Application of Tensor and Matrix Completion on Environmental Sensing Data

Michalis Giannopoulos[1,2], Sofia Savvaki[1,2],
Grigorios Tsagkatakis[1], and Panagiotis Tsakalides[1,2] *

1- Institute of Computer Science
Foundation for Research and Technology Hellas (FORTH)
Heraklion, 70013, Greece

2- Department of Computer Science
University of Crete
Heraklion, 70013, Greece

Abstract. As environmental resources utilization becomes more and more crucial, Wireless Sensor Networks (WSNs) are introduced in order to capture the variation of diverse parameters. However, limitations such as network connectivity, power consumption, and storage capacity lead to missing measurements from such networked sensors. To address this problem, we investigate the potential of recovering high dimensional environmental signals from small sets of observations. To account for the dimensionality of the data, we invoke tensor modelling and we propose a low-rank tensor recovery formulation. Experimental results using real WSN data from an indoor industrial environment as well as from an outdoor natural environment demonstrate that the estimation of missing measurements is much better addressed when structural information is considered.

1 Introduction

Monitoring of environmental signals is critical for understanding the dynamics and for controlling chemical and other processes. For instance, as water resource utilization is becoming more and more crucial, the deployment of Smart Water Networks (SWNs) is of utmost importance in our attempt to efficiently organize these resources. To support the necessity of continuous and dependable monitoring, Wireless Sensor Network (WSN) technologies are introduced. While WSNs can provide automated, robust, and easy-to-deploy monitoring, they are also characterized by severe limitations including limited power supply, packet loses, and noisy measurements.

From a WSN perspective, increased sampling rates can lead to more statistically robust results, at the cost of a dramatic decrease in network lifetime, since acquisition, storage, and transmission are associated with high power consumption. As a result, systems are often forced to operate at less than ideal sampling rates. Furthermore, missing measurements in WSNs are often attributed to communication failures, where packets are lost, or to de-synchronization of sensors, leading to different sampling instances.

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

In this work, we focus on the problem of recovering missing measurements of environmental sensing platforms using two state-of-the-art methods: Matrix Completion (MC) and Tensor Completion (TC). In MC, the objective is to exploit inherent correlations within the data in order to recover low rank matrices from a substantially limited number of observations. MC has been successfully applied in an array of problems. In addition to MC, we also consider the extension of this problem to higher dimensional structural data that can be represented as tensors. The reason for exploring tensors is that two-way matrices are unable to preserve the higher structural complexity needed for simultaneously encoding data from a variety of sources.

2 Matrix and Tensor Completion

Given a data matrix $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$, the goal of Matrix Completion is to recover all its entries from a partially observed fraction of them. More 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 an operator setting all unknown indices to zero:

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

In [1] it was shown that recovery of the missing values from a low rank matrix $\mathbf{X}$ is possible by solving the rank minimization problem:

$$\begin{aligned} &\underset{\mathbf{X}}{\text{minimize}} \; rank(\mathbf{X}) \\ &\text{subject to } \; \mathcal{A}(\mathbf{X}) = \mathcal{A}(\mathbf{M}) \end{aligned} \tag{1}$$

Although the rank minimization can recover the matrix, it is impractical for real-life problems due its NP-hard nature. Fortunately, it has been shown that the nuclear norm, *i.e*, the sum of the singular values, can serve as a proxy to the rank. Thus, the optimization problem in (1) can be reformulated according to

$$\begin{aligned} &\underset{\mathbf{X}}{\text{minimize}} \;\; \|\mathbf{X}\|_* \\ &\text{subject to } \; \mathcal{A}(\mathbf{X}) = \mathcal{A}(\mathbf{M}) \end{aligned} \tag{2}$$

From a theoretical point of view, in order for these methods to recover the desired solution, the sampling set Ω must be chosen uniformly at random and at the same time the data matrix $\mathbf{M}$ must satisfy a low coherence condition. Then, with probability $1 - n^{-3}$, the solution of (2) will converge to the solution of (1), provided that the number of obtained samples obeys $k \geq Cn^{6/5} rlog(n)$, where $n = max(n_1, n_2)$, C is an appropriate constant, and r is the matrix rank.

An alternative approach to tackle the problem of missing measurements is via Tensor Completion. Tensors are generalizations of vectors and matrices that encode high dimensional structural information. In this case, the aforementioned factors leading to missing measurements within a WSN would result in an undersampled $[n_1] \times [n_2] \times [n_3]$ tensor $\mathcal{X}$, which we wish to recover from a fraction k of its entries being available.

Equation (2) 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 $\mathcal{X}$ which agrees with the given data:

$$\begin{aligned} &\underset{\mathcal{X}}{\text{minimize}} \;\; \|\mathcal{X}\|_* \\ &\text{subject to} \;\; \mathcal{A}(X) = \mathcal{A}(\mathcal{T}) \end{aligned} \tag{3}$$

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

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

However, the optimization regime now is tougher than before, as the tensor nuclear norm is not defined as the tightest convex relaxation of the tensor rank, as was the case with matrices. Adopting the approach proposed in [2], one can define the tensor nuclear norm as follows:

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

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 can be defined as the convex combination of the nuclear norms of all matrices unfolded along each of its modes. Under this definition, Eq. (3) can be written as:

$$\begin{aligned} &\underset{\mathcal{X}}{\text{minimize}} \;\; \sum_{i=1}^{n} \alpha_i \|\mathcal{X}_{(i)}\|_* \\ &\text{subject to} \;\; \mathcal{A}(\mathcal{X}) = \mathcal{A}(\mathcal{T}) \end{aligned} \tag{4}$$

3 Experimental Evaluation

In this section, we evaluate the performance of two state-of-the-art methods for Matrix and Tensor Completion on real environmental sensing data. For the MC problem we chose the widely used Augmented Lagrange Multipliers (ALM) method [3] and for the TC one the Low-rank Tensor Completion using Parallel Matrix Factorization approach [4]. The performance is reported in terms of the Normalized Mean Square Error (NMSE) metric.

We evaluate the performance of the MC and the TC recovery for matrices and tensors, ranging from ”sparsely” to ”densely” sampled. To quantify this density range, we introduce the fill-ratio f, which is defined as the number of the non-zero elements divided by the number of all the available entries of our

$[n_1] \times [n_2]$ measurements matrix $f = \frac{\#non-zeroelements}{n_1 \times n_2}$. This performance metric is computed and subsequently plotted versus different sizes of available data revealed to the solvers. We consider two representative environmental sensing datasets, namely an indoor SWN dataset and an outdoor WSN dataset.

3.1 Recovery of SWN Data

The SWN dataset contains measurements of water impedance (in Ohms) recorded in 10 different channels by 5 sensors deployed in a pilot desalination plant [5]. These measurements were collected during a 3 day period, where each sensor recorded 1 measurement per hour. In this experiment, we evaluate the performance of the MC and TC method for recovery, under different fill ratios, as shown in Figure 1. We consider two cases and associated data structures, namely acquisition of one measurement every one and every two hours, resulting in 50×72 and 50×36 matrices for MC and $5 \times 10 \times 72$ and $5 \times 10 \times 36$ tensors for TC.

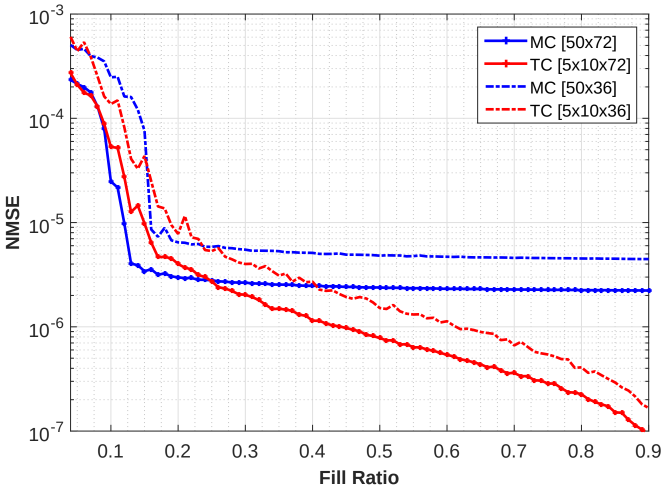

Fig. 1: Normalized MSE for MC and TC as a function of the fill ratio (SWN dataset).

We can observe in Figure 1, that increasing the fill-ratio has a dramatic effect on the reconstruction quality. More precisely, for both one and two hour sampling frequency, increasing the fill ratio leads to lower reconstruction error, as expected. The results also demonstrate that TC achieves a lower NMSE than MC. More importantly, the error rate of TC is monotonically decreasing, unlike the error rate of MC which reaches a plateau. For very low fill-ratios ($f < 0.2$), we observe that MC performs better than TC, a behavior which is is due to the fact that completing a data structure becomes tougher as the dimensionality of the structure increases.

3.2 Recovery of WSN Data

The second dataset is part of the SensorScope network project [6] where we consider temperature data from the Grand-St-Bernard pass between Switzerland and Italy. We selected an array of 19 sensors that provided stable results and collected 288 measurements per day for a period of 10 days.

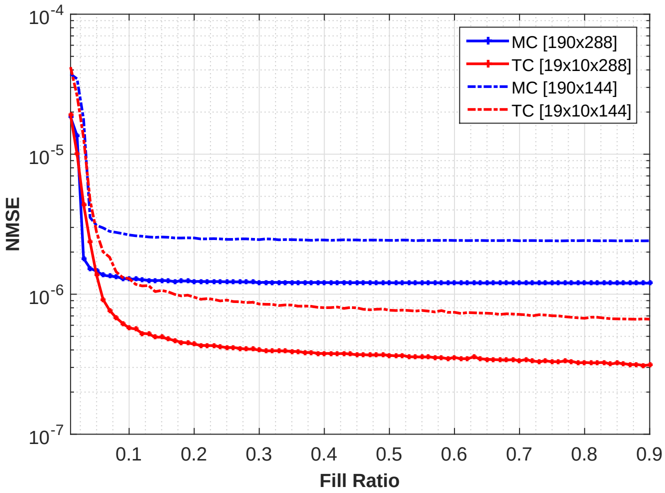

Fig. 2: Normalized MSE for MC and TC as a function of the fill ratio (SensorScope dataset).

Experimental results presented in Figure 2 demonstrate that increasing the fill-ratio leads to considerable performance gains in lower sampling regimes compared to higher ones. The overall NMSE achieved by the TC approach is lower compared to the error obtained via the MC approach, similarly to the SWN dataset. Again, the performance gap between the two approaches is clearly in favor of TC, both in the case of sampling per one hour and per two hours.

In addition, the NMSE achieved by TC is clearly reduced as we increase the fill-ratio, in contrast to the MC approach where an increase in the fill-ratio over 0.2 does not appear to have any considerable effects to its performance. The results also highlight the fact that in this case, where the dataset was quite larger than the one used in the first experiment, the TC approach appears to start outperforming MC for a lower fill-ratio value regime.

To further visualize the performance of the recovery methods, Figure 3 presents ground truth and reconstructed measurements from a single day, sampled at a fill-ratio $f = 0.2$. We observe that significant features are accurately estimated even from such low sampling rates, justifying the potential of the proposed methods for efficient data acquisition and transmission in WSNs.

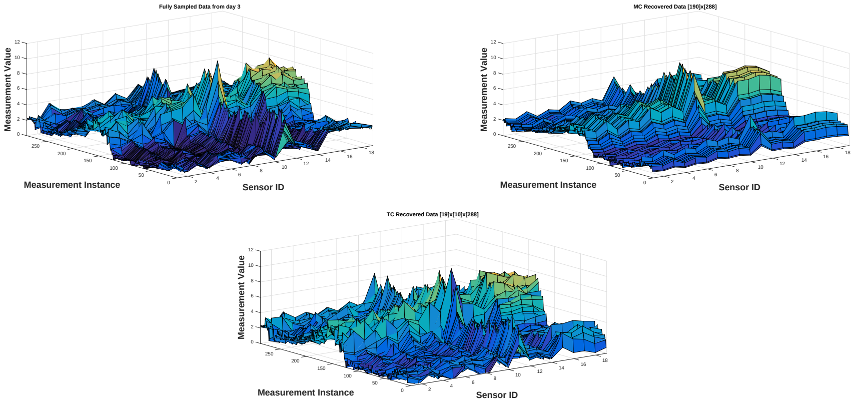

Fig. 3: Exemplary reconstruction from 20% of the measurements.

4 Conclusions

In this work, we investigated the application of Tensor and Matrix Completion for the estimation of missing environmental sensing measurements. The experimental evaluation on two distinct datasets demonstrate that both the Matrix and Tensor Completion approaches are promising methods for achieving that goal, while Tensor Completion can better exploit the structure and availability of the data. Results imply that as we strive to handle higher quantities of data, emerging correlations can be better exploited by higher-order tensors.

References

[1] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[2] Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):208–220, 2013.

[3] Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[4] Yangyang Xu, Ruru Hao, Wotao Yin, and Zhixun Su. Parallel matrix factorization for low-rank tensor completion. *arXiv preprint arXiv:1312.1254*, 2013.

[5] G Tzagkarakis, G Tsagkatakis, D Alonso, E Celada, C Asensio, A Panousopoulou, P Tsakalides, and B Beferull-Lozano. Signal and data processing techniques for industrial cyber-physical systems. *Cyber Physical Systems: From Theory to Practice, DB Rawat, J. Rodrigues, and I. Stojmenovic, Eds. CRC Press, USA*, 2015.

[6] Guillermo Barrenetxea, François Ingelrest, Gunnar Schaefer, Martin Vetterli, Olivier Couach, and Marc Parlange. Sensorscope: Out-of-the-box environmental monitoring. In *Information Processing in Sensor Networks, 2008. IPSN'08. International Conference on*, pages 332–343. IEEE, 2008.