Machine Learning (2021) 110:3157–3175
https://doi.org/10.1007/s10994-021-05985-w

Misalignment problem in matrix decomposition with missing values

Sofia Fernandes[1] · **Mário Antunes**[2] · **Diogo Gomes**[2] · **Rui L. Aguiar**[2]

Received: 28 September 2020 / Revised: 30 March 2021 / Accepted: 12 April 2021 /
Published online: 28 April 2021

Abstract
Data collection within a real-world environment may be compromised by several factors such as data-logger malfunctions and communication errors, during which no data is collected. As a consequence, appropriate tools are required to handle the missing values when analysing and processing such data. This problem is often tackled via matrix decomposition. While it has been successfully applied in a wide range of applications, in this work we report an issue that has been neglected in literature and "degenerates" the quality of the imputations obtained by matrix decomposition in multivariate time-series (with smooth evolution). Briefly, the problem consists of the misalignment of the matrix decomposition result: the missing values imputations fall within an incorrect range of values and the transitions between observed and imputed values are not smooth. We address this problem by proposing a post-processing alignment strategy. According to our experiments, the post-processing adjustment substantially improves the accuracy of the imputations (when the misalignment occurs). Moreover, the results also suggest that the misalignment occurs mostly when dealing with a small number of time-series due to lack of generalization ability.

Keywords Matrix decomposition · Missing values · Multivariate time-series · Imputation

Editors: João Gama, Alípio Jorge, Salvador García.

✉ Sofia Fernandes
ssf@av.it.pt

Mário Antunes
mario.antunes@av.it.pt

Diogo Gomes
dgomes@av.it.pt

Rui L. Aguiar
ruilaa@av.it.pt

[1] Departamento de Electrónica, Telecomunicações e Informática, Universidade de Aveiro, Aveiro, Portugal

[2] Instituto de Telecomunicações, Departamento de Electrónica, Telecomunicações e Informática, Universidade de Aveiro, Aveiro, Portugal

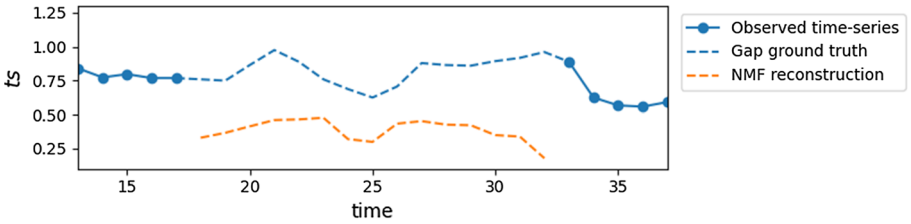

Fig. 1 Example of the matrix decomposition misalignment with respect to the observed time-series (*ts*)

1 Introduction

With the emergence of the Internet of Things (IoT), new data is being collected every day. Mining such data in an efficient and accurate manner is crucial in order to extract value from it. The mining process involves not only analysis and processing tools, but also strategies to handle erroneous and missing data. For example, when collecting sensor data, one recurrent issue is the existence of time gaps in which no measurements were published due to connectivity problems or sensor malfunctions (Karkouch et al., 2016). While there have been considerable efforts on imputing the missing values in time-series, traditional approaches focus mainly in isolated missing values (Stekhoven & Bühlmann, 2011). Nonetheless, in real-world scenarios, data is generally missing in blocks of consecutive timestamps over multiple time-series, thus requiring more robust strategies to approximate the unobserved values (Cao et al., 2018).

One of the approaches to tackle this problem is via matrix/tensor decomposition, also widely considered for image recovery (Zhao et al., 2015) and recommendation systems (Bao et al., 2014). Briefly, given the matrix/tensor modeling the time-series, the goal is to estimate the missing values by accounting for the similarities between the time-series, captured through decomposition techniques. This approach is appropriate when dealing with low rank data (Chen et al., 2013). On the contrary to naive approaches, it is able to deal with gaps consisting of a large number of consecutive timestamps. On the other hand, it does not require a large number of time-series to perform well, contrarily to the emerging deep learning techniques. One of its limitations is its inability of providing an estimate to a time gap in which no time-series were observed (Banerjee & Roy, 2014).

In this work, we report one additional issue, to the best of our knowledge, not addressed in literature, that we refer to as the "misalignment problem in matrix decomposition with missing values". This problem is illustrated in Fig. 1: while the gap estimate obtained through non-negative matrix factorization (NMF) accurately captures the evolution of the time-series in the missing gap, the estimate range of values is distant from the real one. In fact, by analysing the observed time-series values, one notes that the changes between consecutive timestamps are smooth and such smoothness is not preserved between the observed values and the reconstruction.

To tackle this issue, recent works (Yu et al., 2016; Chen & Sun, 2020) have incorporated temporal smoothness constraints into the decomposition algorithms. However, as it is demonstrated in this work, such approaches may still lead to inaccurate imputations. Based on this, we propose a post-processing procedure to correctly align the reconstruction along the observed values. The main novelty of our approach arises from imposing the smoothness at a post-processing stage so that the procedure is algorithm independent. It should be noted that the misalignment problem was detected when imputing a real-world time-series

of indoor climate, collected within the scope of project Smart Green Homes (SGH)[1] with the goal of understanding the tenants' thermal comfort patterns and take advantage of such information to improve energy-efficient use. Briefly, our dataset consisted of three measurements (temperature, humidity and pressure) collected through sensors at 14 distinct houses, with around 50% of missing values. The usage of matrix decomposition was motivated not only by the low rank nature of the data (Chen et al., 2013), but also by the small number of time-series. A special focus is provided to this case study. Nonetheless, we also investigate the presence of this problem, as well as the suitability of the proposed solution, in other datasets.

Based on this, our contributions are as follows:

1. We report the misalignment problem, observed when applying matrix decomposition with missing values to some datasets;
2. We propose a post-processing adjustment strategy to tackle the misalignment problem and carry out a set of experiments not only to validate the proposed solution but also to study the conditions under which the misalignment problem manifests.

The rest of the work is organized as follows. In Sect. 2 we overview the related literature. In Sect. 3 we describe the problem reported. The proposed method is described in Sect. 4 and the experiments are presented in Sect. 5. We conclude in Sect. 6.

2 Related work

The most straightforward approach for imputing values in time-series is the replacement by a fixed value such as the mean. To account for the neighbourhood of the missing gap, linear interpolation and moving average variants may also be considered (Lepot et al., 2017). Nonetheless, such approaches may not be appropriate when the amount of data missing is large. Moreover, these techniques are univariate and do not take advantage of the possible similarities/correlations between the multiple sources of data.

With respect to approaches specifically designed for multivariate time-series imputation, multiple research directions have been explored. In Multiple Imputation Chained Equations (MICE) (Azur et al., 2011), a statistical perspective was considered. In this method, multiple estimates are obtained through regression models and combined to fill the missing gaps.

Recently, a new trend emerged in this field and is based on machine learning techniques, namely deep learning. In this context, multiple Recurrent Neural Networks (RNN) architectures have been proposed (Che et al., 2018; Cao et al., 2018; De Brouwer et al., 2019; Liu et al., 2019). Additionally, Fortuin et al. (2020) proposed a model based on auto-encoders. Briefly, the idea consists of encoding the time-series with missing values to a low-dimensional complete space so that the missing values are recovered from the decoding phase. Another related approach is the usage of Generative Adversarial Nets (GAN) (Luo et al., 2019), which consist of two components: a generator network, that carries out the imputation process, and a discriminator, that distinguishes the observed (known) values from the imputed ones. In order to obtain a high-quality imputation, the discriminator is

[1] http://www.ua.pt/smartgreenhomes/.

trained to minimize the *real vs. imputed* classification error; on the opposite to the generator, which is trained to maximize the discriminator error.

As previously exposed, another perspective, often considered for this task, is matrix and tensor decomposition. In general, typical decomposition methods for handling missing data can be applied to time-series, including non-negative variants (Kim & Choi, 2009) and adaptations of Singular Value Decomposition (Cai et al., 2010), Principal Component Analysis (Shu et al., 2014; Zhang & Balzano, 2016; Balzano et al., 2018) and tensor decomposition methods (Acar et al., 2011; Williams et al., 2018). However, none of such approaches accounts for the temporal relation between observations when modeling time-series. In this context, Yu et al. (2016)) proposed Temporal Regularized Matrix Factorization (TRMF) which incorporates temporal smoothness. In particular, TRMF has the ability to incorporate smoothness not only between two consecutive timestamps, but also over timestamps spaced by a given period (thus, allowing to model the periodic patterns). With an identical goal, Chen and Sun (2020)) proposed Low-Rank Autoregressive Tensor Completion (LATC) in which the multivariate time-series is reshaped into a tensor by adding a temporal dimension to explicitly model periodicity and seasonality.

Despite these efforts to model temporal evolution, we observed that there were scenarios in which there was a misalignment between the matrix/tensor decomposition reconstruction and the observed values. This problem, that substantially degenerates the imputation power of these approaches, is the target of our work and is described in more detail next.

3 Problem formulation

In this work, we expose the misalignment problem in matrix decomposition with missing values, a problem derived from applying matrix decomposition to impute multivariate time-series. Therefore, in order to facilitate the understanding of the problem addressed, we start by introducing both the problem of multivariate time-series imputation and matrix decomposition with missing values.

3.1 Multivariate time-series imputation

From a general point of view, the missing value imputation problem in time-series may be formulated as follows: given a multivariate time-series consisting of N univariate time-series, $\mathcal{X} = \{\mathrm{x}_i\}_{i=1}^{N}$, with T timestamps, some of which were not observed; the goal is to find an accurate estimate for the missing values.

3.2 Matrix decomposition with missing values

Let $\mathbf{X} \in \mathbb{R}^{N \times M}$ be a matrix, the goal of matrix decomposition (also referred to as matrix factorization) is to find matrices $\mathbf{U} \in \mathbb{R}^{N \times R}, \mathbf{V} \in \mathbb{R}^{M \times R}$, so that $\mathbf{X} \sim \mathbf{U}\mathbf{V}^T$ and the approximation/fit error, given by $||\mathbf{X} - \mathbf{U}\mathbf{V}^T||^2$, is minimized (with $||.||$ denoting the Frobenius norm). R is a parameter of the model and is referred to as number of components/factors.

In case there are values missing in the data matrix $\mathbf{X}$, the decomposition must be driven in order to minimize the approximation error in the known values. Consequently, matrices $\mathbf{U}, \mathbf{V}$ are obtained by minimizing:

Table 1 Notation summary

Symbol	Description
x	scalar (nonbold lower case)
$\mathbf{x}$	vector (bold lower case)
$\mathbf{x}[i]$	entry i of vector $\mathbf{x}$
$\mathbf{X}$	matrix (bold upper case)
$\mathbf{X}[i,j]$	entry (i,j) of matrix $\mathbf{X}$
$\mathbf{X}[i,:]$	row i of matrix $\mathbf{X}$
$\mathbf{X}[i,j_1 : j_n]$	entries (i,j_k) with $j_k \in \{j_1,\ldots,j_n\}$ of matrix $\mathbf{X}$

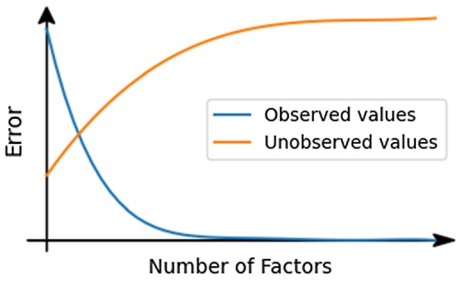

Fig. 2 Error curves for the observed and unobserved values in the temperature measurements of our case study dataset when reconstructing gaps of 48 h (smoothed with polynomial regression for illustration purposes)

$$\left|\left|\mathbf{O} \odot \left(\mathbf{X} - \mathbf{U}\mathbf{V}^T\right)\right|\right|^2 + \alpha_1 ||\mathbf{U}||^2 + \alpha_2 ||\mathbf{V}||^2 \tag{1}$$

(Kim & Choi, 2009), where $\mathbf{O}$ is the matrix mapping the observed values ($O[i,j] = 1$ if entry (i,j) was observed and $O[i,j] = 0$, otherwise) and $\odot$ is the Haddamard product (consisting of the element-wise product). The regularization terms are meant to reduce overfitting to the observed values.

To facilitate the reading of this manuscript, the notation used throughout this work is summarized in Table 1.

3.3 Misalignment problem in matrix decomposition with missing values

Matrix decomposition is often applied to solve the multivariate time-series imputation problem by modelling each time-series as a row of the data matrix (that is, by setting $\mathbf{X}[t,:] = \mathbf{x}_t$).

When resorting to matrix factorization for imputing time-series, one is interested not only in obtaining a model that fits the observed data but, most of all, in a model that has generalization ability. If, on one hand, the higher the number of components the higher the fit of the matrix decomposition model; the same trend does not necessarily hold regarding the quality of the estimated missing values. In particular, in a preliminary analysis of the performance of NMF in our case study, we observed that, despite the regularization terms in (Eq. 1), NMF started to overfit with an extremely low number of components as shown in Fig. 2. In such case, in order to maximize the generalization ability we have to neglect the fitting model, potentially resulting in misplaced reconstructions (recall Fig. 1). We refer to this problem as the *misalignment problem*

in matrix decomposition with missing values. In other words, this problem refers to the misalignment between the matrix decomposition result and the observed values that leads to a misplacement of the missing gaps estimation and consequent poor accuracy.

Whilst in this work we study the misalignment phenomena across multiple decomposition algorithms, our aim is not to perform a comparative study of the methods, but, instead, to show that the misalignment may occur even when considering time-aware decomposition approaches. For a recent comparison study on time-series imputation techniques, the reader may refer to Khayati et al. (2020).

4 Proposed method

To tackle the misalignment problem in matrix decomposition, we propose a post-processing technique to be applied to each of the gaps estimated imputations. Our framework is general and can be applied regardless of the decomposition algorithm considered. In fact, it is general enough so that it is applicable regardless of the imputation strategy considered. However, to study its need/usefulness when applied to other imputation strategies does not fall within the scope of this work and is left for future investigation.

Our method was designed by assuming that the time-series being imputed change smoothly (as in our case study) and consists of two alternative transformations: (i) gap

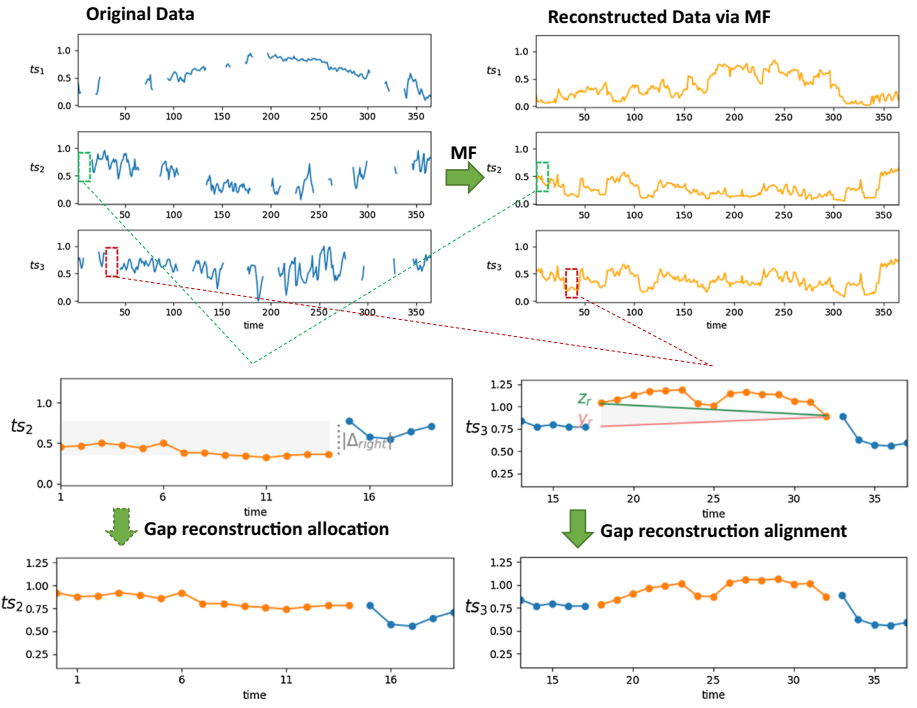

Fig. 3 Adjustment process illustration: for the given gap, matrix factorization (MF) reconstruction is transformed with the goal of smoothing the transition between the observed values and the imputed ones

allocation and (ii) gap alignment. When the missing gap includes the first or last timestamps of the time-series (in which case there is less information regarding the expected range of values), we apply the gap allocation transformation which consists of a shift to the observed gap margin. In the remaining gaps, we apply the alignment transformation, in which the estimated sequence is subject to a shift and a slight distortion in order to simultaneously guarantee that the imputation falls within the expected range of values and both the observed→imputed and imputed→observed transitions are smooth.

The adjustment process is illustrated in Fig. 3 and documented in algorithm 1—for ease of understanding we incorporated the decomposition into the process.

Algorithm 1: Adjusted matrix factorization.

Data: incomplete matrix $\mathbf{X} \in \mathbb{R}^{N \times T}$, number of factors R, decomposition method $factorize$, gaps list $\{ts_i, [t_1^i, \ldots, t_{L_i}^i]\}_i$,
Result: complete matrix $\mathbf{X}' \in \mathbb{R}^{N \times T}$

// Initialize the incomplete matrix with the incomplete matrix
$\mathbf{X}' \leftarrow \mathbf{X}$

// Get reconstruction via factorization
$\hat{\mathbf{X}} \leftarrow fatorize(\mathbf{X}, R)$

// Adjust reconstruction
for *gap i in gaps list* **do**
  $ts \leftarrow ts_i$ // get time-series index of gap i
  $gap \leftarrow [t_1^i, \ldots, t_{L_i}^i]$ // get gap i timestamps
  $\hat{\mathbf{x}} \leftarrow \hat{\mathbf{X}}[ts, gap]$ // get gap reconstruction
  if *there are no observed time-series for the gap first and last timestamp* **then**
    $\hat{\mathbf{x}} \leftarrow linearpadding(\hat{\mathbf{x}}, gap)$ // extend gap linear interpolation
  if $t_1^i = 1$ // gap first missing timestamp is first instant of the time-series **then**
    // Allocate gap reconstruction to the right observed margin
    $\Delta_{right} \leftarrow \mathbf{X}[ts, t_{L_i}^i + 1] - \hat{\mathbf{x}}[L_i]$
    $\hat{\mathbf{x}} \leftarrow \hat{\mathbf{x}} + \Delta_{right}$ // align with the observed right margin
  else if $t_{L_i}^i = T$// gap last missing timestamp is last instant of the time-series **then**
    // Allocate gap reconstruction to the left observed margin
    $\Delta_{left} \leftarrow \mathbf{X}[ts, t_1^i - 1] - \hat{\mathbf{x}}[1]$
    $\hat{\mathbf{x}} \leftarrow \hat{\mathbf{x}} + \Delta_{left}$ // align with the observed left margin
  else
    // Align gap
    $\mathbf{y} \leftarrow linearinterpolate(\mathbf{X}[ts, t_1^i - 1], \mathbf{X}[ts, t_{L_i}^i + 1], L_i)$ // get observed "trend"
    $\mathbf{z} \leftarrow linearinterpolate(\hat{\mathbf{x}}[1], \hat{\mathbf{x}}[L_i], L_i)$ // get reconstructed "trend"
    $\hat{\mathbf{x}} \leftarrow \hat{\mathbf{x}} - (\mathbf{z} - \mathbf{y})$ //adjust the reconstruction
  // Impute the missing gap
  $\mathbf{X}'[ts, gap] \leftarrow \hat{\mathbf{x}}$

Let $\mathbf{X} \in \mathbb{R}^{N \times T}$ be the incomplete time-series matrix with the missing blocks defined as $\{ts_i, [t_1^i, \ldots, t_{L_i}^i]\}_i$, meaning that block i of consecutive timestamps $[t_1^i, \ldots, t_{L_i}^i]$, with $t_j^i = t_{j+1}^i - 1, \forall j < L_i$, is missing for time-series ts_i (no assumption is made on the number of gaps per time-series). We start by decomposing the data with R factors by using the desired decomposition algorithm, which we refer to as *factorize*. Then, we apply the post-processing procedure to each gap. It should be noted that if there are no observed values for a given time-series, then the decomposition does not provide an useful (non-degenerated)

estimation for it and, as consequence, it is of no use to apply the adjustment procedure. Therefore, for simplicity, we assume that there were observed instants for each time-series in the matrix.

Let us denote the estimated imputation for gap i as $\hat{\mathbf{x}}$ (corresponding to the matrix decomposition output on timestamps $[t_1^i, \ldots, t_{L_i}^i]$ of time-series ts_i). Since the gap first and last timestamps might be missing for all the time-series (in which case, the decomposition does not provide a useful estimate), we apply a linear padding to such timestamps (through *linearpadding*). This step is necessary because the proposed adjustment relies on the reconstruction at the gap margins, that is, at instants t_1^i and $t_{L_i}^i$. Therefore, its goal is to obtain a reliable estimation of the reconstruction $\hat{\mathbf{x}}$ at instants t_1^i and $t_{L_i}^i$, whenever no observation was obtained at such instants. For illustration purposes let us assume no time-series were observed at instant t_1^i (but there were observed time-series at $t_{L_i}^i$), then the matrix decomposition was not able to properly estimate $\hat{\mathbf{X}}[ts_i, t_1^i]$, and the use of such margin for the adjustment process would strongly affect the imputation quality. To overcome such issue, we search for the first timestamp $t' \in \{t_1^i + 1, \ldots, t_{L_i}^i - 1\}$, for which at least a time-series was observed. Given $\hat{\mathbf{x}}$ at the corresponding t' and $t_{L_i}^i$ timestamps, we obtain the parameters of the linear interpolation between such points, corresponding to the gap reconstruction non-degenerated margins, and estimate $\hat{\mathbf{x}}$ at t_1^i as the interpolation value at such point. This procedure may be carried out for all timestamps in $\{t_1^i, \ldots, t' - 1\}$, thus encompassing a "padding" to the original reconstruction. In the case where no time-series were observed at instant t_L^i, the procedure is identical with the main difference that t' is computed as the last timestamp in the gap for which at least a time-series was observed. After this preliminary step, one of the following transformations is carried out.

Gap allocation If the missing gap occurs at the first or last timestamps of the time-series, $\hat{\mathbf{x}}$ is shifted by a magnitude corresponding to the difference between the last missing and the first observed values (if the first timestamps are missing) or the last observed and the first missing values (if the last timestamps are missing). No further processing is carried out because there is not enough information to improve the imputation in these cases.

Gap alignment When the gap is preceded and followed by observed values, we can use such (observed) values to improve the imputation quality by imposing smoothness in both observed→imputed and imputed →observed transitions. We are interested not only in imposing smoothness but also in preserving the evolution of $\hat{\mathbf{x}}$, captured via matrix decomposition. Based on this, the idea consists of aligning $\hat{\mathbf{x}}$ according to the gap linear interpolation, $\mathbf{y}$, by shifting $\hat{\mathbf{x}}$ by a magnitude which is dependent on the linear interpolation between $\hat{\mathbf{x}}$ first and last values, $\mathbf{z}$ (as illustrated by the grey area in Fig. 3).

It is noteworthy that in the gap alignment procedure, our goal was to find a transformation which would impose smoothness in the *observed* ↔ *imputed* transitions while applying the least amount of distortion. Based on this, intuitively the proposed transformation aims at approximating a symmetry. Nonetheless, on contrary to symmetry, this transformation preserves the time axis, which is an important feature because we are dealing with discrete time and it would not be appropriate to model, for example, 0.5 timestamps.

We reinforce that the adjustment strategy was developed under the assumption of smoothness in the time-series evolution. Therefore, further investigation must be carried out in order to develop a strategy which is robust to local anomalies in the first and last timestamps of the missing gap. We leave this research direction as future work.

5 Experiments

In order to empirically show the existence of the misalignment problem and to study the usefulness of the proposed adjustment procedure, we carried out multiple experiments, whose details are provided below. For reproducibility purposes, the source code is available at https://github.com/ATNoG/adjusted-mf. It should be noted that the implementations of TRMF and LATC were derived from the corresponding authors public implementations, while public libraries including SVT (Duan, 2020) and CP (Williams et al., 2018) were also considered. Only NMF implementation was written from scratch.

5.1 Datasets

In these experiments, we gave special focus to our case study: the indoor climate dataset. However, additional datasets were considered for the validation of our study. All datasets are characterized as follows.

Indoor climate consists of temperature, humidity and pressure values measured in 14 houses by a period of 13 consecutive months (from March 2019 to April 2020). A measurement (*temperature*, *relativehumidity*, *pressure*) was collected by wireless sensors whenever a considerable change in the values was detected (with a minimum time frequency of an hour), after the sensor connects to an IoT platform named SCoT (Santiago et al., 2019) to share the measurements. The three measurements were made by the same device. Therefore, the main causes for missing data were the sensor malfunction (in which case, no measurements were performed for the corresponding tenant during the malfunction period) and connection issues with the IoT platform (in which case, no measurements were stored for any of the houses). The measurements of each variable were hourly averaged and rearranged into a 14×9528 matrix, for which around 52% of the entries are missing. Temperature, relative humidity and pressure were, respectively, measured in Celsius degrees (C^o), percentage (%) and millibars (*mbar*).

As previously exposed, this dataset was acquired within the scope of Smart Green Homes project. The participants were informed about the goals of the study and provided their consent on the data collection, processing and analysis. The privacy of the participants was also ensured.

Iberian Peninsula average temperature (Menne et al., 2012, 2020) comprehends a set of time-series of daily average temperature across 9 weather stations in the south of the Iberian Peninsula over a period of a year. It should be noted that since the stations locations were close, the time-series assume a similar range of values. In fact, the median correlation between two different time-series in this dataset was ≈ 0.97.

This dataset was provided by the National Oceanic and Atmospheric Administration (NOAA)[2], we refer to it as IBERIAN_TEMP.

World average temperature (Menne et al., 2012, 2020) is similar to IBERIAN_TEMP, as it also consists of daily average temperatures and it was provided by NOAA. However, in this dataset, we considered 12 locations across the different continents. The median pairwise correlation of these time-series was ≈ 0.4, nonetheless, there was a subset of time-series with a pairwise median correlation of ≈ 0.8.

For simplicity, we refer to this dataset as WORLD_TEMP.

Guangzhou traffic (Chen et al., 2018) comprises time-series of traffic speed across multiple road segments in the city of Guangzhou, China. In particular, the dataset is formed by 214 time-series with 8784 timestamps each (corresponding to 61 days of 10 min

[2] https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.ncdc:C00861.

timestamps). In this work, we restricted the dataset to the first 5 days and discarded 5 road segments whose data was missing. The time-series in this dataset were "heterogeneous" with the median pairwise correlation between them being less than 0.65. We refer to this data subset as GUANGZHOU.

Seattle traffic[3] (Cui et al., 2018, 2019) is formed by time-series of traffic speed on a freeway of Seattle. In more detail, there are 323 time-series collected at different freeway mileposts at each 5 min and, for simplicity, we restricted our analysis to the first 500 timestamps. The resulting data was also heterogeneous with the median correlation across time-series being lower than 0.55. The resulting data matrix is dubbed as SEATTLE.

Gas sensor[4] (Fonollosa et al., 2014) is, as the name suggests, a collection of measurements made by 8 gas sensors (acquired for ethylene released at high concentration and zero carbon monoxide concentration). In this dataset, the measurements were arranged into 100 milliseconds timestamps, and we restricted our study to the first 500 timestamps. The median correlation across time-series was ≈ 0.53.

5.2 Baselines

Since our goal is to improve the accuracy of matrix decomposition when imputing missing data, the baseline in this study is the matrix decomposition itself. Based on this we considered NMF for incomplete data (Kim & Choi, 2009), as the main decomposition algorithm in all of the experiments.

When targeting our case study on the indoor climate data, we considered four other decomposition approaches. This is justified by our interest in finding the most accurate imputation for this dataset and by the need of providing evidence that the misalignment is not specific to NMF but, instead, also occurs when considering other decomposition methods. In this context, we considered TRMF (Yu et al., 2016), SVT (Cai et al., 2010) and LATC (Chen & Sun, 2020) as the methods in which each thermal measurement was imputed isolatedly. Moreover, we also considered non-negative CP decomposition by Hierarchical Alternating Least Squares (Cichocki & Phan, 2009; Williams et al., 2018), a tensor decomposition approach in which the three measurements were jointly decomposed. For the sake of simplicity we refer to this method as CP.

5.3 Evaluation metric

Since in imputation tasks it is not possible to account for the accuracy of the imputation of the unobserved values (due to the lack of ground truth), a common strategy consists of removing a set of observed values from the data and account for the imputation quality on such values. Based on this, in order to measure the imputation quality, we resorted to the root mean squared error (RMSE). In more detail, for each dataset and experimental setting, we removed sequences of consecutive values $\mathbf{y}_i = \mathbf{X}[ts_i, t_1^i : t_{L_i}^i]$, applied matrix decomposition to the "corrupted" datasets and measured the imputation error as:

[3] https://github.com/zhiyongc/Seattle-Loop-Data.

[4] https://archive.ics.uci.edu/ml/datasets/Gas+sensor+array+exposed+to+turbulent+gas+mixtures.

$$RMSE = \sqrt{\frac{1}{\sum_{i=1}^{n} L_i} \sum_{i=1}^{n} \sum_{k=1}^{L_i} (\mathbf{y}_i[k] - \hat{\mathbf{y}}_i[k])^2} \tag{2}$$

where $\hat{\mathbf{y}}_i$ denotes the corresponding estimation of $\mathbf{y}_i$. Also, $\sum_{i=1}^{n} L_i$ accounts for the total number of missing entries in the data matrix.

It should be noted that the setting considered for obtaining RMSE was dependent on the completeness of the dataset under study. When dealing with incomplete data, each $\hat{\mathbf{y}}_i$ was obtained from the decomposition of the corrupted dataset obtained by discarding $\mathbf{y}_i$, similarly to (Tarpey, 2000). In datasets with no missing values, a model was obtained from a corrupted version of the dataset, in which the set of all gaps, $\{\hat{\mathbf{y}}_i\}_{i=1}^{n}$, was missing. In this case, only one decomposition was performed, but it involved the imputation of multiple gaps. We note that, in datasets with no missing values, if we removed only a single gap at the time, the amount of remaining data for the decomposition model to learn would still be large (> 90%), which would facilitate the problem.

5.4 Experimental setting

This study involved sets of experiments with some variations between them. Nonetheless, the main set up was similar: (i) for each dataset, we generated 10 random samples of gaps from the observed values (each sample of gaps encompasses blocks of consecutive timestamps across the multiple time-series, as defined in Sect. 4, with an additional constraint of fixed length size: $L_i = L, \forall i$; no restriction was considered regarding overlapping gaps, that is, some timestamps may be missing for more than one time-series); (ii) for each sample, we applied the procedure described in Sect. 5.3 and computed the RMSE using (Eq. 2); (iii) finally, we computed the median of the RMSE results over the 10 samples. We note that we present the median RMSE for the sake of simplicity. Still, we analysed the results of each sample individually and we observed that the median RMSE was representative of the behaviour observed in the majority of the samples.

For the indoor climate dataset, a sample consisted of around 208 gaps of observed sequences and we considered fixed length gaps. Multiple experiments were carried out by varying the gap size between 12 and 96 h. In the remaining datasets, the gap sizes considered ranged from 5 to 40 timestamps and the number of gaps was selected so that the rate of missing timestamps was around 50%.

In all of the experiments, we varied the number of factors in the matrix decomposition techniques from 1 to 12. For the sake of simplicity, we only show the results for the model associated with the minimal median RMSE. With respect to method specific parameters, we carried out a (non-exhaustive) search on the parameter space to find the parameter values that lead to more accurate imputations. Based on the results, we used a regularization factor of 0.1 for NMF. In TRMF we considered all regularization terms ($\lambda_f, \lambda_x, \lambda_w$) to be 0.75 and a lag set ($\mathcal{L}$) of $\{1, 2, 3\}$. In LATC, we considered a lag set of $\{1\}$ and the remaining parameters were set as $\alpha = \{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}, \rho = 10^{-4}, \lambda = 5 \times 10^{-4}, \theta = 15, \epsilon = 10^{-3}$. In the remaining methods, the parameter values considered coincided with the defaults in the corresponding implementations[5].

[5] The reader should refer to Duan (2020) for details on the SVT implementation and to Williams et al. (2018) for details on the CP implementation.

Moreover, we observed that the scaling of the data affected the imputation quality. While it is out of the scope of the work to discuss such issue, we searched for the scaling that best suited the method for each dataset. The final scaling used is presented in Table 2.

5.5 Results

The main goal of this work is to study the misalignment problem, found when imputing our indoor climate dataset. Thus, within a first line of research, we are interested in addressing our indoor climate data case study. In particular, we aim at addressing the following questions:

Q1.A: Does the adjustment procedure improve the imputation accuracy in our case study?

Q1.B: Is the imputation accuracy obtained by the adjusted matrix decomposition competitive in our case study?

Additionally, we are also interested in understanding if the misalignment problem occurs in other types of datasets, and what factors may be causing it. Based on this, we also target the following question:

Q2: Does the adjustment procedure improve the imputation accuracy in other datasets?

To address **Q1.A**, we applied multiple decomposition methods (as exposed in Sect. 5.2), and for each imputation we applied our adjustment strategy. We note that the datasets were processed by season as it helps capture local dynamics (Xie et al., 2016). The results are depicted in Fig. 4, according to which:

- in TEMP the adjustment consistently led to more accurate imputations, with the exception of LATC, in which the adjustment effect was very reduced;
- in HUM the adjustment led to more accurate imputations in all settings but 72h in TRMF. Moreover, the level of improvement after the adjustment was higher for the gaps with the smaller sizes;
- in PRE, when considering NMF, TRMF and CP, the adjustment was only beneficial for gap sizes of at most 36 h. In the remaining methods, the adjustment always led to an improvement.

It should be noted that SVT accuracy was (comparably) poor in the three measurements, a result which is in accordance with comparison studies in literature (Khayati et al., 2020).

From a general perspective, we observed that the imputations refined through the adjustment process were more accurate in the generality of the scenarios, with the main exceptions occurring for gap sizes greater or equal than 48 timestamps in PRE measurements. It is noteworthy, that among the three measurements, pressure was the one in which the measurements were more similar across the multiple houses. As a consequence, the overfitting to the observed values may have less impact and the misalignment may be less severe.

These results show that the matrix decomposition imputation misalignment occurs in the indoor climate datasets, even when time-aware methods such as TRMF and LATC are

Table 2 Scaling used for each of the methods and datasets under study (./*s* scaling consists of dividing the data matrix by *s*)

	TEMP	HUM	PRE	Others
NMF	min–max	min–max	min–max	min-max
TRMF	none	./10	./50	–
SVT	min–max	min–max	min–max	–
LATC	none	none	./50	–
CP	min–max	min–max	min–max	–

considered. Moreover, the results also provide evidence that the adjustment approach is suitable.

Additionally, we were also interested in understanding how the level of improvement varies from the original imputation to the adjustment and how competitive are the adjusted imputations obtained by the different models. Although Fig. 4 already provides some insights, in order to facilitate the comparison across the methods, we re-organized the results in Figs. 5, 6 and 7, in which two baselines were also considered: linear interpolation and exponential weighted moving average (EWMA) (Moritz et al., 2019). It is noteworthy that we also tried machine learning based approaches, namely, GP-VAE (Fortuin et al., 2020). Nonetheless, the time required to carry out the experiments exceeded the time limit imposed (8 h). Still, we analysed the results acquired so far and observed that the imputation quality was poor. For example, for a gap size of 12 h, the RMSE obtained for TEMP was around 2.5, for HUM was around 10.0 and for PRE was around 9.0. Since these methods rely on large training sets, we believe that the small number of time-series available may had compromised the accuracy of these type of models. Finally, it should also be noted that the scales for original and adjusted plots are different to facilitate the visualization. Moreover, in this analysis we focus only on the cases for which the adjustment led to improvements.

By analysing the imputation error of the multiple decomposition models (without the adjustment), we observed that, generally, the most accurate methods were the ones modelling temporal dependencies (that is, TRMF and LATC). Nonetheless, when we applied the adjustment procedure to such imputations, we verified that the best accuracy was attained by NMF and CP. In other words, this means that the adjustment procedure introduced a higher improvement on non-time-aware approaches and that the resulting adjustments, when improving the original imputation, were generally more accurate than the ones obtained by time-aware methods and corresponding adjustments.

These results also allows us to answer **Q1.B**: we observed that, for most of the scenarios, the methods were more competitive regarding the baselines when the adjustment was applied (recall Figs. 5, 6 and 7).

Finally, we studied the misalignment problem in other datasets. In particular, if a strong misalignment occurs, our adjustment strategy is expected to improve the accuracy of the imputation, due to its gap re-allocation. Therefore, to tackle **Q2**, we applied our adjustment strategy to matrix decomposition in five other datasets. For the sake of simplicity, we considered only NMF. Moreover, similarly to the indoor climate datasets, IBERIAN_TEMP and WORLD_TEMP were also processed by season. In Fig. 8, we show the results for three of such datasets and we can observe that the results were very different across them. In particular, the adjustment procedure led to poor quality imputation in the IBERIAN_TEMP. We believe that, similarly to what occurred in PRE, the misalignment did not occur because the different time-series not only have an identical evolution (being highly correlated) but also assume an identical range of values. To deal with

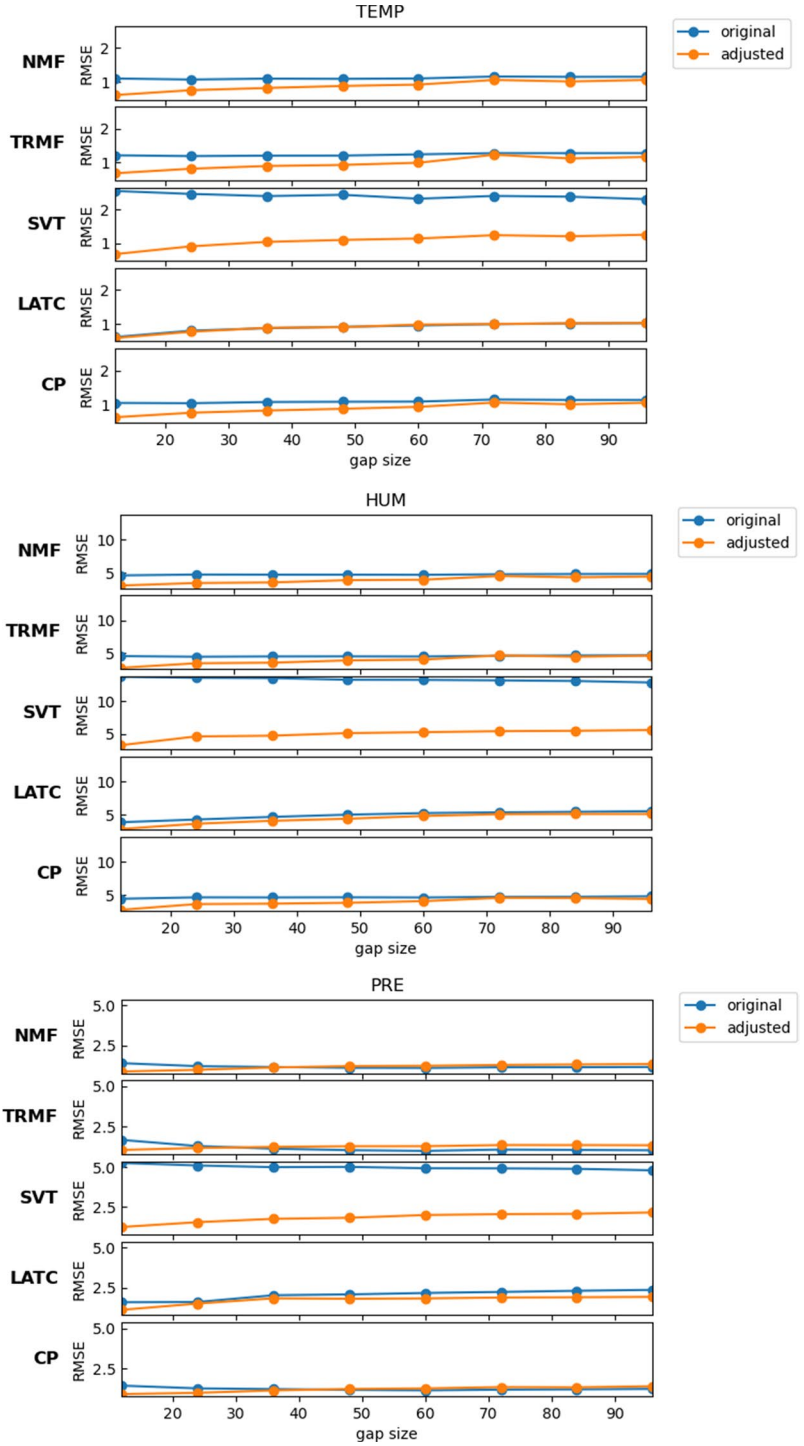

Fig. 4 RMSE on the original and the (post-processed) adjusted imputations for five distinct decomposition approaches (NMF, TRMF, SVT, LATC and CP) on the three indoor climate measurements (TEMP, HUM, and PRE)

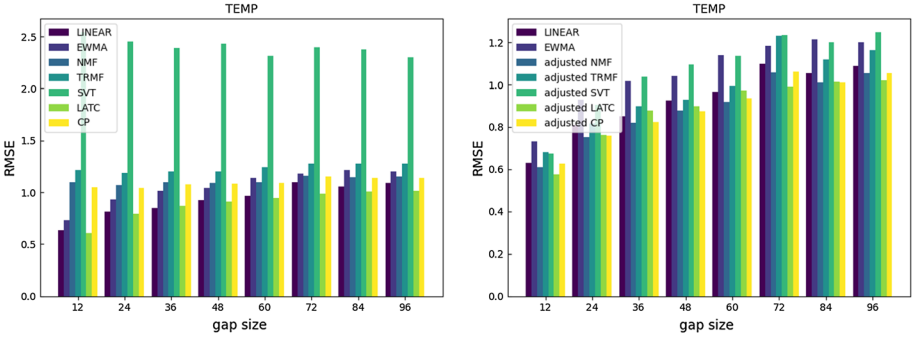

Fig. 5 RMSE per gap size on the linear, EWMA and original/adjusted (left/right) NMF, TRMF, SVT, LATC and CP imputations on the TEMP indoor climate measurement

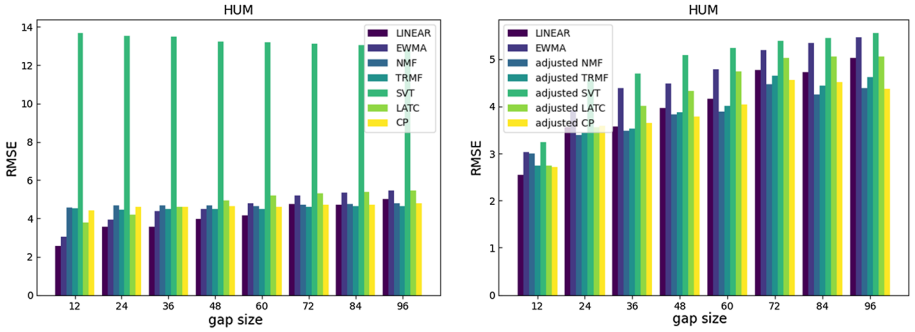

Fig. 6 RMSE per gap size on the linear, EWMA and original/adjusted (left/right) NMF, TRMF, SVT, LATC and CP imputations on the HUM indoor climate measurement

these cases, in which the misalignment does not occur, it would be important to incorporate in our method a strategy to infer whether the adjustment is needed or not based on the smoothness of the transitions between observed and imputed values. Nonetheless, such issue does not fall within the scope of this work and is left as future work. In WORLD_TEMP, we observed that the adjustment was useful when imputing small gaps and in GAS_SENSOR, the adjustment led to a general imputation accuracy improvement regardless of the gap size.

With the goal of having a deeper understanding of the scenarios in which the adjustment is beneficial, we carried out an additional study on the datasets of GUANGZHOU and SEATTLE, which have a larger number of time-series. In particular, we studied the adjustment need according to the number of time-series. In order to carry out this study, we considered a subset of the original datasets by randomly selecting $\hat{N}$ time-series, with $\hat{N} \in \{15, 60, 100, 200\}$. We considered 10 samples of subsets for each $\hat{N}$ so that we were able to perceive the dominant pattern. For the sake of simplicity, we exhibit the results of one of such samples per number of time-series considered (see Fig. 9). Nonetheless, we ensured that the sample presented reflected the behaviour *original vs. adjusted* observed across the majority of the samples. Interestingly, we observed the same pattern in both datasets: when considering only 15 timestamps, the adjustment always improved the original

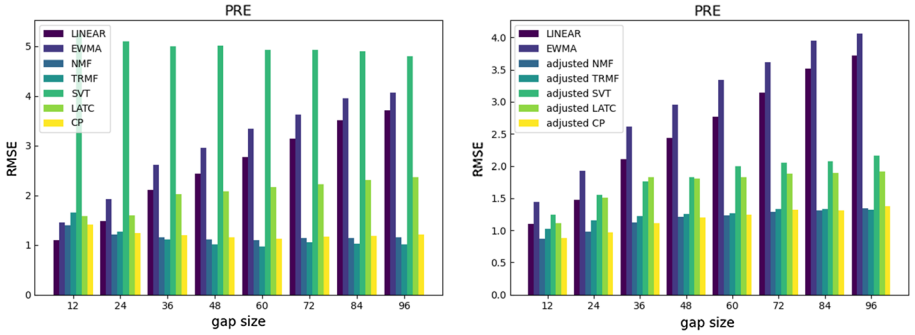

Fig. 7 RMSE per gap size on the linear, EWMA and original/adjusted (left/right) NMF, TRMF, SVT, LATC and CP imputations on the PRE indoor climate measurement

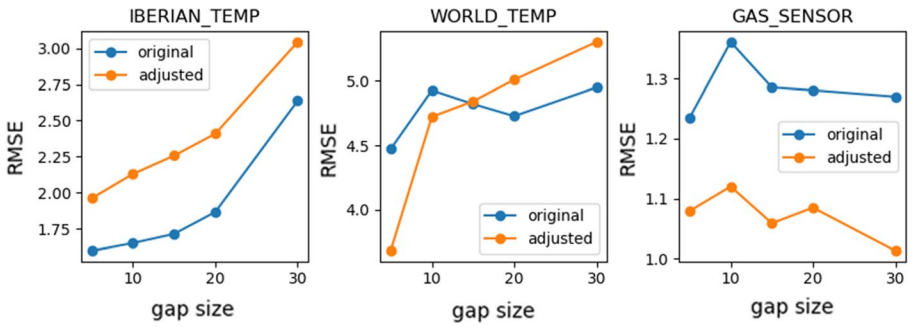

Fig. 8 RMSE per gap size on the original and the (post-processed) adjusted imputations for NMF in IBERIAN_TEMP, WORLD_TEMP and GAS_SENSOR datasets

imputation, but , as we increased the number of time-series in the datasets, the adjustment became less beneficial (especially in the larger gaps). These results reinforce the evidence that the misalignment problem arises from overfitting on datasets with few time-series.

General observations Since, when a strong misalignment occurs (such as in Fig. 1), shifting the imputation to the right range of values would substantially improve the imputation, then we can claim that poor adjustment quality is implicitly associated with few misalignment. Therefore, according to our experiments, we verified that:

- The misalignment problem was not algorithm specific, occurring in multiple decomposition techniques, even when those techniques incorporated temporal constraints.
- When there was high similarity (namely regarding the correlation and range of values) between the time-series, the misalignment was residual/not observed (as it occurred in the IBERIAN_TEMP case, in which the time-series had a pairwise correlation of around 0.97).
- The misalignment was more likely to occur in datasets with few time-series as it was empirically suggested by analysing the results on the traffic datasets: while the adjustment was always beneficial when considering 15 time-series, the same did not occur when considering a data matrix of 200 time-series (specially in SEATTLE).

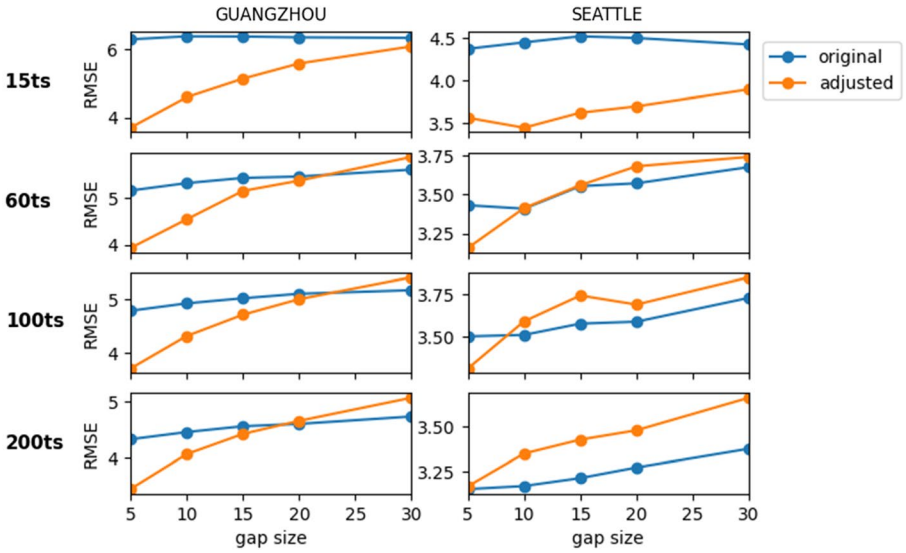

Fig. 9 RMSE per gap size on the original and the (post-processed) adjusted imputations for NMF in GUANGZHOU and SEATTLE datasets, for varying number of time-series (15, 60, 100 and 200)

6 Conclusions

As previously stated, with the emergence of the IoT there is an increase in the number of data sources available. Extracting relevant knowledge from these data sources requires not only analysis and processing tools, but also strategies to handle erroneous and missing data. In this work, we reported the misalignment problem of matrix decomposition with missing values, according to which matrix decomposition may misplace the imputations, thus compromising their quality. Moreover, we proposed an adjustment strategy to tackle this problem.

We carried out several experiments that demonstrated not only the existence of the misalignment problem in different scenarios but also the suitability of our adjustment approach. In this context, the experimental results provided some insights on the causes of misalignment. In particular, there was empirical evidence that this problem may derive from considering a small number of time-series. While one may argue that nowadays we deal mostly with Big Data and consequently, this problem does not hold, we believe that this work may be useful for small-scale studies and cases in which the access to data is limited.

Finally, with respect to future work, one of the main goals is to build a strategy that automatically decides the need for adjustment or not. We believe that such decision may be made based on the level of change between the observed values and between the observed and imputed values on the gap margins.

Funding This work is funded by FCT/MEC through national funds under the project PTDC/EEI-TEL/30685/2017.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Acar, E., Dunlavy, D. M., Kolda, T. G., & Mørup, M. (2011). Scalable tensor factorizations for incomplete data. *Chemometrics and Intelligent Laboratory Systems*, *106*(1), 41–56.

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, *20*(1), 40–49.

Balzano, L., Chi, Y., & Lu, Y. M. (2018). Streaming pca and subspace tracking: The missing data case. *Proceedings of the IEEE*, *106*(8), 1293–1310. https://doi.org/10.1109/JPROC.2018.2847041.

Banerjee, S., & Roy, A. (2014). *Linear algebra and matrix analysis for statistics*. CRC Press

Bao, Y., Fang, H., & Zhang, J. (2014). Topicmf: Simultaneously exploiting ratings and reviews for recommendation. *AAAI*, *14*, 2–8.

Cai, J. F., Candès, E. J., & Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, *20*(4), 1956–1982.

Cao, W., Wang, D., Li, J., Zhou, H., Li, L., & Li, Y. (2018). Brits: Bidirectional recurrent imputation for time series. In *Advances in neural information processing systems* (pp. 6775–6785).

Che, Z., Purushotham, S., Cho, K., Sontag, D., & Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, *8*(1), 1–12.

Chen, G., Liu, X.Y ., Kong, L., Lu, J.L., Gu, Y., Shu, W., & Wu, M. Y. (2013). Multiple attributes-based data recovery in wireless sensor networks. In *2013 IEEE global communications conference (GLOBECOM)* (pp. 103–108). IEEE.

Chen, X., & Sun, L. (2020). Low-rank autoregressive tensor completion for multivariate time series forecasting. 2006.10436

Chen, X., Chen, Y., & He, Z. (2018). Urban traffic speed dataset of guangzhou, China. https://doi.org/10.5281/zenodo.1205229.

Cichocki, A., & Phan, A. H. (2009). Fast local algorithms for large scale nonnegative matrix and tensor factorizations. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, *92*(3), 708–721.

Cui, Z., Ke, R., & Wang, Y. (2018). Deep bidirectional and unidirectional lstm recurrent neural network for network-wide traffic speed prediction. arXiv preprint arXiv:180102143

Cui, Z., Henrickson, K., Ke, R., & Wang, Y. (2019). Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting. *IEEE Transactions on Intelligent Transportation Systems*, *21*(11), 4883–4894.

De Brouwer, E., Simm, J., Arany, A., & Moreau, Y. (2019). Gru-ode-bayes: Continuous modeling of sporadically-observed time series. In *Advances in neural information processing systems* (pp. 7379–7390).

Duan, T. (2020) Lightweight python library for in-memory matrix completion. https://github.com/tonyduan/matrix-completion

Fonollosa, J., Rodríguez-Luján, I., Trincavelli, M., Vergara, A., & Huerta, R. (2014). Chemical discrimination in turbulent gas mixtures with mox sensors validated by gas chromatography-mass spectrometry. *Sensors*, *14*(10), 19336–19353.

Fortuin, V., Baranchuk, D., Rätsch, G., & Mandt, S. (2020). Gp-vae: Deep probabilistic time series imputation. In *International conference on artificial intelligence and statistics* (pp. 1651–1661). PMLR.

Karkouch, A., Mousannif, H., Moatassime, H. A., & Noel, T. (2016). Data quality in internet of things: A state-of-the-art survey. *Journal of Network and Computer Applications*, *73*, 57–81. https://doi.org/10.1016/j.jnca.2016.08.002.

Khayati, M., Lerner, A., Tymchenko, Z., & Cudré-Mauroux, P. (2020). Mind the gap: An experimental evaluation of imputation of missing values techniques in time series. *Proceedings of the VLDB Endowment*, *13*(5), 768–782.

Kim, Y., & Choi, S. (2009). Weighted nonnegative matrix factorization. In *2009 IEEE international conference on acoustics, speech and signal processing* (pp. 1541–1544).

Lepot, M., Aubin, J. B., & Clemens, F. H. (2017). Interpolation in time series: An introductive overview of existing methods, their performance criteria and uncertainty assessment. *Water*, *9*(10), 796.

Liu, Y., Yu, R., Zheng, S., Zhan, E., & Yue, Y. (2019). Naomi: Non-autoregressive multiresolution sequence imputation. In *Advances in neural information processing systems* (pp. 11238–11248).

Luo, Y., Zhang, Y., Cai, X., & Yuan, X. (2019). E2gan: End-to-end generative adversarial network for multivariate time series imputation. In *Proceedings of the 28th international joint conference on artificial intelligence* (pp. 3094–3100). AAAI Press.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., & Houston, T. G. (2012). An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, *29*(7), 897–910. https://doi.org/10.1175/JTECH-D-11-00103.1.

Menne, M. J., Durre, I., Korzeniewski, B., McNeal, S., Thomas, K., Yin, X., Anthony, S., Ray, R., Vose, R., Gleason, B. E., & Houston, T. G. (2020). Global historical climatology network-daily (ghcn-daily). Version 3.38, NOAA National Climatic Data Center. https://doi.org/10.7289/V5D21VHZ,.

Moritz, S., Moritz, M. S., & ByteCompile, T. (2019). Package "imputets". cran r-project org.

Santiago, A. R., Antunes, M., Barraca, J. P., Gomes, D., & Aguiar, R. L. (2019). SCoTv2: Large scale data acquisition, processing, and visualization platform. In *2019 7th International conference on future internet of things and cloud (FiCloud)*. IEEE. https://doi.org/10.1109/ficloud.2019.00053.

Shu, X., Porikli, F., & Ahuja, N. (2014). Robust orthonormal subspace learning: Efficient recovery of corrupted low-rank matrices. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3874–3881).

Stekhoven, D. J., & Bühlmann, P. (2011). MissForest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597.

Tarpey, T. (2000). A note on the prediction sum of squares statistic for restricted least squares. *The American Statistician*, *54*(2), 116–118.

Williams, A. H., Kim, T. H., Wang, F., Vyas, S., Ryu, S. I., Shenoy, K. V., Schnitzer, M., Kolda, T. G., & Ganguli, S. (2018). Unsupervised discovery of demixed, low-dimensional neural dynamics across multiple timescales through tensor component analysis. *Neuron*, *98*(6), 1099–1115.

Xie, K., Ning, X., Wang, X., Xie, D., Cao, J., Xie, G., & Wen, J. (2016). Recover corrupted data in sensor networks: A matrix completion solution. *IEEE Transactions on Mobile Computing*, *16*(5), 1434–1448.

Yu, H.F., Rao, N., & Dhillon, I. S. (2016). Temporal regularized matrix factorization for high-dimensional time series prediction. In *Advances in neural information processing systems* (pp. 847–855).

Zhang, D., & Balzano, L. (2016). Global convergence of a grassmannian gradient descent algorithm for subspace estimation. In *AISTATS* (pp. 1460–1468).

Zhao, Q., Zhang, L., & Cichocki, A. (2015). Bayesian cp factorization of incomplete tensors with automatic rank determination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(9), 1751–1763.