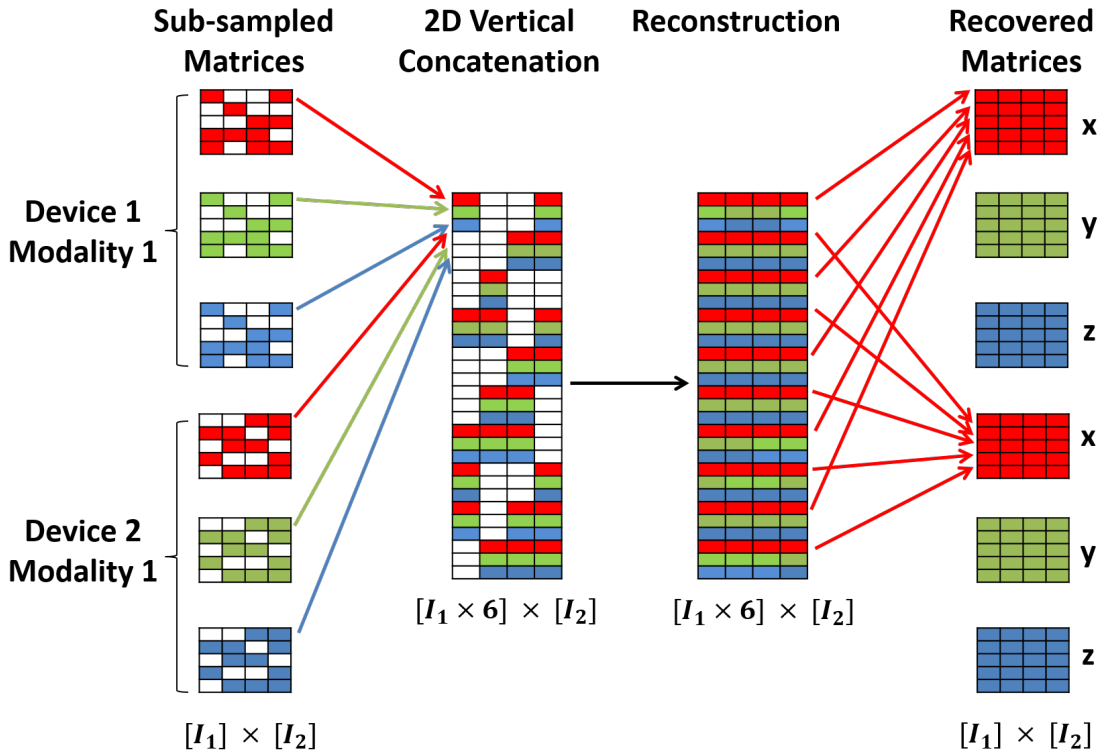

Figure 3.5: Instance of the 2D recovery module for Scenario 2, regarding the 3-axial modality 1 in a test-bed of 2 sensing devices.

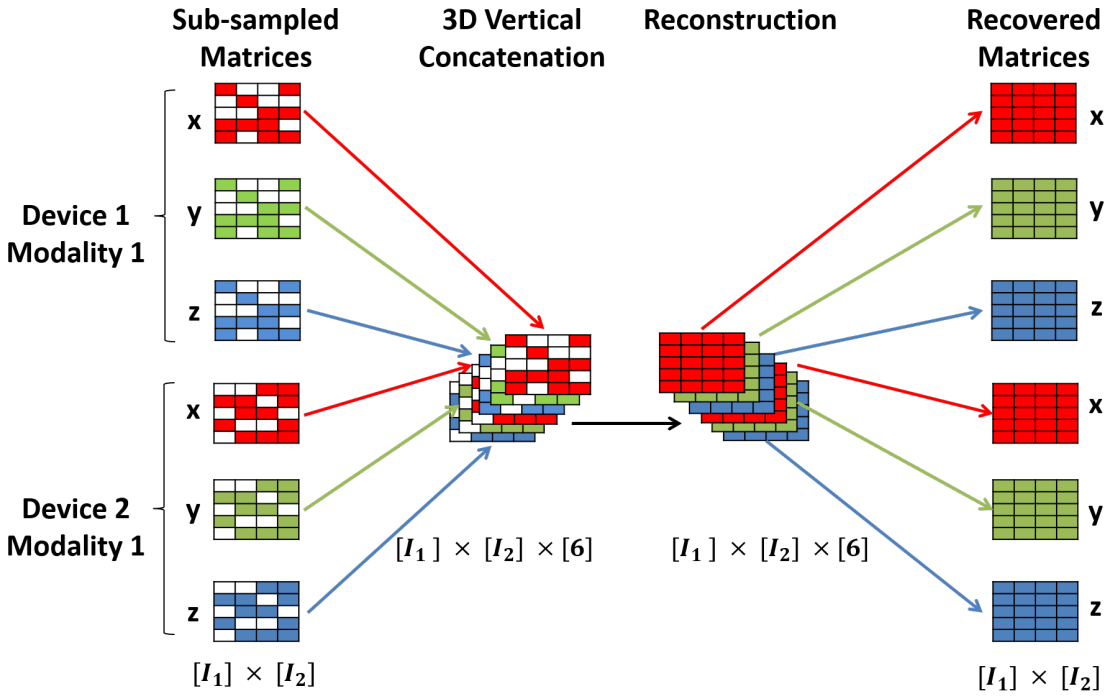

Figure 3.6: Instance of the 3D recovery module for Scenario 2, regarding the 3-axial modality 1 in a test-bed of 2 sensing devices.

Chapter 4

Case study 1: Matrix Completion for the recovery of water treatment data colected by a WSN

Recent advances in Cyber-Physical Systems (CPS) have revolutionized water management in urban areas. Nevertheless, literature reports minor progress in introducing CPS-based systems in industrial water treatment plants, responsible for water purification. Such environments would greatly benefit by adopting CPS technologies in general, and Wireless Sensor Networks (WSNs) in particular. However, WSNs suffer from a series of industrial monitoring constraints, described in Section 1.2, which inevitably lead to missing measurements. In this chapter, we study the problem of efficient estimation of missing water treatment data, collected by a WSN deployed in a water desalination plant. Our goal is to examine how data redundancies can be used for the recovery of extremely under-sampled matrices, by applying the theory of low-rank Matrix Completion via the method of the Augmented Lagrange Multipliers, analysed in Chapter 2. To this end, we implement a realistic framework for experimentation, which is composed by a subset of the modules presented in the previous chapter. We consider three key issues related to the performance of our considered method; namely, efficient recovery of missing measurements, single-device versus collective recovery of measurements matrices, and the problem of temporal super-resolution.

4.1 The proposed framework for SWM

The process of imputing missing water treatment data is incorporated into the framework depicted in Figure 4.1. Our focus is on the sampling and recovery modules. In this section, we extensively describe the methods that we have utilized for their instantiation within the specified application framework.

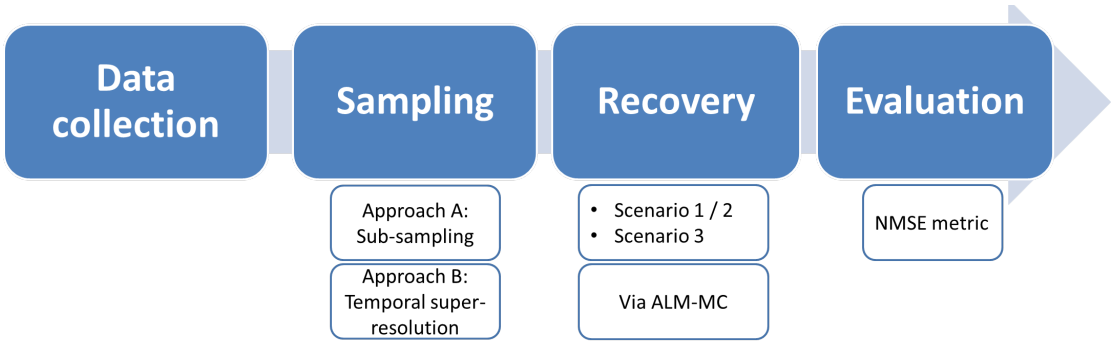

Figure 4.1: Case study 1 - The proposed framework for water treatment data.

4.1.1 The data collection module

Data collection took place at the La Tordera pilot desalination plant, owned by Acciona Agua [1], by deploying a customized WSN solution. The deployment environment and WSN modules are illustrated in Figure 4.2. Each device features an IEEE-802.15.4-based protocol stack, that employs a customized CSMA-based solution for adaptive link scheduling [57] and the IEFT Standard for Routing over Low Power Lossy Networks (RPL) [58].

Figure 4.2: Case study 1 - The desalination plant (left) and the WSN module used for the collection of water desalination data (right).

4.1.2 The sampling module

The sampling module is implemented by using the two approaches presented in Section 3.1, namely via the artificial introduction of missing values via random sub-sampling and by temporal super-resolution of the WSN's maximum sampling rate.

A significant aspect of the sampling process is related to the amount of the unobserved measurements. Therefore, we modify this module accordingly, by engaging numerous cases of missing measurements with the purpose of reaching robust conclusions regarding the efficiency of our proposed recovery method. The key parameter herein, dictating the percentage of missing values within the data at hand, is the fill ratio f, defined in the previous Chapter. High values

[1] `http://www.acciona-agua.com/`

of f result in relatively "dense" matrices lacking only a few measurements, while low fill ratios lead to "sparse" matrices where the majority of measurements is unobserved.

4.1.3 The recovery module

For the recovery module we used the ALM-based MC method, described in Section 2. Note that, in this case study we do not implement the aforementioned Hankelization process, since we are dealing with rather low-rank data. Thus, instead of structuring the available measurements into Hankel matrices, we adopt the simple approach of the vertical concatenation of the data into one large sub-sampled measurements matrix of dimensions $[I_1 \times n] \times [I_2]$, as depicted in Figure 4.3. I_1 stands for the number of devices included in the WSN, n for the number of the data streams, namely channels, captured by each device, and I_2 for the timestamps of the measurements. Moreover, the data are grouped per device, and we refer to the "virtual" frequency of each device as a *sensing modality*. This allows us to generalize the specific study and its conclusions, to other cases of WSN-based monitoring beyond water treatment.

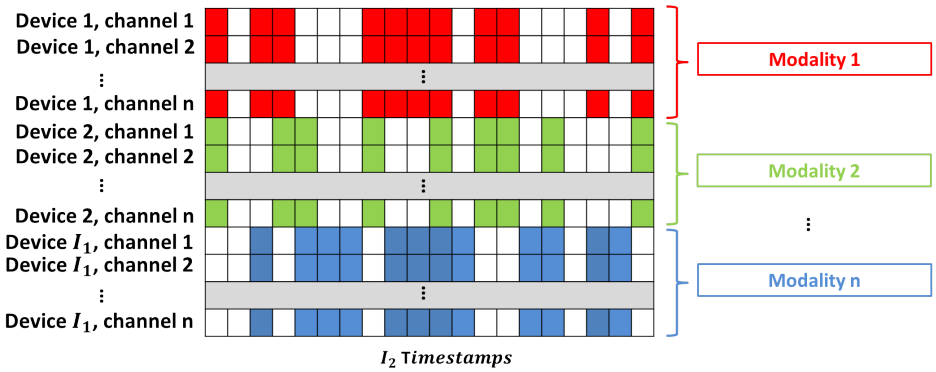

Figure 4.3: Case study 1 - The proposed structuring of data (general case).

As far as the recovery scenarios are concerned, the implementation of the recovery module herein concentrates on the number of devices providing the measurements. Notice that, since each device in this case study represents a modality, Scenario 2, i.e., the collective recovery per modality, and Scenario 3, i.e., the overall collective recovery, are identical. Thus, the performance of the MC-based reconstruction is evaluated, when considering data from a single device (Scenario 1), as opposed to Scenario 2/3, where collective recovery is performed using measurements from all the devices/virtual modalities. According to the first option, MC-based recovery takes place locally at each device. Practically speaking, in this scenario, recovery takes place on a sub-matrix of the measurements matrix shown in Figure 4.3, by employing only the data corresponding to a specified device. The second option, incorporates a sink and applies MC to the collective measurements matrix, $I_1 \times I_2$.

4.1.4 The evaluation module

A key issue for the implementation of the evaluation module, is related to the specific metric that we use in order to quantify the obtained results. We evaluate the proposed framework based upon the commonly used Normalized Mean Square Error (NMSE), defined as the mean squared error between the initial fully-populated and the reconstructed measurements matrix, normalized with respect to the l_2 norm.

4.2 Experimental evaluation of the proposed framework

The goal of this section is to assess the effectiveness of the ALM algorithm for MC on WSN-based water desalination data against various aspects, which are related to the volume of the data, the percentage of missing entries, which arise either due to sub-sampling or temporal super-resolution, and the number of devices providing measurements.

4.2.1 Dataset

The dataset considered for our evaluation purposes contains water impedance measurements (Ohm), sampled at $n = 10$ different channels per device, by $I_1 = 5$ devices at different stages of the desalination process. WSN measurements were collected for a 3 day period (25th, 26th, 27th October 2014) and the sampling rate of the nodes was set to 1 measurement per hour per node. Hence, our initial, fully-populated measurements matrix, contains in total $[I_1 = 5 \times n = 10] \times [I_2 = 3 \times 24] = 3600$ measurements. Thus, the general underlying matrix containing the available measurements, illustrated in Figure 4.3, is instantiated as presented in Figure 4.4, according to the dataset at hand.

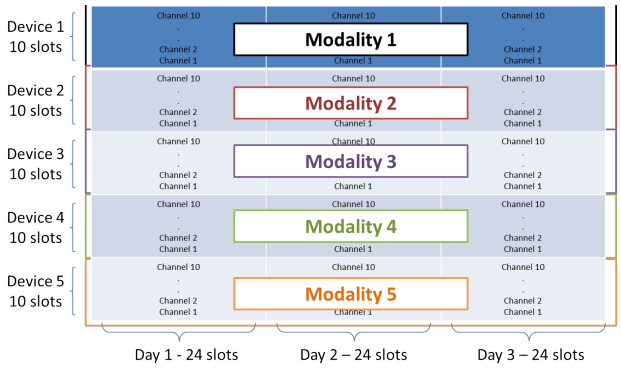

Figure 4.4: Case study 1 - The proposed structuring of data, based upon the water dataset at hand.

4.2.2 Effects of measurement matrix dimensions

In this experiment, we investigate the Matrix Completion recovery capabilities as a function of the sub-sampling factor, f, on the measurements matrix. The objective is to assess how the

MC-based recovery performance is associated with the volume of the data, by using four different sizes of measurements matrices, generated from the initial fully-populated [50] × [72] matrix.

Therefore, we gradually reduce the dimensionality by a factor of two, by assuming that for each day the number of sampling instances, *i.e.*, columns of the matrix, is halved, while the number of rows, *i.e.*, number of modalities, remains constant. As a result, the measurements matrix dimensions at the second experiment is [50] × [36 (2 hours sampling interval)], at the third experiment is [50] × [18 (4 hours sampling interval)], and so on. For each data size, f was initiated at 0.1 and was iteratively incremented up to 0.9, with a step size equal to 0.02.

Figure 4.5, depicts the recovery performance measured by the NMSE as a function of f. We can observe that increasing f has a positive effect on the reconstruction quality. Regardless the size of the measurements matrix, the amount of missing values affects crucially the reconstruction performance. The results also show that the NMSE converges at high values of f, for all 4 sizes of the measurements matrix. Moreover, it is demonstrated that, for larger data matrices, the convergence of NMSE is much smoother.

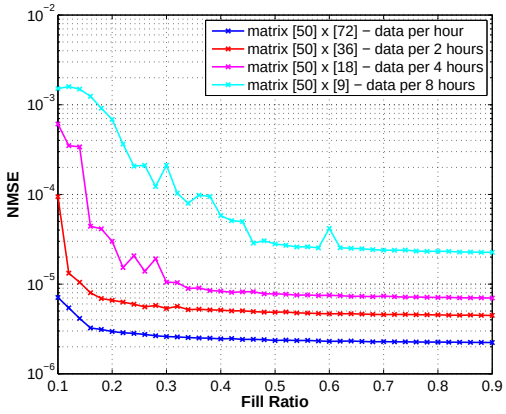

Figure 4.5: Case study 1 - Normalized MSE as a function of f with respect to 4 different measurements matrix sizes.

Figure 4.6(a) depicts a fully sampled [50] × [72] measurements matrix and Figures 4.6(b)-4.6(d) present the MC reconstructed matrices when f equals to 0.1, 0.5, and 0.9, respectively. It is apparent that higher fill ratios lead to more accurate measurements reconstruction, in accordance with both the theoretical models, as well as the quantified results presented at Figure 4.5. Furthermore, we observe that although the reconstruction at 0.1 exhibits noisy artifacts, one can still get an meaningful overall sense of the behaviour of the data. Finally, it is shown that even 50% of the measurements can produce very accurate estimations. This suggests that a dramatic reduction in the sampling requirements for this application is possible through MC reconstruction.

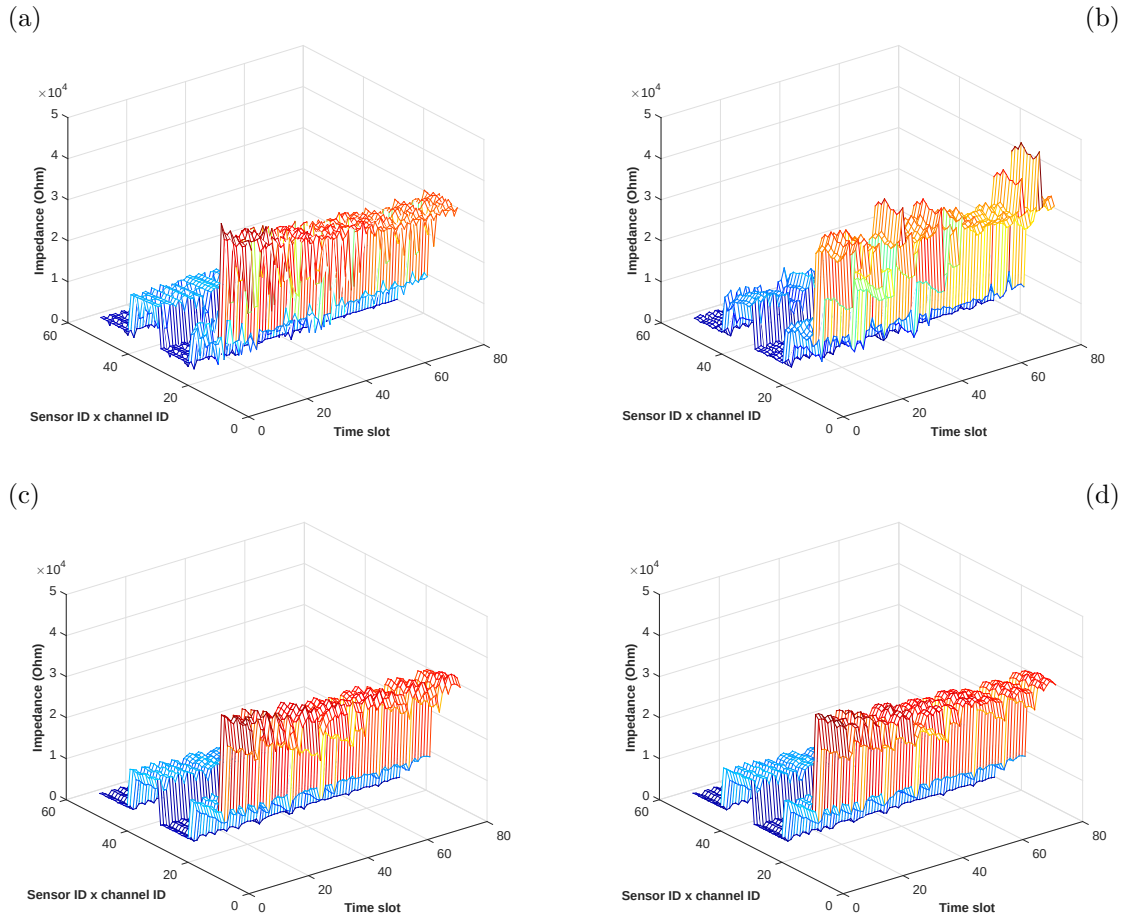

Figure 4.6: Case study 1 - (a) The original fully-sampled measurements matrix and the MC-recovery results: (b) reconstructed matrix from 10% of the measurements, (c) reconstructed matrix from 50% of the measurements, (d) reconstructed matrix from 90% of the measurements.

4.2.3 MC for temporal super-resolution

In this set of experimental results, we attempt to approach the problem of missing measurements from a different angle, related to the low temporal sampling characteristics of this application's underlying WSN. Herein, the missing values naturally arise as we increase the temporal sampling rate beyond its maximum value, thus performing temporal super-resolution, as described in Section 3.1.2.

More specifically, the $[50] \times [72]$ fully-populated measurements matrix corresponds to a sampling rate of 1 measurement every 60 minutes, for all 5 devices and 10 channels, which is the maximum sampling rate of our testbed. Increasing this sampling rate to, *e.g.*, one measurement every 15 minutes, leads to an increase in the dimensionality of the measurements matrix to $[50] \times [4 * 72]$. The number of columns is quadrupled, but there are not enough measurements obtained to fully-populate the expanded measurements matrix, thus leading to zero-value filled cells. In this case, f is equal to 0.25. Table 4.1 presents the dimensionality, and corresponding values of f for the sampling rates used in this experiment.

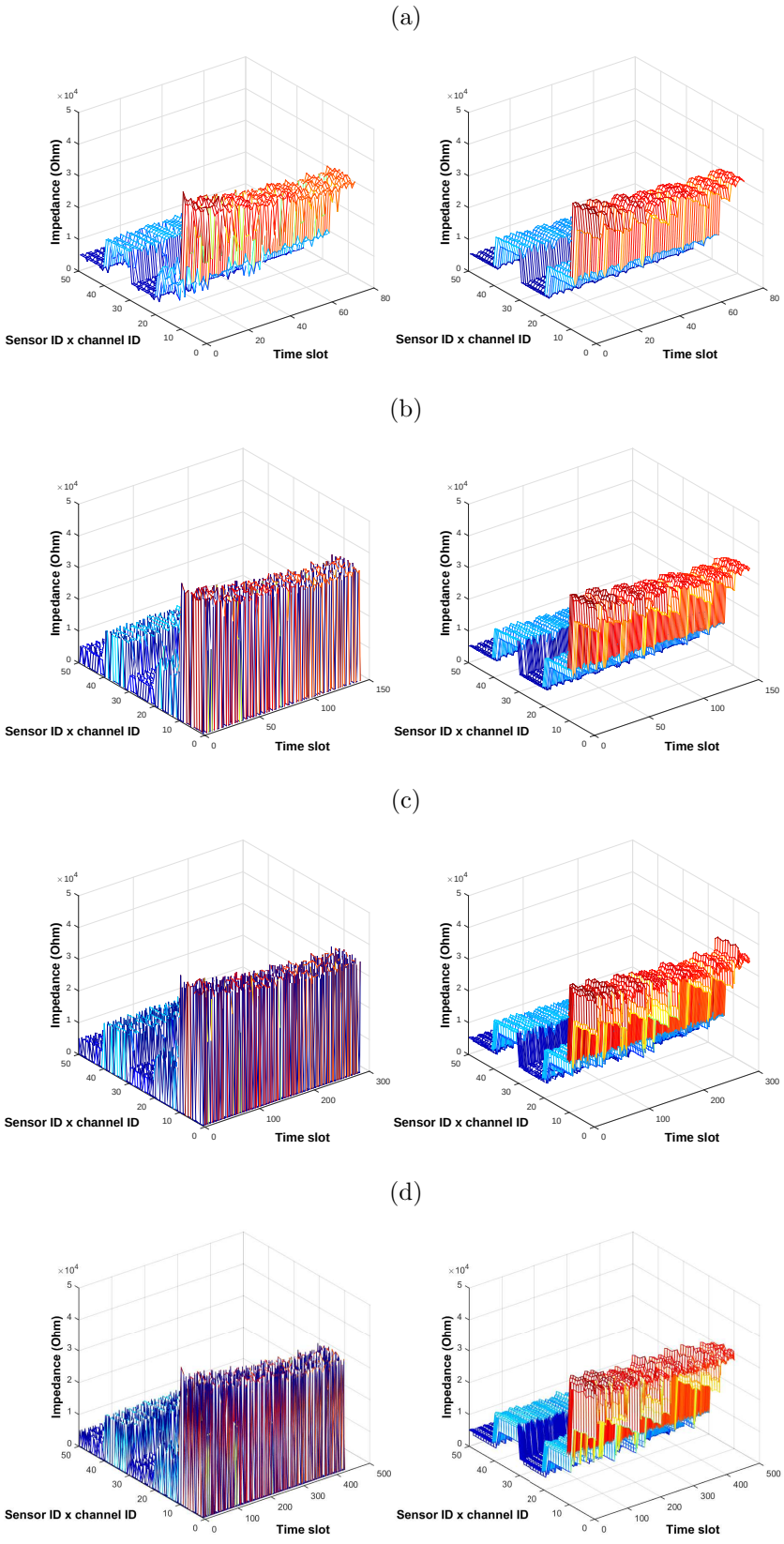

Figure 4.7: Case study 1 - The initial (left) and reconstructed (right) when the dimensions of the measurements matrix are (a) [50] × [72], (b) [50] × [144], (c) [50] × [288], (d) [50] × [432].

Table 4.1: Case study 1 - Dimensionality and fill ratio for different sampling rates.

Sampling rate	Dimensionality	Fill ratio (f)
60 minutes	$[50] \times [72]$	1
30 minutes	$[50] \times [144]$	0.5
15 minutes	$[50] \times [288]$	0.25
10 minutes	$[50] \times [432]$	0.16

Herein, it is not possible for us to calculate the NMSE metric, since the reference measurements matrix does not exist. Nevertheless, we have conducted this experiment, in order to *visually* assess the performance of MC-based recovery. This corresponds to the realistic conditions, where the reference measurements matrix will not be available. We performed the MC reconstruction experiment for 4 different sizes of matrices, presented in Table 1. Results are shown in Figures 4.7(a)-(d).

Since we super-resolve in the temporal domain, the lack of ground truth data means that we cannot estimate the performance via some error metric. However, the visual observations made suggest that, while the dimensionality of the matrices increases and the fill ratio decreases, the MC-reconstructed data maintain their smoothness and distribution to a great extent, compared to the initial fully-populated measurements matrix of dimensions $[50] \times [72]$. This 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.

4.2.4 Impact of the number of devices

In the final set of experiments, we evaluate the performance of MC-based reconstruction when considering data from a single device (Scenario 1), as opposed to scenarios where collective recovery is performed using measurements from all the devices/virtual modalities (Scenario 2/3).

Figures 4.8 and 4.9 illustrate the recovery performance with respect to f for devices 1, 2, and 5, respectively. Note here that the NMSE is calculated between the initial fully populated sub-matrix of each device and recovered sub-matrix according to Scenario 1 and Scenario 2/3, respectively.

Regarding devices 2 and 5, we observe that, as we move on to higher values of f, *i.e.*, above 0.4 and 0.5 respectively, the single-device recovery achieves better reconstruction results. This behaviour is different than the case of device 1, where the reconstruction quality in the collective recovery scenario is better compared to the single-device case for high values of f. The experimental results suggest that, by exploiting the intra-device correlation, collective reconstruction can achieve superior performance compared to the single-device case for most situations, although at high sampling rates, the performance gain is marginal.

As far as device 1 is concerned, it is demonstrated that high reconstruction quality in the single-device case can indeed be accomplished using just local measurements. Nevertheless, collective recovery achieves better reconstruction, over all different values of f. This result suggests

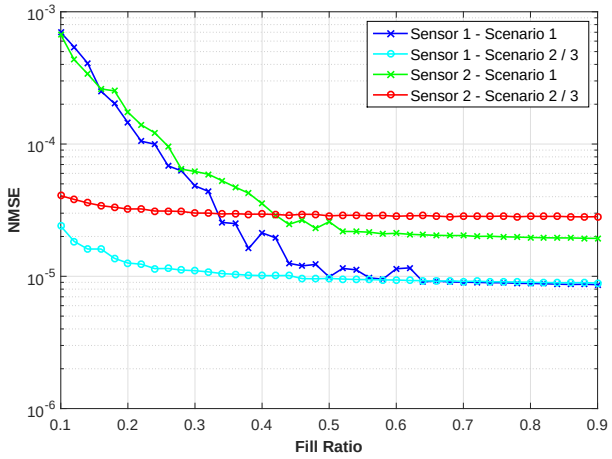

Figure 4.8: Case study 1 - NMSE w.r.t. f for single-device vs. collective recovery regrading the devices 1 and 2.

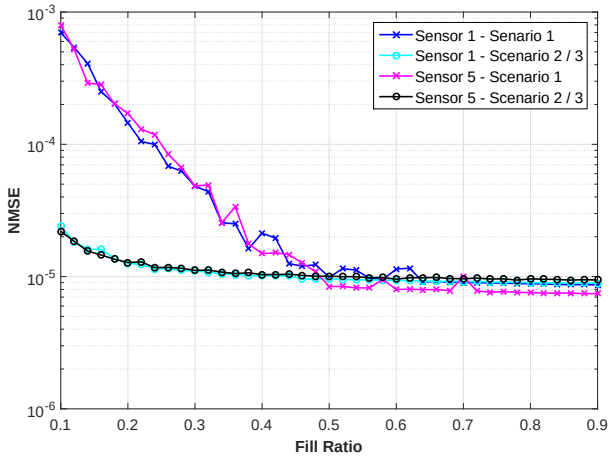

Figure 4.9: Case study 1 - NMSE w.r.t. f for single-device vs. collective recovery regrading the devices 1 and 5.

that collective MC recovery has the ability to fully utilize the correlation that exists among devices, even if such correlations are not explicitly encoded into the recovery process, thus highlighting the generalization ability of our proposed schemes.

Chapter 5

Case Study 2: Matrix and tensor completion for recovering multi-modal HAR data

Sensor-based human activity recognition (HAR) is encountered in many applications in the area of pervasive healthcare and plays a crucial role in biomedical research. However, a major challenge related to this domain lies in the poor performance of machine learning algorithms in the common case of unobserved or missing measurements. In this chapter, we study the problem of accurate estimation of missing multi-modal physical data and we propose a complete framework for data structuring, reconstruction, classification, and assessment of the overall recognition process in the scenario of unobserved values. We introduce the concept of organizing the available data streams into low-rank Hankel structures and we exploit data redundancies using sophisticated recovery techniques, with an emphasis on Matrix and Tensor Completion. Moreover, we examine the interaction between the data reconstruction and the subsequent classification steps, by experimenting with several state-of-the-art classifiers. The proposed framework is evaluated with respect to various levels of missing values, that are uniform or non-uniform per sensing device, as well as different collective recovery scenarios in terms of data structuring and origin. Finally, the influence of sub-sampling on the battery consumption of Shimmer sensing platforms is reviewed. Our experimental findings rely on two public datasets containing physical data, that extend to numerous activities, multiple sensing modalities and devices.

5.1 The proposed framework for HAR

Activity recognition in the presence of reconstructed values is incorporated into a complete and modular classification framework depicted in Figure 5.1. Similar to other machine learning applications, our proposed classification framework contains a training phase for the classifiers to produce their predictive models, prior to the testing phase which evaluates the performance of

the system.

Empirical work on handling missing values within classification frameworks has primarily addressed the challenge of induction from incomplete training data [32], [59]. Nevertheless, in this case study, we depart from this approach by proposing a novel framework, where the unobserved measurements are introduced in the evaluation stage. This consideration corresponds to realistic scenarios, where the system training can be performed off-line in ideal sensing conditions, yet the evaluation stage is dynamically affected by numerous constraints that lead to missing values. Hence, both the training and the testing stages involve the modules of data partitioning, hankelization, feature extraction, and classification, while the recovery module is integrated at the testing phase. The following sections contain a thorough description of the aforementioned framework modules.

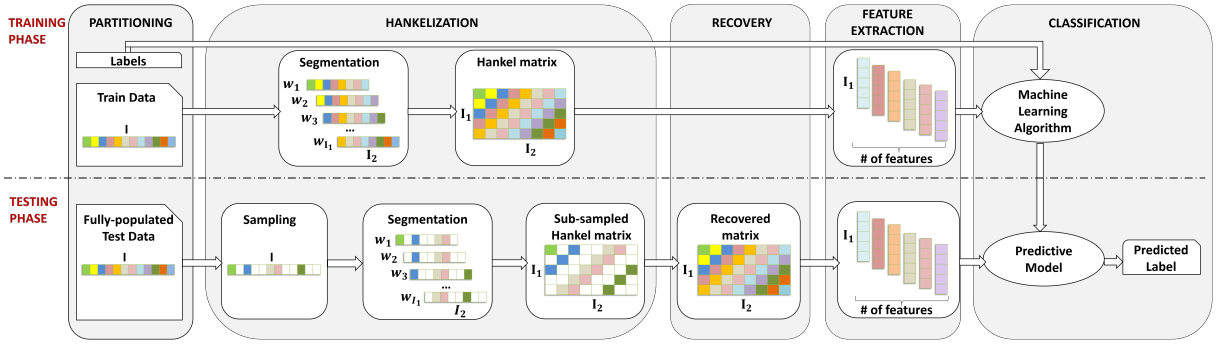

Figure 5.1: Case study 2 - The proposed HAR framework. Reconstruction is incorporated in the testing phase. Subsequently, feature extraction and classification is performed using the predictive models formed on training phase.

5.1.1 The partitioning module

Datasets at hand are randomly and per user partitioned into non-overlapping training and testing sets, imposing that each subset encompasses at least one instance from each of the activities. More specifically, for the training phase a subset of the users IDs is randomly selected. The time-series data streams corresponding to the traces of these users are acquired from the database and constitute the input to the training phase. The remaining user IDs and their respective data are utilized for evaluation at the testing phase.

5.1.2 The Hankelization module

During the training phase, the $[I_3]$ fully-populated sensor streams are segmented to $[I_1]$ consecutively lagged temporal windows of I_2 samples and structured to form their respective (2D) Hankel matrices or (3D) Hankel-sliced tensors, depending on the subsequently applied recovery technique, through the Hankelization process described in Section 3.2.

During the testing phase, prior to segmentation, data sampling is applied in order to artificially

introduce missing values in the test data streams. We apply random zero-placement at the same instances for all data streams of each sensing modality, following the rational described in Section 3.1. Hence, the succeeding Hankelization process on the sub-sampled test data streams generates 2D or 3D structures containing unobserved values that need to be reconstructed.

5.1.3 The recovery module

This is the most crucial module of the overall procedure, where the recovery techniques are applied on the previously formed sub-sampled Hankel structures. In this case study, we focus on the two recovery techniques, namely Matrix Completion (MC) and Tensor Completion (TC), described in Chapter 2. We also provide a direct comparison of MC and TC, to another sophisticated method for reconstruction, that of the Regularized Expectation Maximization (RegEM) [60]. In the following, we give a brief description of the RegEM imputation technique.

RegEM is a sophisticated iterative method for finding maximum likelihood estimates of parameters, where the model depends on unobserved latent variables. The RegEM iteration alternates between filling in the missing values in the sub-sampled matrix with their conditional expectation values, given the available data and the estimates of the mean and of the covariance matrix, and by revising the estimates of these parameters. These estimates are then used to determine the distribution of the latent variables in the next iteration. The distribution parameters of the regression model are computed by an individual ridge regression for each missing value, until they converge or until a predetermined maximum number of iterations is reached. This is a rather computationally complex, yet very effective algorithm [60], applied for the reconstruction of vital signs, as well as health care data [56], [61].

Another key question we aim to answer herein, apart from the optimal utilized recovery technique, is related to the origin and structuring of the data involved in each reconstruction instance. To this end, we consider the three scenarios presented in Section 3.3, involving data from a single or multiple sensing modalities and devices. For each scenario, we employ the proposed methods for recovery and we evaluate the reconstruction quality using the Normalized Mean Square Error (NMSE) metric for various cases of missing data determined by the fill ratio f.

5.1.4 The feature extraction & classification modules

During training, the previously formed high dimensional fully-populated structures are transformed to a lower dimension feature space through the procedure of feature extraction. For each time window of training data, a vector of 22 informative and non-redundant statistical features, namely the mean, standard deviation, min, max, 1st component of principal component analysis, interquartile range, variance, kurtosis, skewness, median, zero crossing rate, and a histogram of 10 bins, is extracted and given as input to each machine learning algorithm. Then, the respective activity recognition models are produced. Likewise, at the testing phase, the extracted feature

set of the partially observed test Hankel matrices is evaluated on each of the previously trained learning models, generating the predicted activity label.

5.1.4.1 Utilized classifiers

For classification, we employ the following state-of-the-art off-the-shelf classifiers: [1]

Decision trees: This classifier builds a hierarchical model in which each internal node represents a test on an attribute, e.g. whether a coin flip comes up heads or tails, each branch stands for the outcome of the test, and each leaf node corresponds to a class label, namely the decision taken after computing all attributes. The paths from root to leaf represent classification rules. We use Gini's diversity index as the impurity reduction criterion of our decision tree and set the maximum number of splits to 50.

K-Nearest Neighbours (K-NN): The K-NN classification algorithm uses the principle of similarity, i.e. distance, between the training set and the new observation to be classified. The new observation is assigned to the most common class through a majority vote of its K nearest neighbours. The distance of the neighbours of an observation is calculated using a distance metric called similarity function. In this work, we engage two versions of this classifier, namely the Euclidean and the Cosine distanced K-NN respectively, with $K = 10$ neighbours.

Support Vector Machines (SVMs): SVMs rely on kernel functions that project all instances to a higher dimensional space with the aim of maximizing the margin around a decision boundary, i.e. hyperplane, to partition the data. In their standard formulation, SVMs are linear classifiers. However, non-linear classification can be achieved by extending SVM through kernel methods. The key idea of kernel methods is to project the data from the original data space to a high dimensional space called feature space by using a given non-linear kernel function. We apply two versions of SVMs, using *Gaussian* and *Quadratic* kernels, respectively.

The aforementioned parameters have been fine-tuned through experimentation and provide the best predictive model for each of the above listed classifiers. In addition, we use classification accuracy as our evaluation metric to assess the effectiveness of the classification algorithms in use.

5.2 Experimental evaluation

The main objective of this chapter is to empirically evaluate the effect of test data recovery on the accuracy of the subsequent classification. We start by describing the datasets in use and we follow up with the set-up of each conducted experiment and the demonstration and discussion of the derived results.

[1] `http://www.mathworks.com/help/stats/classificationlearner-app.html`.

5.2.1 Datasets

In this study, we consider the following two popular databases for the classification of activities of daily living, shortly described in Table 5.1:

- ***HAR Smartphones Dataset*** [62], [63]: This dataset contains body motion recordings of thirty subjects performing 5 physical activities (Walking, Climbing stairs, Sitting, Standing, Laying). A smart phone (Samsung Galaxy S II) placed on the subject's waist was used to capture 3-axial linear acceleration and angular velocity motion signals. The raw data were pre-processed by applying noise filters.
- ***MHEALTH Shimmer Dataset*** [64], [65], [66] : This dataset was created from the sensor recordings of ten subjects performing 12 daily activities (standing still, sitting and relaxing, lying down, walking, climbing stairs, waist bends forward, frontal elevation of arms, knees bending, cycling, jogging, running, jumping front & back). Shimmer2 wearable sensors attached on the subject's chest, right wrist, and left ankle were utilized as the devices recording multi-local 3-axial linear acceleration, rate of turn, and magnetic field motion signals. We consider this dataset as a generalization of HAR in terms of the variety of the number, intensity and execution speed of activities, as well as the diversity of sensing devices and modalities involved.

Data streams of both datasets were obtained at a sampling rate of 50 Hz, while the experiments were labelled manually through video-recordings.

Table 5.1: Comparison of the employed datasets for Human Activity Recognition.

	HAR	**MHEALTH**
Device	Smart phone (Samsung Galaxy S II)	Shimmer2
Number of locations	One	Multiple
Body Parts	Waist	Chest Right wrist Left ankle
Number of activities	5	12
Modalities	Accelerometer Gyroscope	Accelerometer Gyroscope Magnetometer

In the following experimental sections, the datasets at hand were partitioned to 70% for training and 30% for testing, respectively. The available time-series data streams per dataset were segmented in I_1 consecutive temporal windows of $I_2 = 128$ samples, to form the aforementioned Hankel structures with a 50% overlap. Our experimentation (see Appendix A) demonstrated these as the optimum parameters, in accordance to what is reported in the bibliography [67].

5.2.2 Effects of measurement matrix dimensions on MC-based NMSE

In this set of experiments, we investigate the matrix completion recovery abilities as a function of f and the size of the measurements matrix. The objective is to assess how the NMSE is associated to the size of the data, by evaluating the recovery as a function of different sizes of measurements matrices ranging from $I_1 = 34$ up to 340, consecutive windows of $I_2 = 128$ samples, and fill ratios from $f = 0.1$ to 0.9 with a step size of 0.1. The presented results come from the x-axis data stream of the available modalities in the HAR dataset. Our observations are extended to the MHEALTH dataset.

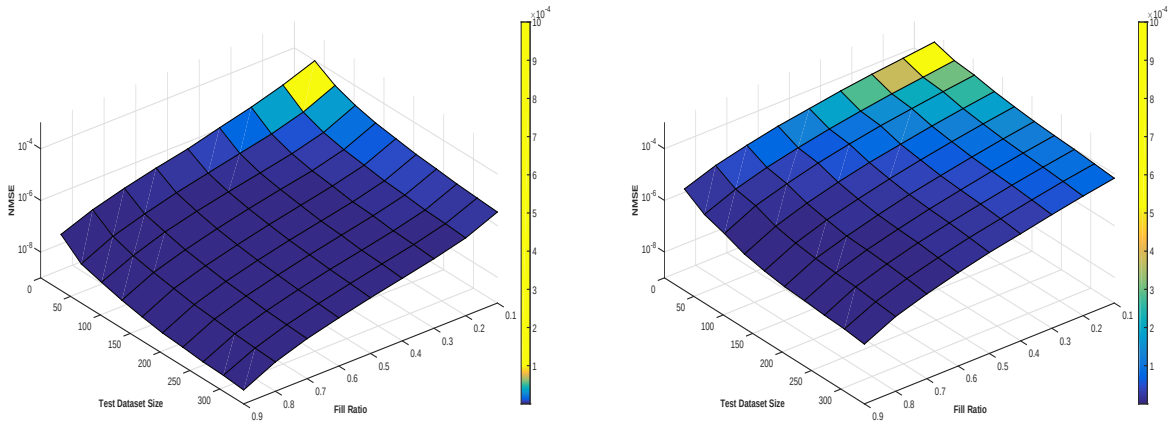

Figure 5.2: Case study 2 - NMSE as a function of f and test data size for the x-axis accelerometer (left) and the x-axis gyroscope (right) of the HAR dataset, considering Scenario 1.

Figure 5.2 illustrates the recovery performance measured by the NMSE as a function of f and the size of test data for Scenario 1. We observe that higher fill ratios lead to more accurate data reconstruction, as expected. Moreover, the size of the measurements matrix plays a crucial role to the recovery performance, since larger matrices are clearly shown to present lower reconstruction error. This is reasonable, considering that larger data matrices contain a greater number of observed measurements, which can be exploited by the MC method for more precise reconstruction of the unobserved values. However, one cannot fail to notice that there is an important trade-off concerning the associated computational complexity, as the matrices grow to higher dimensions.

A final significant comment regarding this experiment, is related to the relative magnitude of the NMSE between the two modalities under scope on HAR. It is noticed that the gyroscope channel presents much higher overall NMSE, than that of the accelerometer for all fill ratios and matrix sizes. This can be explained by examining the magnitude of the singular values of each channel individually, as depicted in Figure 5.3. As demonstrated, the accelerometer data exhibit a much higher linear correlation than the gyroscope ones, manifested by the smaller number of dominating singular values. The resulting observed superior performance of MC upon

the accelerometers comes in accordance to the underlying theoretical model for this recovery method, described in Section 2.2.

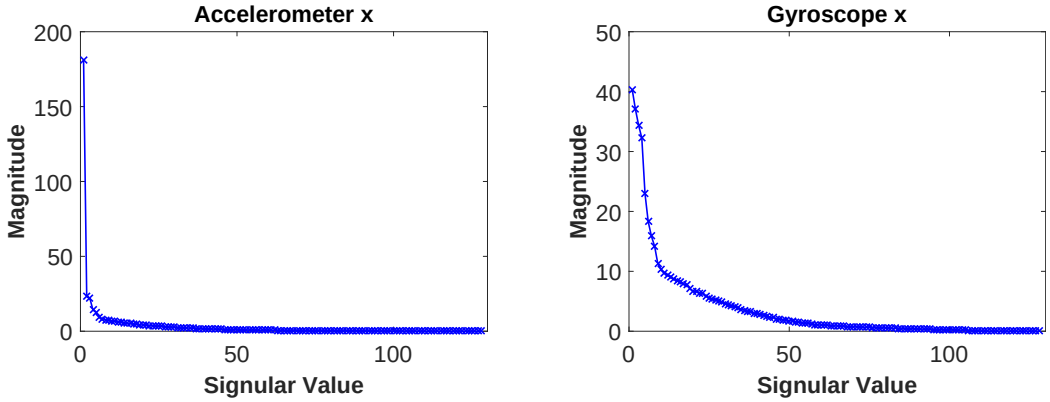

Figure 5.3: Case study 2 - The magnitude of the singular values of x-axis accelerometer (left) and x-axis gyroscope (right) channels of the HAR dataset.

5.2.3 Sufficient training time per dataset

The objective of this experiment is to define the sufficient training period for all of the employed classifiers, on fully-populated test data streams, i.e. $f = 1$, in order to provide a ground truth for our later experiments on partially observed data. Obviously, in this experiment we omit the sampling step of the testing phase in order to obtain full test Hankel matrices.

The system is trained with numerous sizes of randomly selected data per user, corresponding to up to 70% of each dataset. This percentage is proportionate to the data of up to 21 and up to 7 distinctive subjects of the HAR and MHEALTH datasets, respectively. Since our data streams are captured at a constant rate of 50Hz, 7.25 minutes are needed to capture the data of an average user ($I_1 = 340$ windows) of the HAR dataset and 11.2 minutes per user ($I_2 = 525$ windows) for the MHEALTH dataset. Subsequently, the classifiers are tested on the resulting 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 considers the data of one randomly selected user.

Figure 5.4 illustrates the performance of each classifier measured by the classification accuracy as a function of training time. As expected, the increase of time for training has a pronounced effect on the system's learning. Is it observed that all considered classifiers present stable performance when trained for at least 100 minutes of non-recurring data for the HAR dataset. For the MHEALTH dataset, the lower bound of efficient training time is 60 minutes. Moreover, results suggest that Support Vector Machines outperform all other employed classifiers and manage accuracy of over 90% for both datasets when trained for sufficient time.

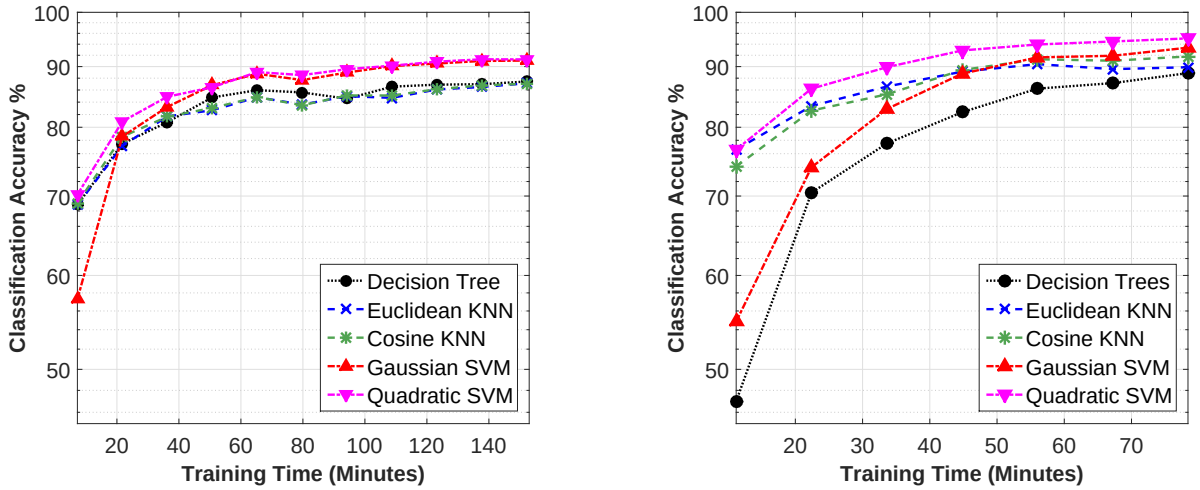

Figure 5.4: Case study 2 - HAR (left) and MHEALTH (right) classification accuracy w.r.t. training time for all classifiers on fully-populated test matrices. (Mean of 10 runs.)

5.2.4 Comparing Different Recovery Techniques

In this set of experiments, we are interested in comparing the performance of the three considered methods for reconstruction, namely MC and RegEM for matrices and TC for tensors. Figure 5.5 illustrates comparative plots of the NMSE (top) and the corresponding running times (bottom) of all applied recovery techniques as a function of f, on HAR (left) and MHEALTH (right), respectively. The presented results are cumulative for the total of the available streams per dataset and regard the *single-device recovery* (Scenario 1). Results indicate that RegEM achieves the best reconstruction quality on both datasets, followed by MC, while TC has the worse performance at all fill ratios, in terms of the NMSE metric.

Regarding the HAR dataset (left), we notice that RegEM and MC perform almost identically in terms of NMSE, with RegEM slightly outperforming MC at low fill ratios (top-left). However, the superiority of RegEM comes with a significant computational complexity that leads to a remarkable increase in its running time (bottom-left). Indicatively, at $f = 0.4$ RegEM is executed in 220.7 seconds $\simeq$ 3.6 minutes, while MC needs only 2.7 seconds. On the other hand, TC is inferior concerning the cumulative NMSE metric, especially at low fill ratios. In terms of the recorded TC running time, it is much faster than RegEM, yet slower than MC, since it is executed on approximately 20 seconds at all fill ratios.

Similar observations hold for the MHEALTH dataset (right), where RegEM once again slightly outperforms MC but it needs two orders of magnitude more running time. It is noteworthy that at $f = 0.4$, MC runs for 13.7 seconds while RegEM requires 2101 seconds, i.e. $\simeq$ 35 minutes on this larger dataset.

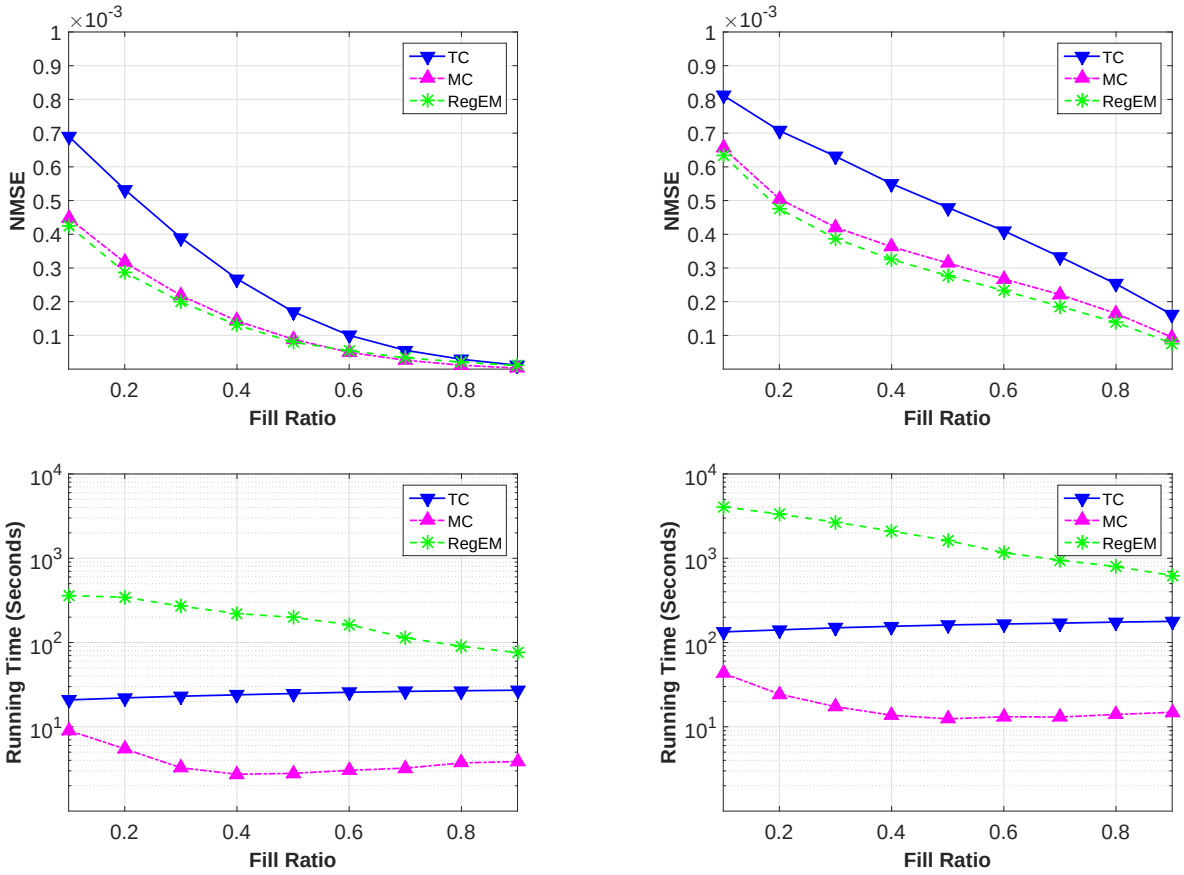

Figure 5.5: Case study 2 - Cumulative NMSE (top) and corresponding running times (bottom) w.r.t. f of all applied reconstruction methods for HAR (left) and MHEALTH (right). (Scenario 1 - Mean of 10 runs.)

5.2.5 Running Time per Recovery Scenario

The goal of this experiment is to further examine the running time of the employed reconstruction methods, especially when moving from the relatively "small scale" collective matrices/tensors of Scenario 1 to higher volumes of data required by Scenarios 2 & 3. The results concerning the corresponding cumulative reconstruction times for all data streams are presented in Figure 5.6. Note that for the HAR dataset, *single-device recovery* (Scenario 1) is identical to *collective recovery* (Scenario 3), since in this dataset there is only one sensing device.

Figure 5.6 (top) shows that TC exhibits an almost constant running time for all fill ratios, which increases slightly for higher values of f. This outcome suggests that the computational complexity of the TC algorithm is not directly related to the volume of missing data to be recovered. However, it is strongly associated to the size of the underlying tensor, as expected, since Scenario 3 requires slightly longer execution times than Scenarios 1 & 2. The difference in the size and structuring of each tensor per scenario is dictated by its third dimension, namely the number of vertically concatenated Hankel-slices. This can range from $I_3 = 3$ available channels for the sensor attached on the subject's chest (for an instance of Scenario 1) to $I_3 = 21$ (Scenario

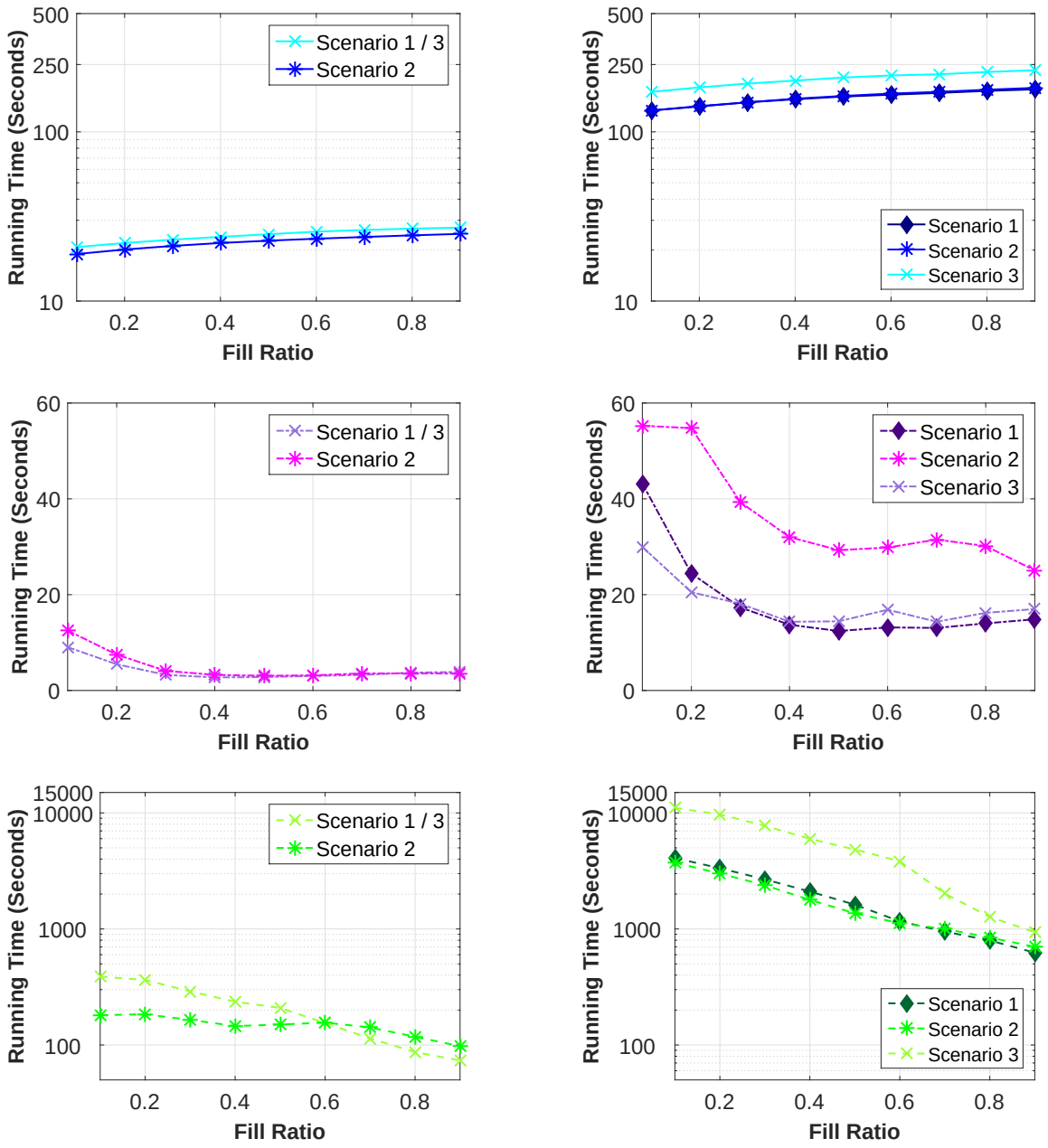

Figure 5.6: Case study 2 - TC (top), MC (middle) and RegEM (bottom) running time for HAR (left) and mHealth (right) w.r.t. f, for Scenario 1,2,3. (Mean of 10 runs.)

3), which is the total number of data streams included in the MHEALTH dataset.

MC on the other hand, demonstrates entirely different behaviour, as highlighted in Figure 5.6 (middle). The running time decreases when moving to fuller matrices, denoting that the computational complexity is strongly associated to the amount of missing data. Moreover, one would expect MC to need higher running times for *collective recovery*, since it is dealing with measurements matrices of greater size. Notwithstanding, Scenarios 1 & 3 for MC are the most efficient also in terms of time. Specifically, instead of performing MC reconstruction individually for each vertically concatenated matrix per modality (Scenario 2), we perform only one recon-

struction call of higher computational complexity (Scenario 3), which is more time-effective. On HAR (middle-left) this observation is more noticeable at low fill ratios, where there are greater volumes of data to be recovered. However, on MHEALTH (middle-right) the difference between the MC running time between Scenario 2 and Scenarios 1 & 3 is prominent.

Finally, the RegEM algorithm presents an exponential increase in its running time when shifting to the overall collective scenario, as illustrated in Figure 5.6 (bottom). Specifically, the cumulative running time of all fill ratios of Scenario 3 is $\simeq 32.5$ minutes on the HAR dataset (bottom-left), whereas on MHEALTH it scales up to 13.2 hours in total (bottom-right). Consequently, despite its superiority in terms of NMSE, we consider RegEM inefficient for large-scale data recovery on commodity hardware.

5.2.6 Impact of MC & TC on Classification Accuracy

In this experimental section, we aim to evaluate the classification accuracy of our proposed framework in the presence of TC and MC reconstructed test measurements. Herein, we display our findings corresponding to the totality of the classifiers we have experimented upon. More analytically:

- Figure 5.7 presents our results regarding the Decision Trees.
- Figure 5.8 shows the classification accuracy of a Euclidean distanced K-NN classifier.
- Figure 5.9 illustrates the performance of a Cosine distanced K-NN classifier.
- Figure 5.10 demonstrates our findings related to an SVM using a Gaussian kernel.
- Figure 5.11 is assigned to the accuracy of an SVM using a Quadratic kernel.

Each figure depicts the accuracy of the corresponding classifier w.r.t. to f, considering MC (left) and TC (right) for all recovery scenarios on HAR (top) and MHEALTH (bottom). To demonstrate the efficiency of our suggested schemes, we also provide a direct comparison of the achieved performance per scenario to the one managed by the classifiers on features extracted from fully-populated structures, i.e. $f = 1$ (green straight curves), which can be considered as ground truth.

With respect to the engaged scenarios for MC reconstruction (left), Scenario 1/3 is shown to be the most effective data structuring technique for the HAR dataset. Specifically, on HAR (top-left), Scenario 1/3 outperforms Scenario 2 at all fill ratios. On MHEALTH, results also report collective scenarios as the most promising ones, alternating between Scenario 2 and Scenario 3 depending on the classifier in use. On MHEALTH (bottom-left) the superiority of the collective scenarios is obvious on extremely under-sampled structures, i.e. $f < 0.4$, whereas as the fill ratio increases, all scenarios perform identically. This outcome suggests that MC can fully utilize the correlation that exists among diverse modalities or diverse devices, even if such correlations are not explicitly encoded into the recovery process.

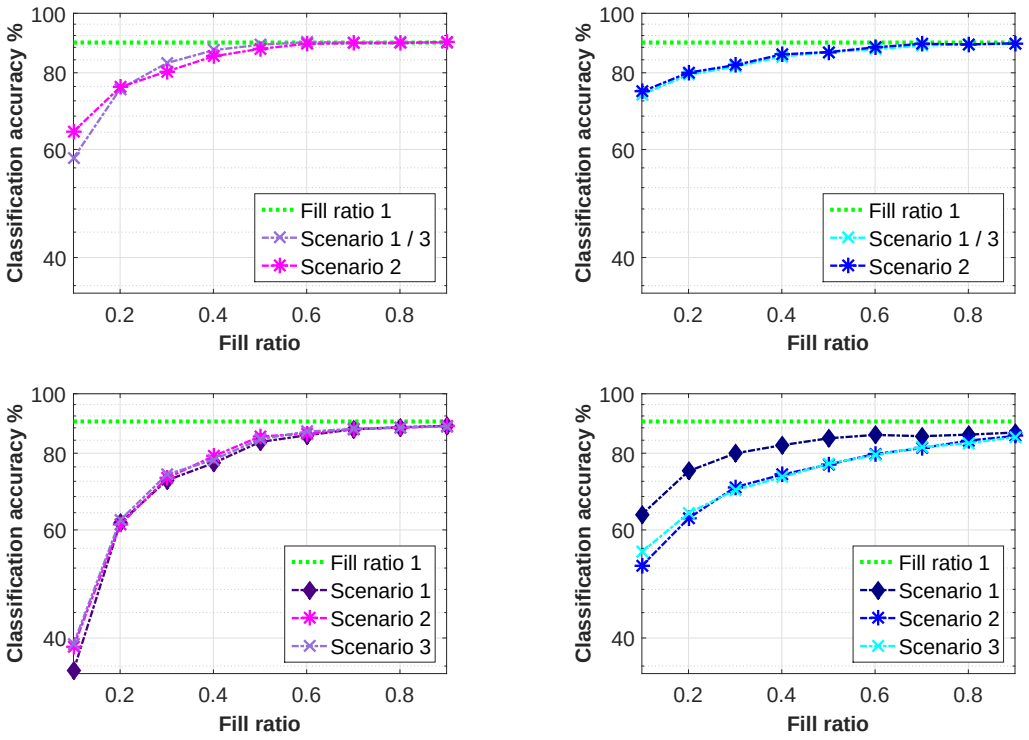

Figure 5.7: Case study 2 - Classification accuracy of Decision Trees on HAR (top) and MHEALTH (bottom) w.r.t. to f, considering ground truth, and MC (left) or TC (right) reconstruction for all scenarios. (Mean of 10 runs.)

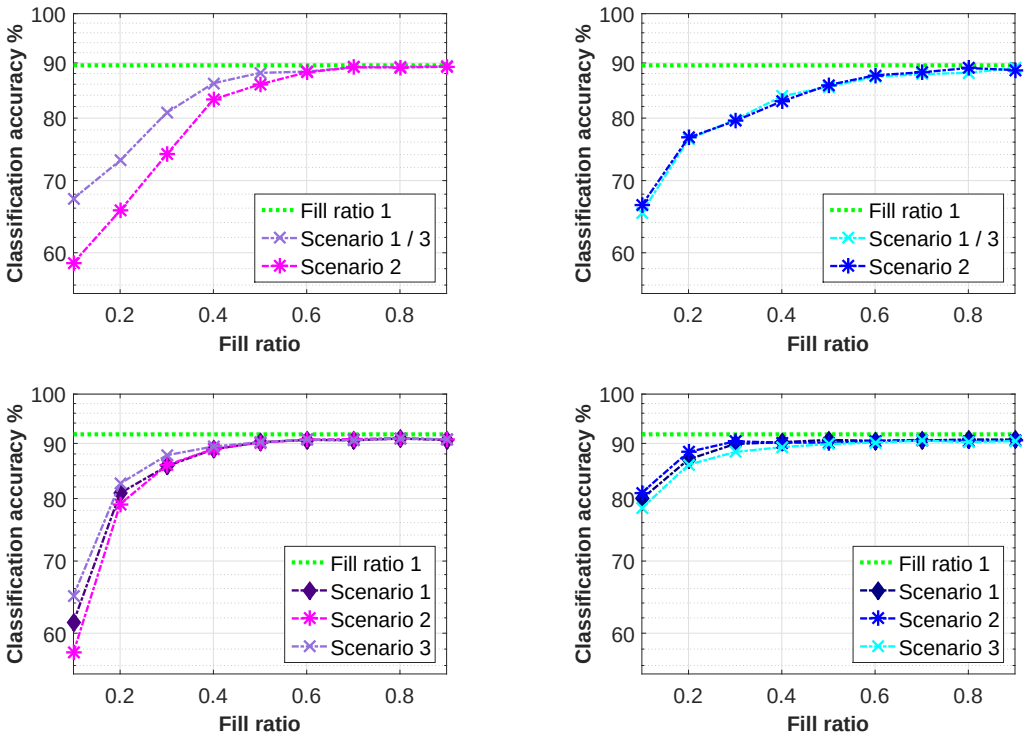

Figure 5.8: Case study 2 - Classification accuracy of Euclidean K-NN on HAR (top) and MHEALTH (bottom) w.r.t. to f, considering ground truth, missing (non-recovered) data, and MC (left) or TC (right) reconstruction for all scenarios. (Mean of 10 runs.)

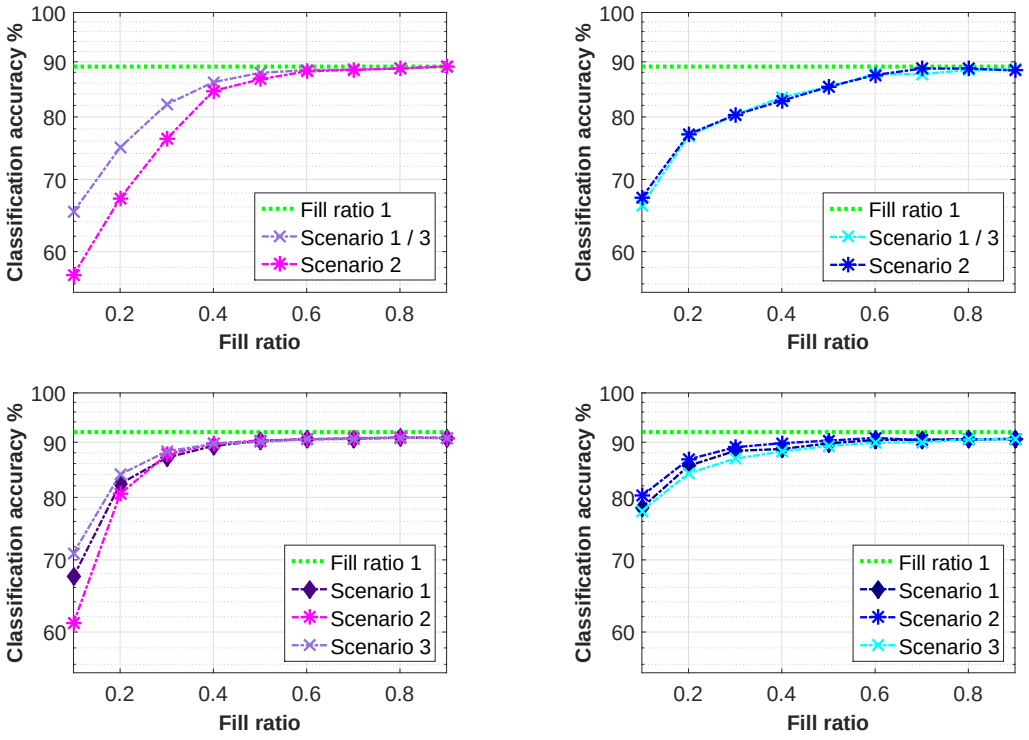

Figure 5.9: Case study 2 - Classification accuracy of Cosine KNN on HAR (top) and MHEALTH (bottom) w.r.t. to f, considering ground truth, and MC (left) or TC (right) reconstruction for all scenarios. (Mean of 10 runs.)

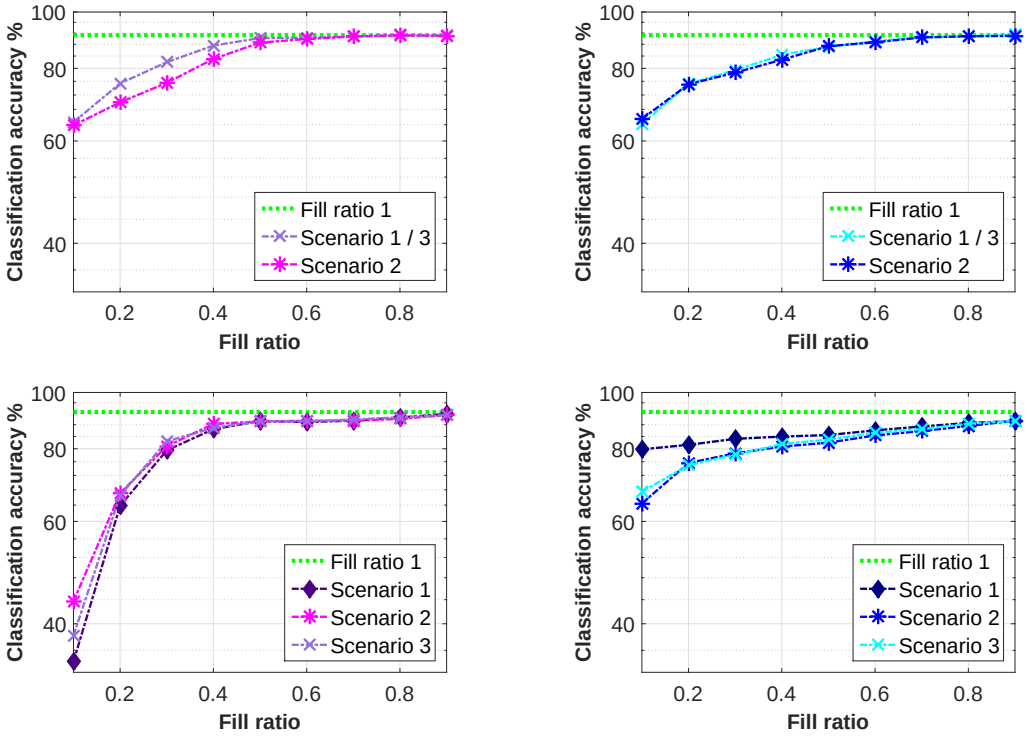

Figure 5.10: Case study 2 - Classification accuracy of Gaussian SVM on HAR (top) and MHEALTH (bottom) w.r.t. to f, considering ground truth, and MC (left) or TC (right) reconstruction for all scenarios. (Mean of 10 runs.)

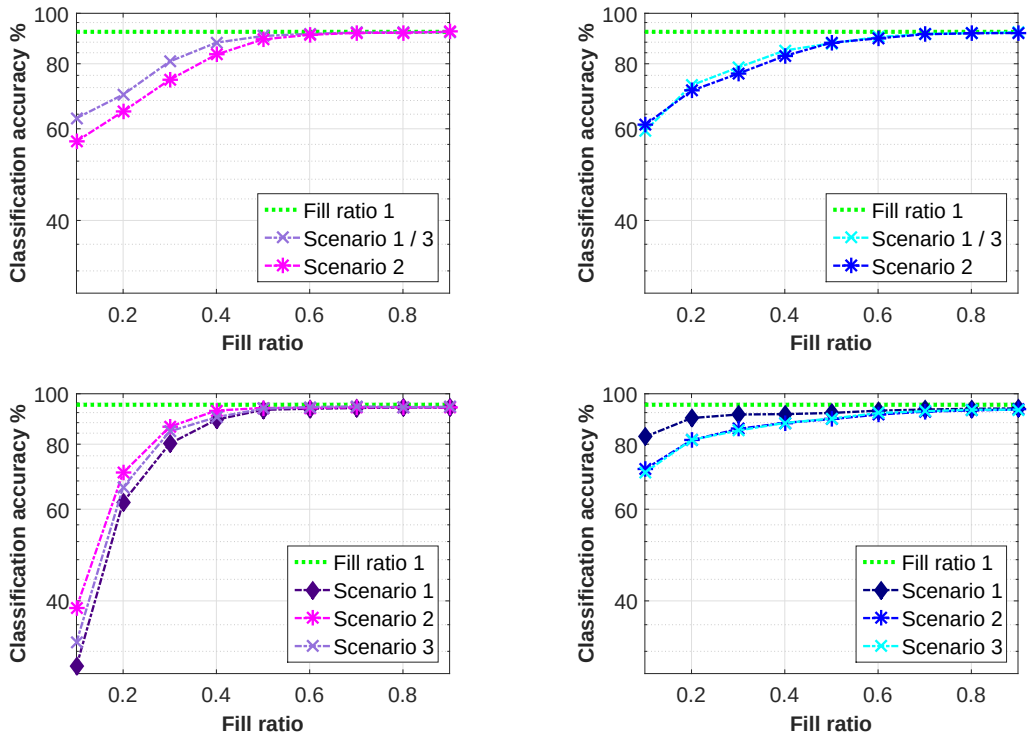

Figure 5.11: Case study 2 - Classification accuracy of Quadratic SVM on HAR (top) and MHEALTH (bottom) w.r.t. to f, considering ground truth, and MC (left) or TC (right) reconstruction for all scenarios. (Mean of 10 runs.)

However, for TC reconstruction (right), experimental results of the reconstruction scenarios are marginal depending on the dataset and the considered classifier. On HAR (top-right), their performance is almost identical for all classifiers in use, but on MHEALTH (bottom-right) there are variations on the prevailing scenario according to the employed classifier. More analytically, on both K-NNs, Scenario 2 slightly outperforms Scenarios 1 & 3 at low fill ratios, i.e. $f < 0.4$. On the other hand, SVMs and Decision Tress clearly report Scenario 1 as the most efficient structuring method for reconstruction. This remark indicates that, as the data extend to diverse sensing devices, the rank of the underlying tensor increases, thus leading to inferior collective recovery for TC for the majority of the employed classifiers.

Considering MC (left) as compared to TC (right), the results demonstrate the latter as the most efficient recovery technique, on extremely under-sampled structures, meaning $f \leq 0.3$. The superiority of TC is more noticeable for the MHEALTH dataset. Thereby, there is at least one recovery scenario, where the classification accuracy in the presence of TC-recovered data exceeds 80% for only 10% of the measurements available, for almost all of the classifiers in use. This outcome contradicts our aforementioned intuition on the performance of TC, depending on the cumulative NMSE metric. Such a result can be explained, though, by once again examining the existing correlations among the data. The Singular Value Decomposition of each data stream revealed that some data channels are rather low-rank, while others are not. TC can recover

the strongly correlated streams with high precision, but fails for higher-ranked data, resulting in an increased total NMSE than that of MC. Nonetheless, the superior classification accuracy achieved by TC, especially on MHEALTH, suggests that the accurately predicted low-rank data streams, i.e. accelerometers, are much more dominant and statistically significant than the poorly recovered less correlated ones, gyroscopes and magnetometers.

Moreover, Table 5.2 shows the direct numerical comparison of the classification accuracy managed in the presence of missing data, to the one achieved by Scenario 3 MC-based reconstruction for $f = 0.4$ on an indicative subset of classifiers per dataset. The substantial improvement in the performance of all classifiers in use highlights the significance as well as the effectiveness of recovery in truly lost or unavailable measurements. Furthermore, Cosine K-NN is shown as the most missing values-tolerant classifier, unlike Gaussian SVM which performs poorly under such circumstances.

Table 5.2: Case study 2 - Comparison of classification accuracy of the proposed classification framework for non-recovered vs. MC-recovered (Scenario 3) data.

	f = 1	**Missing Data f = 0.4**	**MC-Recovered Data f = 0.4**
HAR Tree	90.0	76.2	86.4
HAR Eucl. K-NN	90.2	80.4	86.2
HAR Cosine K-NN	89.5	82.5	87
MHEALTH Gaussian SVM	93.6	58.6	87.2
MHEALTH Quadr. SVM	95.8	78.4	89.5

Finally, Figure 5.12 illustrates the classification accuracy of an indicative subset of the employed classifiers per dataset for Scenario 3 MC & TC with respect to f, compared to the one achieved from fully-populated structures.

Regarding MC (left), we notice that on both datasets the utilized classifiers manage accuracy of only $1 - 2\%$ lower than the ground truth at $f = 0.5$, whereas at $f = 0.6$ they reach optimal performance. Thus, a significant observation related to the overall performance of MC within our proposed framework is that, near-optimal classification accuracy is feasible even on extremely under-sampled matrices, where 50% or more of their observations are recovered.

As far as TC is concerned (right), ground-truth accuracy is nearly reached at $f = 0.6$ on both datasets. Moreover, the above discussed high quality results of TC at low fill ratios of Euclidean K-NN are extended to the rest of the considered classifiers. Regarding $0.3 \leq f \leq 0.6$, MC slightly outperforms TC, whereas at high fill ratios they exhibit similar performance. A noteworthy remark is the accuracy of $\simeq 85\%$ achieved by Cosine K-NN for $f = 0.2$. This is a very promising outcome highlighting the effectiveness of TC structuring and recovery method on multi-modal and extremely under-sampled tensors, containing just 20% of their measurements.

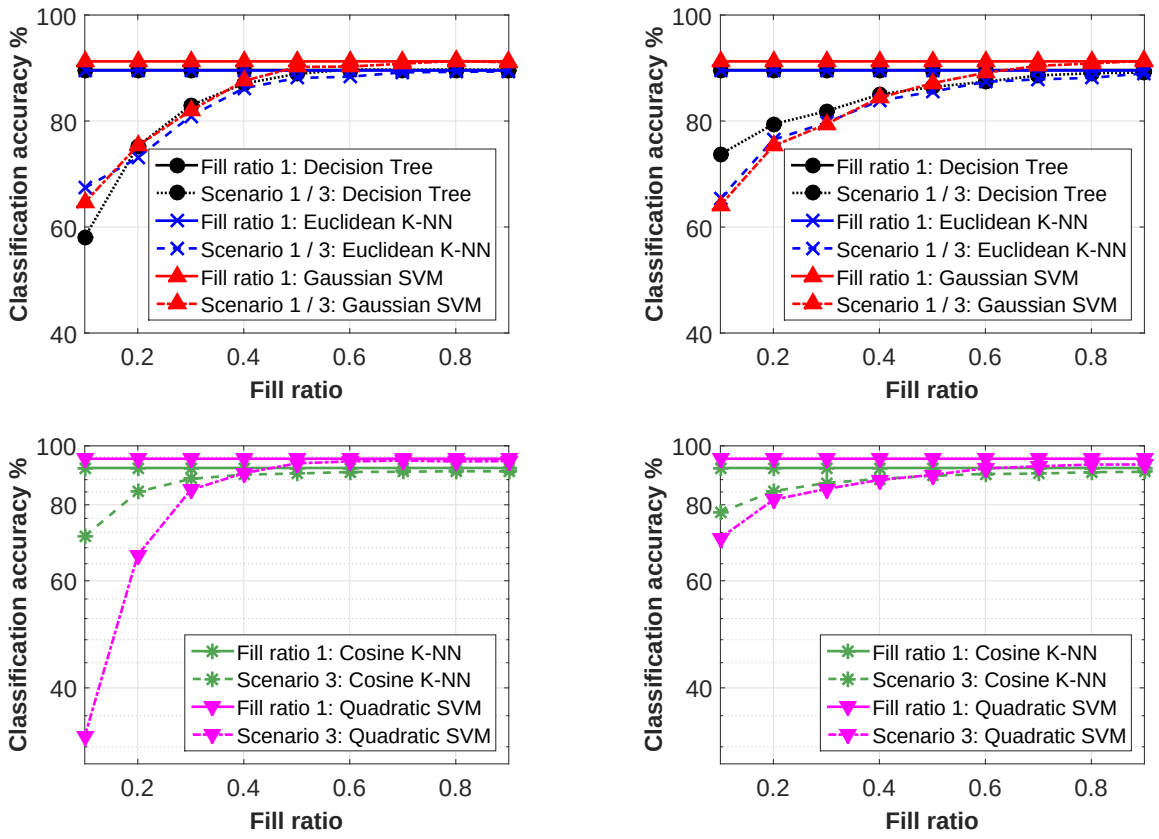

Figure 5.12: Case study 2 - Classification accuracy for MC (left) and TC (right) reconstruction w.r.t. to f, considering Scenario 3 for HAR (top) and MHEALTH (bottom). (Mean of 10 runs.)

5.2.7 Different Sampling per Device

The objective of this experiment is to address more realistic scenarios, in terms of the occurrence of missing values. Let us consider a number of sensing devices attached on a subject. With high probability, energy limitations could easily lead to different volumes of missing values per device. Our goal is to examine the impact of these conditions on our proposed framework.

We implemented such scenarios by adjusting the sampling step of the preprocessing module accordingly. The experiment was conducted on the MHEALTH dataset, where there are three sensing devices and the utilized recovery methods were MC and TC. Missing values were introduced by performing different sampling per device, e.g. f=0.3 available chest data, f=0.5 for the right wrist, and f=0.7 for the data acquired from the left ankle. Table 5.3 summarizes our experimental findings regarding the SVM classifier with a Gaussian kernel for nine different sampling sets, grouped per three mean values of fill ratio, namely f=0.2, f=0.5, and f=0.8.

A first observation concerning the classification accuracy in the presence of missing values, is that the device attached to the right wrist provides the classifiers with the most "distinguishable" data among all of the considered mean f values. Higher fill ratios on the right wrist, followed by less left ankle data and even more limited data originating from the chest, are shown as the

Table 5.3: Case study 2 - Classification accuracy of Gaussian SVM considering different sampling per device for MC/TC-recovered data.

Ground Truth, f=1	Sampling per Device Ch=Chest, LA=Left Ankle, RW=Right Wrist		Missing Values	Scenario 1 MC / TC	Scenario 2 MC / TC	Scenario 3 MC / TC
92.53	Mean f=0.2	Ch=0.1, LA=0.2, RW=0.3	**46.24**	60.37 / **82.00**	**64.27** / 75.26	59.81 / 74.77
		Ch=0.3, LA=0.2, RW=0.1	45.79	58.42 / 80.44	61.78 / 74.58	60.77 / 75.14
		Ch=0.1, LA=0.1, RW=0.4	42.12	44.51 / 79.58	44.49 / 69.05	38.80 / 69.82
	Mean f=0.5	Ch=0.3, LA=0.5, RW=0.7	**66.51**	**88.47** / **84.82**	84.70 / 81.29	83.87 / 81.94
		Ch=0.7, LA=0.5, RW=0.3	65.29	86.54 / 84.73	84.70 / 83.64	85.67 / 84.20
		Ch=0.4, LA=0.4, RW=0.7	58.26	87.03 / 83.81	87.19 / 80.72	87.00 / 80.91
	Mean f=0.8	Ch=0.6, LA=0.8, RW=1	**87.61**	89.96 / 87.05	90.81 / 87.20	90.75 / 86.89
		Ch=1, LA=0.8, RW=0.6	86.42	90.16 / **88.69**	90.92 / 88.12	**91.25** / 88.41
		Ch=0.7, LA=0.7, RW=1	85.39	89.55 / 86.70	90.62 / 85.89	90.50 / 85.75

most advantageous data acquisition scenario in this framework.

Moreover, our remarks regarding the need for data recovery are extended from uniform to non-uniform sampling per device experiments. For each sampling set considered, there is at least one reconstruction scheme, that significantly outperforms the classification accuracy achieved at the presence of non-recovered data. The most prominent paradigm is the performance of Scenario 1 TC at mean $f = 0.2$, where the accomplished accuracy is $\simeq 35\%$ higher than the ground truth.

It is also shown, that there is not a specific sampling set per mean fill ratio producing the optimal accuracy. For mean $f = 0.2$ and $f = 0.5$, right wrist is indicated as the most data-dominant body part, whereas for mean $f = 0.8$ increasing chest data, results in the best performance. With respect to the structuring scenarios, TC reports Scenario 1 as the most promising, since it prevails at all mean fill ratios (green bold), while the respective results are marginal for MC depending on the distribution of the sampling set (red bold). Finally, our experimental findings do not indicate a specific sensing body-part as more low rank, thus more "appealing" to MC or TC reconstruction.

5.2.8 Effects of Sub-sampling on Battery Consumption of Shimmer Sensing Platforms

Our goal is to address the limitations leading to missing measurements from a different, yet more optimistic, perspective in terms of energy consumption. We argue that the deliberate introduction of missing values through sub-sampling, combined with an accurate reconstruction method, could actually be used as an energy efficient data acquisition and communication approach, resulting in favour of the lifetime of the involved devices.

To facilitate exposition, we experimented on Shimmer3 sensing platforms, devices broadly used for capturing human activity. We employed the ShimmerCapture host application to monitor and log data on a Windows 8 laptop. Our test-bed consisted of four Shimmer3 platforms capturing 3-axial accelerometer, gyroscope, magnetometer and battery voltage data, in accor-

dance to the available physical data streams provided by the MHEALTH dataset. Each device operated at a different sampling rate s and streamed the obtained data over Bluetooth radio. The recorded voltage status of the lithium polymer 3.7V - 450mAh Shimmer battery after 9 hours of operation per device is illustrated at Table 5.13. Note that, the voltage values are mapped to the corresponding remaining battery capacity according to the Shimmer3 documentation [68].

Table 5.4: Case study 2 - Shimmer3 battery status after 9 hrs of use.

	Device 1	**Device 2**	**Device 3**	**Device 4**
Sampling Rate s(Hz)	50	100	256	512
Voltage (mV)	3710	3668	3635	3625
Remaining Capacity %	25.3	15.5	7.5	5.3

Now, let us consider that sampling at $s = 512$ Hz produces a fully-populated data matrix of $f = 1$. Reducing the sampling rate to $s = 256$ would produce an under-sampled dataset of $f = 0.5$. With this consideration in mind, Figure 5.13 illustrates a direct mapping of the classification accuracy achieved by Cosine K-NN at four different fill ratios, for Scenario 3 TC-recovered data, with respect to the remaining battery capacity.

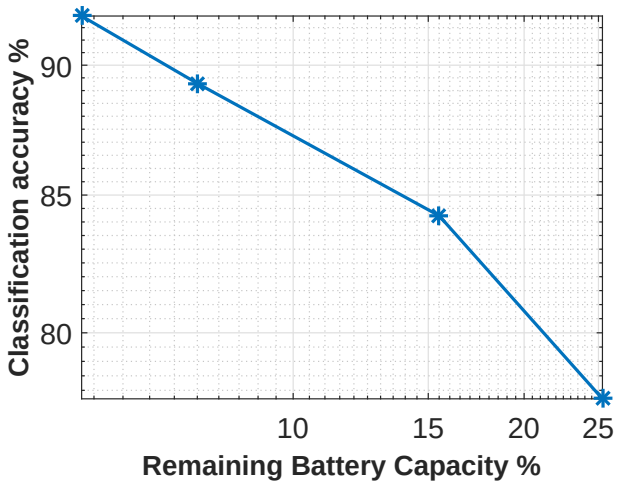

Figure 5.13: Case study 2 - Classification accuracy of Cosine K-NN for TC-Scenario 3 reconstruction w.r.t. the remaining battery capacity of a Shimmer3 platform after 9 hours of use.

Results show that reducing the sampling rate can in fact positively affect the lifetime of the device, especially at the data rates commonly used for Human Activity Recognition, i.e. $50-100$ Hz. Specifically, reducing the sampling rate from 100 to 50 Hz, provides us with a capacity gain of about 10%, which is translated into over 1 hour of extra lifetime for the underlying platform, at the cost of just $\simeq 2.5\%$ classification accuracy. This is a very useful outcome showing us that the considered recovery methods can in fact "create" truly or deliberately lost measurements.

Chapter 6

Conclusions and future work

6.1 Concluding remarks

In this thesis, we have addressed the problem of estimating missing measurements in Wireless Sensor Networks by employing novel mathematical techniques for data structuring and reconstruction, while experimenting on two different application domains, namely, Smart Water Management and Human Activity Recognition.

On the first case study, we have investigated the application of the theory of Matrix Completion for the estimation of missing measurements in a water treatment plan. We have considered three scenarios of paramount importance for the current operation of a WSN-based monitoring paradigm; namely recovery from missing entries, single-device vs. collective recovery of measurements matrices, and temporal super-resolution. Based on our experimental findings, we were able to infer that MC is a viable approach for estimating missing measurements, where such missing measurements are either attributed to lost/unobserved measurements or, in the case of super-resolving, to non-acquired data. Additionally, collective recovery has been proved to be able to achieve better reconstruction than single-device recovery.

Regarding the second application field, we have investigated the effects of various 2D and 3D data structuring and reconstruction methods on a proposed physical activity classification framework. We have experimented on two publicly available datasets, with different dynamics in the variety of activities, sensing modalities, and devices. Based on our experimental findings, Matrix Completion comes out as a very stable and fast reconstruction method, exhibiting high performance with less than 50% of the available measurements. Moreover, Tensor Completion achieves most promising results on multi-modal data acquired by multiple devices, performing at near-optimal levels even with more than 80% of the data lost or unavailable. Results, once again, outline the effectiveness of collective recovery for MC, as it better exploits the correlations introduced among the data through our proposed Hankelization structuring process. Finally, sub-sampling is demonstrated as a significant factor directly contributing to the increase of the network lifetime.

6.2 Future directions

Our proposed methods and modules on both domains are very generic by nature. In the future they can be easily applied to the recovery of other types of WSN-based measurements with an emphasis on IoT-related fields within infrastructure systems, such as smart homes, intelligent transportation, and the smart grid.

Additionally, a very interesting possible future research aspect could be related to the embedding of our proposed modules into energy-efficient WSN-sampling architectures. This could be accomplished by studying the quantification of the trade-off between the energy gain achieved by sub-sampling at the sensing device and the consumption of resources used for the reconstruction of the missing data collected by the host.

Regarding WSN-based classification problems, future work could involve the implementation of an extra module, that of low-rank data streams selection, where only the most correlated data will be utilized on the subsequent reconstruction. Thus, the considered MC an TC recovery methods, which assume low-rankness of the underlying data, will be able to achieve even better recovery, leading to improved classification accuracy. For the rest of the channels, classification could rely on the available measurements or to reconstructed ones using an alternative method of imputation.

Another future direction for research could include the incorporation of feature selection algorithms into our proposed classification framework. Since the extracted features are directly related to the quality of the reconstructed raw values, the selection of the most representative features per imputation method, could have a significant effect on the performance of the system.

Finally, the real time aspect of the considered methods should definitely by addressed in the future. For efficient real time recovery, data should be segmented to smaller windows in order to enable the construction of matrices or tensors of lower dimensionality. The framework must also involve a "memory" module, shifting the data structures through time, so that at the current time instance only a small percentage of data needs to be recovered, while the rest of the data were recovered and stored at previous time instances.

Appendix A

Analysis on setting the optimum parameters for the Hankelization module

The key parameters for the Hankelization module are the window size I_2 and the overlap l, as discussed in Section 3.2. They dictate the form and shape of the resulting Hankel structure, as well as the redundancies introduced among the data. In the experiments presented in Chapter 5, these parameters where set to $I_2 = 128$ samples and $l = 50\% \times I_2$ for the available time-series data streams per dataset. These values were not set arbitrarily. They were drawn through the experimentation presented in this Appendix.

In the following experiments, we considered the six data streams available in HAR dataset, namely the accelerometer and the gyroscope measurements on x,y, and z axis. Our goal was to assess the performance of MC on various Hankel matrices, resulting from different values of I_2 and l. Note that recovery takes place on each Hankel matrix locally, i.e. per stream of data.

The first experimental set, examines the NMSE rising from matrices created by a specified value for the window size, i.e. $I_2 = 128$ samples, and various values for the overlap parameter, regarding several fill ratios from $f = 0.1$ to 0.9 with a step size of 0.1. The results corresponding to the totality of the data streams within HAR dataset, are illustrated in Figure A.1. It is thereby demonstrated that the optimum results, meaning the lower NMSE, is derived at overlap $l = 64 = 50\% \times I_2$.

The next experiments are in direct concordance to the aforementioned ones, nevertheless, at this case the varying parameter is the window size, while the overlap remains constant and set to $l = 50\% \times I_2$. The results, presented in Figure A.2 show that the majority of streams report the value $I_2 = 128$, as the optimum value with reference to the window size.

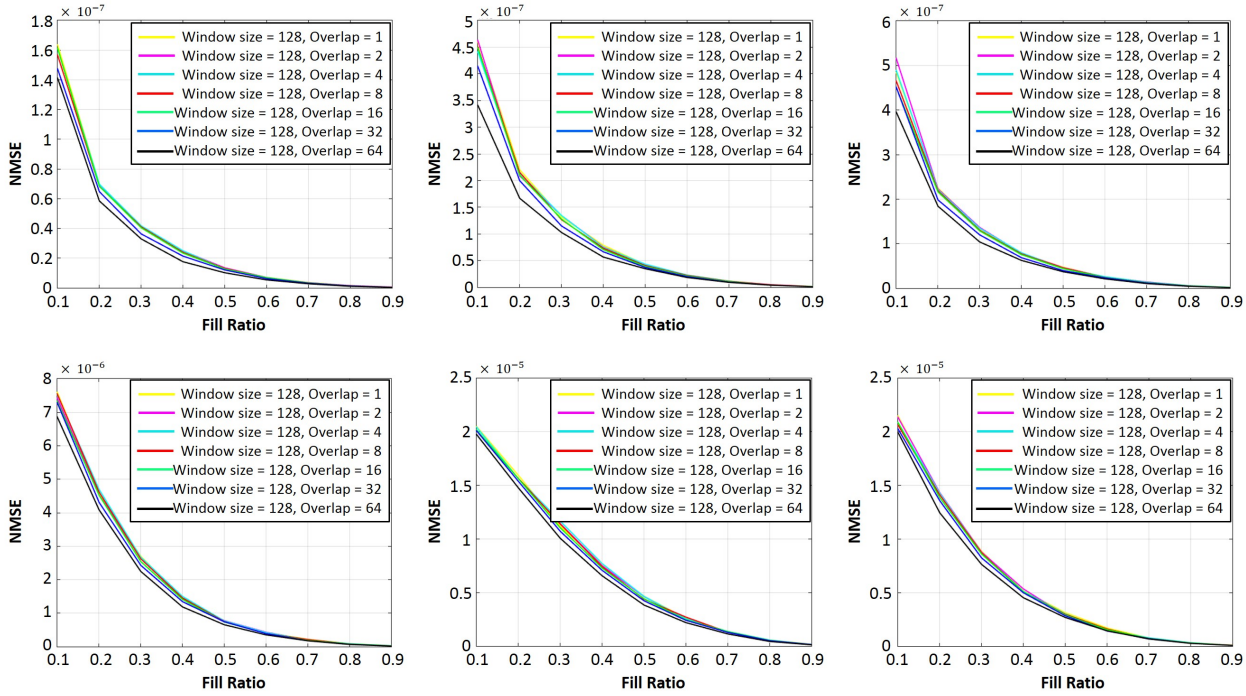

Figure A.1: NMSE w.r.t. f for window size set to 128 samples and various overlaps, regarding the x(left)/y(middle)/z(right)-axis accelerometers (top) and gyroscopes (bottom) in HAR dataset.

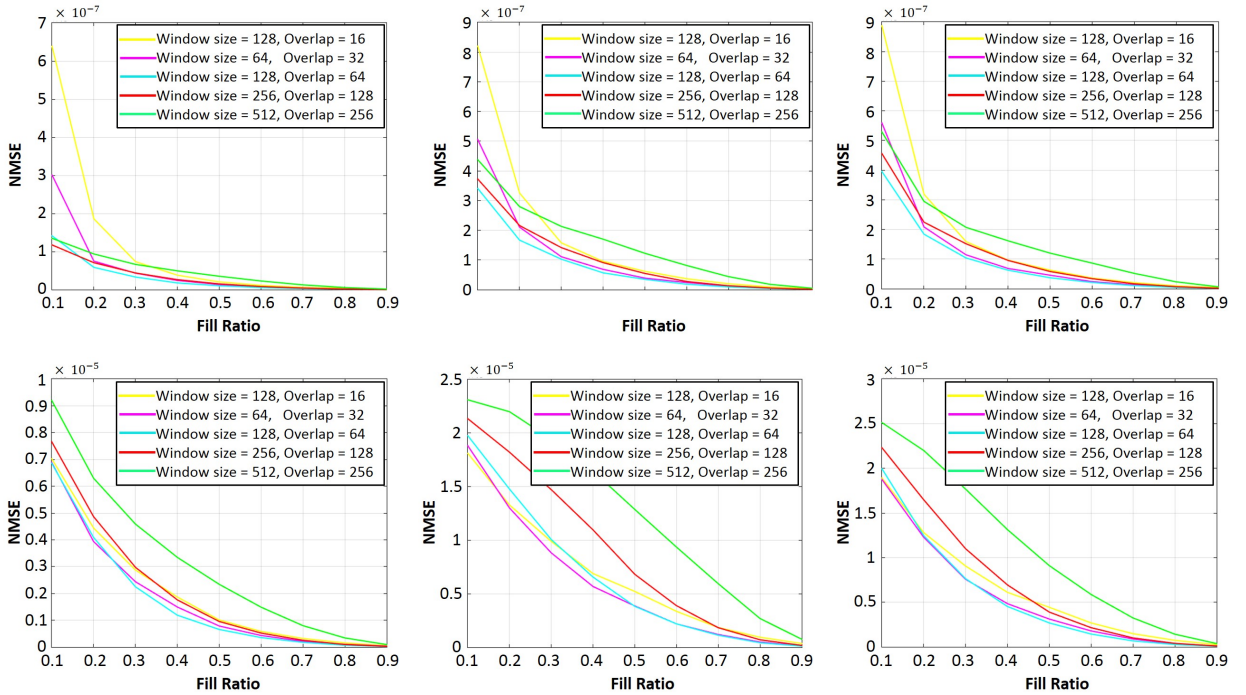

Figure A.2: NMSE w.r.t. f for overlap set to 50% of the window size and various sizes of windows, regarding the x(left)/y(middle)/z(right)-axis accelerometers (top) and gyroscopes (bottom) in HAR dataset.

Bibliography

[1] i. t. I. M. S. B. Wireless Sensor Networks project team, "Internet of things: Wireless sensor networks," *6-th white paper*, 2014.

[2] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[3] L. Mainetti, L. Patrono, and A. Vilei, "Evolution of wireless sensor networks towards the internet of things: A survey," in *Software, Telecommunications and Computer Networks (SoftCOM), 2011 19th International Conference on.* IEEE, 2011, pp. 1–6.

[4] Smart water networks. [Online]. Available: http://www.swan-forum.com/

[5] A. Hauser and F. Roedler, "Interoperability: the key for smart water management," *Water Science and Technology: Water Supply*, vol. 15, no. 1, pp. 207–214, 2015.

[6] C. D. Beal and J. Flynn, "Toward the digital water age: Survey and case studies of australian water utility smart-metering programs," *Utilities Policy*, vol. 32, pp. 29–37, 2015.

[7] G. Xu, G. Q. Huang, and J. Fang, "Cloud asset for urban flood control," *Advanced Engineering Informatics*, vol. 29, no. 3, pp. 355–365, 2015.

[8] S. S. Mane and M. K. Mokashi, "Survey on real-time flash-flood monitoring, alerting and forecasting system using data mining and wireless sensor network," *International Journal*, vol. 2, no. 12, 2014.

[9] T. Robles, R. Alcarria, D. Martín, A. Morales, M. Navarro, R. Calero, S. Iglesias, and M. López, "An internet of things-based model for smart water management," in *Advanced Information Networking and Applications Workshops (WAINA), 2014 28th International Conference on.* IEEE, 2014, pp. 821–826.

[10] J. Harou, P. Garrone, A. Rizzoli, A. Maziotis, A. Castelletti, P. Fraternali, J. Novak, R. Wissmann-Alves, and P. Ceschi, "Smart metering, water pricing and social media to stimulate residential water efficiency: Opportunities for the smarth2o project," *Procedia Engineering*, vol. 89, pp. 1037–1043, 2014.

[11] T. Tsung-Te Lai, W.-J. Chen, K.-H. Li, P. Huang, and H.-H. Chu, "Triopusnet: automating wireless sensor network deployment and replacement in pipeline monitoring," in *Information Processing in Sensor Networks (IPSN), 2012 ACM/IEEE 11th International Conference on.* IEEE, 2012, pp. 61–71.

[12] I. Stoianov, L. Nachman, A. Whittle, S. Madden, and R. Kling, "Sensor networks for monitoring water supply and sewer systems: Lessons from boston," in *Proceedings of the 8th Annual Water Distribution Systems Analysis Symposium*, 2006, pp. 1–17.

[13] B. O'Flynn, F. Regan, A. Lawlor, J. Wallace, J. Torres, and C. O'Mathuna, "Experiences and recommendations in deploying a real-time, water quality monitoring system," *Measurement Science and Technology*, vol. 21, no. 12, p. 124004, 2010.

[14] N. Metje, D. N. Chapman, D. Cheneler, M. Ward, and A. M. Thomas, "Smart pipes-instrumented water pipes, can this be made a reality?" *Sensors*, vol. 11, no. 8, pp. 7455–7475, 2011.

[15] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, and P. J. Havinga, "A survey of online activity recognition using mobile phones," *Sensors*, vol. 15, no. 1, pp. 2059–2085, 2015.

[16] C. Torres-Huitzil and A. Alvarez-Landero, "Accelerometer-based human activity recognition in smartphones for healthcare services," in *Mobile Health*. Springer, 2015, pp. 147–169.

[17] A. Solanas, C. Patsakis, M. Conti, I. S. Vlachos, V. Ramos, F. Falcone, O. Postolache, P. A. Pérez-Martínez, R. Di Pietro, D. N. Perrea *et al.*, "Smart health: a context-aware health paradigm within smart cities," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 74–81, 2014.

[18] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013.

[19] M. Zhang and A. A. Sawchuk, "A customizable framework of body area sensor network for rehabilitation," in *2009 2nd International Symposium on Applied Sciences in Biomedical and Communication Technologies*. IEEE, 2009, pp. 1–6.

[20] S. Patel, H. Park, P. Bonato, L. Chan, and M. Rodgers, "A review of wearable sensors and systems with application in rehabilitation," *Journal of neuroengineering and rehabilitation*, vol. 9, no. 1, p. 1, 2012.

[21] A. Rasekh, C.-A. Chen, and Y. Lu, "Human activity recognition using smartphone," *arXiv preprint arXiv:1401.8212*, 2014.

[22] 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," in *International Workshop on Ambient Assisted Living*. Springer, 2012, pp. 216–223.

[23] G. Filios, S. Nikoletseas, C. Pavlopoulou, M. Rapti, and S. Ziegler, "Hierarchical algorithm for daily activity recognition via smartphone sensors," in *Internet of Things (WF-IoT), 2015 IEEE 2nd World Forum on*. IEEE, 2015, pp. 381–386.

[24] A. Burns, B. R. Greene, M. J. McGrath, T. J. O'Shea, B. Kuris, S. M. Ayer, F. Stroiescu, and V. Cionca, "Shimmer-a wireless sensor platform for noninvasive biomedical research," *IEEE Sensors Journal*, vol. 10, no. 9, pp. 1527–1534, 2010.

[25] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Computing Surveys (CSUR)*, vol. 46, no. 3, p. 33, 2014.

[26] U. Maurer, A. Smailagic, D. P. Siewiorek, and M. Deisher, "Activity recognition and monitoring using multiple sensors on different body positions," in *International Workshop on Wearable and Implantable Body Sensor Networks (BSN'06)*. IEEE, 2006, pp. 4–pp.

[27] M. Zhang and A. A. Sawchuk, "A feature selection-based framework for human activity recognition using wearable multimodal sensors," in *Proceedings of the 6th International Conference on Body Area Networks*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2011, pp. 92–98.

[28] L. Pan, J. Li *et al.*, "K-nearest neighbor based missing data estimation algorithm in wireless sensor networks," *Wireless Sensor Network*, vol. 2, no. 02, p. 115, 2010.

[29] L. Kong, M. Xia, X.-Y. Liu, G. Chen, Y. Gu, M.-Y. Wu, and X. Liu, "Data loss and reconstruction in wireless sensor networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 11, pp. 2818–2828, 2014.

[30] L. Pan, H. Gao, H. Gao, and Y. Liu, "A spatial correlation based adaptive missing data estimation algorithm in wireless sensor networks," *International Journal of Wireless Information Networks*, vol. 21, no. 4, pp. 280–289, 2014.

[31] Y. Wang, J. Wang, and H. Li, "An interpolation approach for missing context data based on the time-space relationship and association rule mining," in *2011 Third International Conference on Multimedia Information Networking and Security*. IEEE, 2011, pp. 623–627.

[32] A. Farhangfar, L. Kurgan, and J. Dy, "Impact of imputation of missing values on classification error for discrete data," *Pattern Recognition*, vol. 41, no. 12, pp. 3692–3705, 2008.

[33] Y. Xu, R. Hao, W. Yin, and Z. Su, "Parallel matrix factorization for low-rank tensor completion," *arXiv preprint arXiv:1312.1254*, 2013.

[34] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[35] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, vol. 52, no. 3, pp. 471–501, 2010.

[36] R. Meka, P. Jain, C. Caramanis, and I. S. Dhillon, "Rank minimization via online learning," in *Proceedings of the 25th International Conference on Machine learning*. ACM, 2008, pp. 656–663.

[37] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *Information Theory, IEEE Transactions on*, vol. 56, no. 5, pp. 2053–2080, 2010.

[38] M. Fazel, H. Hindi, and S. P. Boyd, "Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices," in *American Control Conference, 2003. Proceedings of the 2003*, vol. 3. IEEE, 2003, pp. 2156–2162.

[39] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[40] 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.

[41] S. Ma, D. Goldfarb, and L. Chen, "Fixed point and bregman iterative methods for matrix rank minimization," *Mathematical Programming*, vol. 128, no. 1-2, pp. 321–353, 2011.

[42] K.-C. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pacific Journal of Optimization*, vol. 6, no. 615-640, p. 15, 2010.

[43] Y.-J. Liu, D. Sun, and K.-C. Toh, "An implementable proximal point algorithmic framework for nuclear norm minimization," *Mathematical programming*, vol. 133, no. 1-2, pp. 399–436, 2012.

[44] Z. Liu and L. Vandenberghe, "Interior-point method for nuclear norm approximation with application to system identification," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 3, pp. 1235–1256, 2009.

[45] M. Roughan, Y. Zhang, W. Willinger, and L. Qiu, "Spatio-temporal compressive sensing and internet traffic matrices (extended version)," *IEEE/ACM Transactions on Networking*, vol. 20, no. 3, pp. 662–676, 2012.

[46] 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," *IEEE Transactions on Wireless Communications*, vol. 12, no. 2, pp. 850–861, 2013.

[47] 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.

[48] C. Chen, B. He, and X. Yuan, "Matrix completion via an alternating direction method," *IMA Journal of Numerical Analysis*, vol. 32, no. 1, pp. 227–245, 2012.

[49] D. P. Bertsekas, *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[50] C. J. Hillar and L.-H. Lim, "Most tensor problems are np-hard," *Journal of the ACM (JACM)*, vol. 60, no. 6, p. 45, 2013.

[51] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 208–220, 2013.

[52] S. Gandy, B. Recht, and I. Yamada, "Tensor completion and low-n-rank tensor recovery via convex optimization," *Inverse Problems*, vol. 27, no. 2, p. 025010, 2011.

[53] B. Huang, C. Mu, D. Goldfarb, and J. Wright, "Provable models for robust low-rank tensor completion," *Pacific Journal of Optimization*, vol. 11, no. 2, pp. 339–364, 2015.

[54] M. Fazel, T. K. Pong, D. Sun, and P. Tseng, "Hankel matrix rank minimization with applications to system identification and realization," *SIAM Journal on Matrix Analysis and Applications*, vol. 34, no. 3, pp. 946–977, 2013.

[55] B. Li, O. I. Camps, and M. Sznaier, "Cross-view activity recognition using hankelets," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1362–1369.

[56] S. Yang, K. Kalpakis, C. F. Mackenzie, L. G. Stansbury, D. M. Stein, T. M. Scalea, and P. F. Hu, "Online recovery of missing values in vital signs data streams using low-rank matrix completion," in *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, vol. 1. IEEE, 2012, pp. 281–287.

[57] D. Alonso-Romän, E. Celada-Funes, C. Asensio-Marco, and B. Beferull-Lozano, "Improving reliability and efficiency of communications in wsns under high traffic demand," in *2013 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2013, pp. 268–273.

[58] T. Winter, "Rpl: Ipv6 routing protocol for low-power and lossy networks," 2012.

[59] I. A. Gheyas and L. S. Smith, "A neural network-based framework for the reconstruction of incomplete data sets," *Neurocomputing*, vol. 73, no. 16, pp. 3039–3065, 2010.

[60] T. Schneider, "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values," *Journal of Climate*, vol. 14, no. 5, pp. 853–871, 2001.

[61] T. Razzaghi, O. Roderick, I. Safro, and N. Marko, "Fast imbalanced classification of healthcare data with missing values," in *Information Fusion (Fusion), 2015 18th International Conference on*. IEEE, 2015, pp. 774–781.

[62] Har dataset. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

[63] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in *ESANN*, 2013.

[64] O. Banos, R. Garcia, J. A. Holgado-Terriza, M. Damas, H. Pomares, I. Rojas, A. Saez, and C. Villalonga, "mhealthdroid: a novel framework for agile development of mobile health applications," in *Ambient Assisted Living and Daily Activities*. Springer, 2014, pp. 91–98.

[65] Mhealth dataset. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/MHEALTH+Dataset

[66] O. Banos, C. Villalonga, R. Garcia, A. Saez, M. Damas, J. A. Holgado-Terriza, S. Lee, H. Pomares, and I. Rojas, "Design, implementation and validation of a novel open framework for agile development of mobile health applications," *Biomedical engineering online*, vol. 14, no. S2, pp. 1–20, 2015.

[67] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *International Conference on Pervasive Computing*. Springer, 2004, pp. 1–17.

[68] Shimmer user manual. [Online]. Available: http://www.shimmersensing.com/images/uploads/docs/Shimmer_User_Manual_rev3m.pdf