Automation in Construction 119 (2020) 103322

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Fusing imperfect experimental data for risk assessment of musculoskeletal disorders in construction using canonical polyadic decomposition

Amrita Dutta[a], Scott P. Breloff[b], Fei Dai[a,*], Erik W. Sinsel[b], Robert E. Carey[b], Christopher M. Warren[b], John Z. Wu[b]

[a] Department of Civil and Environmental Engineering, West Virginia University, P.O. Box 6103, Morgantown, WV 26506, United States of America
[b] National Institute for Occupational Safety and Health, 1095 Willowdale Road, Morgantown, WV 26505, United States of America

ARTICLE INFO

Keywords:
Tensor decomposition
Risk assessment
Data fusion
Construction safety

ABSTRACT

Field or laboratory data collected for work-related musculoskeletal disorder (WMSD) risk assessment in construction often becomes unreliable as a large amount of data go missing due to technology-induced errors, instrument failures or sometimes at random. Missing data can adversely affect the assessment conclusions. This study proposes a method that applies Canonical Polyadic Decomposition (CPD) tensor decomposition to fuse multiple sparse risk-related datasets and fill in missing data by leveraging the correlation among multiple risk indicators within those datasets. Two knee WMSD risk-related datasets—3D knee rotation (kinematics) and electromyography (EMG) of five knee postural muscles—collected from previous studies were used for the validation and demonstration of the proposed method. The analysis results revealed that for a large portion of missing values (40%), the proposed method can generate a fused dataset that provides reliable risk assessment results highly consistent (70%–87%) with those obtained from the original experimental datasets. This signified the usefulness of the proposed method for use in WMSD risk assessment studies when data collection is affected by a significant amount of missing data, which will facilitate reliable assessment of WMSD risks among construction workers. In the future, findings of this study will be implemented to explore whether, and to what extent, the fused dataset outperforms the datasets with missing values by comparing consistencies of the risk assessment results obtained from these datasets for further investigation of the fusion performance.

1. Introduction

Work-related musculoskeletal disorders (WMSDs) are one of the most common causes of days away from work and physical disabilities in the construction industry [1]. An increased exposure to risk factors in the workplace can enhance the likelihood of WMSDs; hence, proper identification of possible risk exposures and developing injury prevention strategies are essential to alleviate WMSDs.

Collecting risk exposure data with human subject involvement is most often accepted as the gold standard for understanding risky behaviors and conditions that may expose workers to WMSD risks on construction sites [2]. Generally, these data are collected in laboratory settings or real construction sites by technologies such as optical motion capture systems or surface electromyography sensors. However, data collected from these technologies often suffer from 'drop out', a phenomenon in which data is missing due to technology-induced errors (e.g., disconnection of sensors, errors in communicating with the database server, instrument failures), human-induced errors (e.g., accidental human omission) or other unknown reasons [3]. The result is incompleteness of the collected risk exposure data that may lead to invalid conclusions on the effects of the potential WMSD risk factors. Missing data is a common problem associated with data collection in ergonomic risk assessment using technologies, regardless of the quality of the research design [4]. Therefore, it should be carefully handled. In doing so, reserving the interrelation among the potential risk factors and the risk indicators across multiple datasets is necessary. Among several benefits of data fusion, one is revealing the latent pattern of the data and leveraging collaborative relationships among various factors within multiple datasets based on that pattern. This benefit can be utilized to reserve the interrelation among different factors and the risk indicators across the datasets to fill in the missing data. This study proposes a method for dealing with multiple imperfect and incomplete datasets by applying a Canonical Polyadic Decomposition (CPD) technique to treat the imperfect data for WMSD risk assessment. CPD

* Corresponding author.
E-mail addresses: amdutta@mix.wvu.edu (A. Dutta), sbreloff@cdc.gov (S.P. Breloff), fei.dai@mail.wvu.edu (F. Dai), ESinsel@cdc.gov (E.W. Sinsel), rcarey@cdc.gov (R.E. Carey), cpw4@cdc.gov (C.M. Warren), ozw8@cdc.gov (J.Z. Wu).

https://doi.org/10.1016/j.autcon.2020.103322
Received 14 February 2020; Received in revised form 28 May 2020; Accepted 9 June 2020

decomposes the incomplete datasets based on the latent relationship among different risk factors and the risk indicators, then reconstructs a new dataset through fusion as a high-order tensor [5]. This newly reconstructed dataset is referred to as fused dataset, which can then be used for assessing the risk of WMSDs. To validate the effectiveness of the CPD-based method in assessing WMSDs, two WMSD risk-related datasets collected from prior experimental studies (original datasets) were intentionally modified to represent incomplete datasets. Then CPD was applied for fusion and to reconstruct the fused datasets. The risk assessment results obtained using the fused datasets were further compared to those obtained by using the original datasets to evaluate the performance of the fusion treatment.

2. Background

2.1. Importance of research

Missing data is a common problem in research studies that involve human subjects and technologies for data collection. They can reduce statistical power of a study and lead to erroneous conclusions [6]. To potentially mitigate this issue, the sample size for data collection is typically increased. However, this is not always possible due to research design, limitations in budget and human resources. It is not always feasible to regenerate the data by repeating the experiment, as it can be costly and time-consuming. There are existing methods of handling missing data in literature such as simply omitting observations with the missing data [7], regression imputation [8], mean substitution [9], replacing the missing data with the last observed values [10], and estimating the missing data using conditional distribution of the other variables [11]. While useful, these methods may not be optimal for tackling data missing in WMSD risk assessment. Simply omitting observations with missing data typically decreases the sample size and may adversely affect the statistical power [7]. Regression imputation predicts a missing value from other variables, but it barely adds any new information except increasing the sample size and compromising the standard error [8]. If there is a great inequality in the proportion of missing values of different variables, mean substitution may lead to inconsistent biases and underestimation of errors [9]. Replacing missing data with the last observed values assumes that there will be no changes in the outcome [10]. Using conditional distribution of the other variables becomes inappropriate when a large amount of data is missing, as in this situation, the relationship among the variables needs to be properly computed [7,11].

Moreover, in real life scenarios, a phenomenon can be described and characterized with a set of factors. Each of these factors is referred to as the ‘dimension’ of that phenomenon [12]. When a research problem includes multiple dimensions, it is considered ‘multidimensional’. This is often the case in WMSD risk studies, in which risk exposure data can be collected in terms of different measurements, often referred to as risk indicators, and working conditions for the purpose of in-depth understanding of risks. The working conditions may include various work settings and postures, and the risk indicator measurements may include kinematics, kinetics and electromyography (EMG) measurements. Quite often, multiple risk indicators can provide a more detailed insight into the worker's risk-inducing behavior, compared to what a single risk indicator can provide as these multiple risk indicators may be correlated. In WMSD risk studies that involves analyzing the effects of working conditions on risk indicators, the working conditions and the risk indicators can be termed as a ‘dimension’ of the risk phenomenon. This type of data can be considered as high-dimensional data. When data is missing in multiple high-dimensional risk-related datasets, the existing methods will not be suitable as the interrelation among different risk indicators and potential risk factors across multiple datasets may not be well captured. Fuzzy modeling has been used in multidimensional missing data imputation [13]. However, for high-dimensional datasets where capturing the latent relationship among various dimensions is essential during imputation, the performance of Fuzzy algorithms degrades [14,15]. Besides, the performance of Fuzzy models in dealing with multiple datasets is yet to be tested. In contrast, data fusion can be an efficient method in this regard that integrates information based on the interrelation among the factors and risk indicators in such a way that it can produce a more complete representation of the measurements of the risk indicators, compared to that of an incomplete dataset [16].

2.2. State of research on data fusion in construction

In the construction industry, data fusion methods have been used for automated identification, location estimation, dislocation detection of construction materials in jobsites [17,18], automated progress tracking of construction projects [19], structural health monitoring [20], and damage identification of civil structures [21]. In construction ergonomics, a computationally efficient approach using data fusion was developed to recognize construction workers' awkward postures [22]. Fusion of data from continuous remote monitoring of construction workers' location and physiological status was used to identify safe and unsafe behaviors of construction workers [23]. Fusion of spatio-temporal and workers' thoracic posture data has been done for understanding the worker's activity type for productivity assessment [24]. A position and posture data fusion method was proposed for evaluation of construction workers' behavioral risks [25]. However, no study has been done that uses data fusion of incomplete risk-related datasets for assessing WMSD risks.

2.3. State of research on tensor decomposition for data fusion

To integrate multidimensional data, it is essential to learn the relationships between those dimensions to understand the multidimensional nature of a phenomenon [12]. There are currently several existing approaches for analyzing multiple datasets. Zheng et al. [26] summarized the methods of fusing multiple datasets into three categories. The first category is stage-based fusion methods, where different datasets are used at different stages of data mining tasks [27]. The second category is deep learning-based fusion methods that use a neural network to extract original features from multiple datasets and learn a new representation of those features for classification and prediction purposes [28]. The third category is semantic meaning-based fusion methods that identify the association between different features across multiple datasets and carry the semantic meaning, i.e., the latent relationship between the features during fusion [26]. Similarity-based data fusion is one type of the semantic meaning-based fusion methods, in which similarity between the features are measured from multiple datasets. These similarities can leverage the correlation among the features collectively and help obtain any missing information of a feature based on the information available for another feature. Based on the similarities, different datasets can be fused [26]. This study is interested in fusion of risk-related datasets containing different features that represent WMSD risk factors in construction and are inherently related to each other. Understanding the latent relationship between different features across multiple datasets plays a vital role in proper fusion. Also, learning the similarities between those features is useful to impute the missing values. As a result, the similarity-based data fusion methods lend themselves well to tackling the fusion problem in this study.

Tensor decomposition is one type of similarity-based data fusion methods [29]. Tensors are generalizations of matrices to higher dimensions, and are powerful to model multidimensional data [30]. Tensors are multidimensional arrays of numerical values that are widely used in different applications in data analysis and machine learning, including filling in missing values [31], anomaly detection [32], object profiling [33], discovering patterns [34], and predicting evolution [35]. Matrix factorization was previously used in filling out

missing data [36]; however, it only works with two-dimensional data and may not be applicable to high-dimensional data like those used in WMSD risk studies. Tensor decomposition is a useful tool to deal with high-dimensional datasets. It accurately extracts the correlations among various dimensions from different datasets and learns the latent structures and collaborative relationships among the dimensions to approximate the pattern of the data [37].

With increases in dimensions of the datasets, the number of the elements in a tensor also increases. This increase in elements makes the dataset difficult to deal with in terms of computational and memory requirements. By decomposing the high-dimensional data presented in form of a tensor, the problems associated with high-dimensionality can be alleviated or even removed [38]. Also, tensor decomposition can be used as an efficient tool for missing value prediction [39,40]. Though tensor decomposition has been abundantly applied in areas such as scientific computing, signal processing, and social media data analysis, its use in construction industry-related datasets is limited. One application of tensor decomposition with construction datasets was concerned with awkward posture recognition in which worker's motion data was presented as high order tensor [22]. However, this technique has yet to be established as an alternative method of treating imperfect experimental data to provide meaningful and accurate risk assessment results in general WMSD risk assessment studies.

The two most widely used tensor decomposition algorithms are *canonical polyadic decomposition* (CPD) and *Tucker decomposition* (TD). CPD is generally advised for use with latent parameter estimation, while the TD is suggested in use of subspace estimation, compression, and dimensionality reduction [5]. As the objective of this study is to fuse two incomplete datasets based on the latent correlations among all dimensions of those datasets, CPD is more suitable than TD in fulfilling the purpose of this study. Moreover, CPD is very effective in capturing the interactions among various dimensions of a high-dimensional datasets and therefore, it can effectively be used for imputing missing data [31]. Considering many other tensor decomposition techniques are based on CPD and TD and CPD is more suitable for this study, the following subsection will only discuss CPD for tensor decomposition.

2.3.1. Canonical polyadic decomposition (CPD)

In order to impute missing data, CPD extrapolates the latent structures and collaborative relationships among different dimensions, such as rows and columns of a tensor. The CPD method factorizes a tensor into a sum of component rank-one tensors [37]. For example, in CPD, a given three-dimensional tensor $\boldsymbol{Y} \in \mathbb{R}^{I\times J\times K}$ can be written as:

$$\boldsymbol{Y} \approx \sum_{r=1}^{R} \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r \equiv [\![\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]\!] \tag{1}$$

where R is a positive integer, which is referred to as the rank of a tensor. The rank of tensor $\boldsymbol{Y} = R$ can be defined as the smallest number of rank-1 tensor that is required to represent the tensor $\boldsymbol{Y}$ as their sum [5] (Fig. 1); Rank-1 tensor is defined as a decomposition of an N-dimensional tensor into one outer product of N vectors. R is also called latent factor. I, J and K are the sizes of the dimensions of the tensor $\boldsymbol{Y}$ where, $\boldsymbol{a}_r \in \mathbb{R}^I$, $\boldsymbol{b}_r \in \mathbb{R}^J$ and $\boldsymbol{c}_r \in \mathbb{R}^K$ for $r = 1\ldots R$. '$\circ$' indicates the vector outer product. $\boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r \in \mathbb{R}^{I\times J\times K}$ is an outer product of three vectors $\boldsymbol{a}_r$, $\boldsymbol{b}_r$ and $\boldsymbol{c}_r$ and is referred to as a rank-1 tensor. $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ are the factor matrices obtained for each dimension after decomposition of the tensor $\boldsymbol{Y}$, and are referred to as the combination of the vectors from the rank-1 components. Here, the factors matrices $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_R] \in \mathbb{R}^{I\times R}$, $\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_R] \in \mathbb{R}^{J\times R}$ and $\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_R] \in \mathbb{R}^{K\times R}$. Their column vectors are expressed explicitly as: $\boldsymbol{a}_r = [a_{1r}, a_{2r}, a_{3r}, \ldots, a_{Ir}]^T$, $\boldsymbol{b}_r = [b_{1r}, b_{2r}, b_{3r}, \ldots, b_{Jr}]^T$, $\boldsymbol{c}_r = [c_{1r}, c_{2r}, c_{3r}, \ldots, c_{Kr}]^T$.

Elementwise, Eq. (1) can be expressed as:

$$Y_{ijk} \approx \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} \text{ for } i = 1\ldots I;\ j = 1\ldots J;\ k = 1\ldots K. \tag{2}$$

CPD of a 3-dimensional tensor can be formalized as follows:

$$\min_{\hat{\boldsymbol{Y}}} \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|; \text{ Where, } \hat{\boldsymbol{Y}} = \sum_{r=1}^{R} \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r \equiv [\![\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]\!] \tag{3}$$

where $\boldsymbol{Y}$ is the original tensor and $\hat{\boldsymbol{Y}}$ is a low rank approximation of tensor $\boldsymbol{Y}$. CPD of a tensor is typically computed using the alternating least square (ALS) algorithm where each factor matrix is iteratively solved using a least square method. This algorithm fixes all factor matrices, except one, to optimize for the non-fixed matrix and then repeats this procedure for each matrix repeatedly until it converges [37,41].

The CPD of a three-dimensional tensor is illustrated in Fig. 1.

3. Problem statement and research objective

For assessing WMSD risks among construction workers, human-based data can potentially suffer from missing data points due to dropout from the data collection technology. However, an in-depth method in handling multiple imperfect datasets for assessing WMSD risks is missing in the existing literature. Tensor decomposition-based data fusion can be potentially useful in this regard. It may help understand the data distribution of each dataset and consider the correlation among risk indicators captured at multiple work settings or conditions to fill in the missing data points. Therefore, the objective of this research is to develop a method that applies CPD tensor decomposition techniques to fuse multiple imperfect datasets and replace missing data by considering the correlation among the risk indicators for assessing the WMSD risks among construction workers.

4. Proposed method

Fig. 2 provides a schematic overview of the proposed method. First, multiple risk-related incomplete datasets captured in multiple experimental settings are represented as high-dimensional tensors. Then these tensors are fused by applying the CPD tensor decomposition. The CPD first decomposes these tensors into factor matrices that represent the latent structures and collaborative relationships among all dimensions. Based on the correlation among different dimensions, the CPD then reconstructs a new dataset with all missing values filled. This fused new

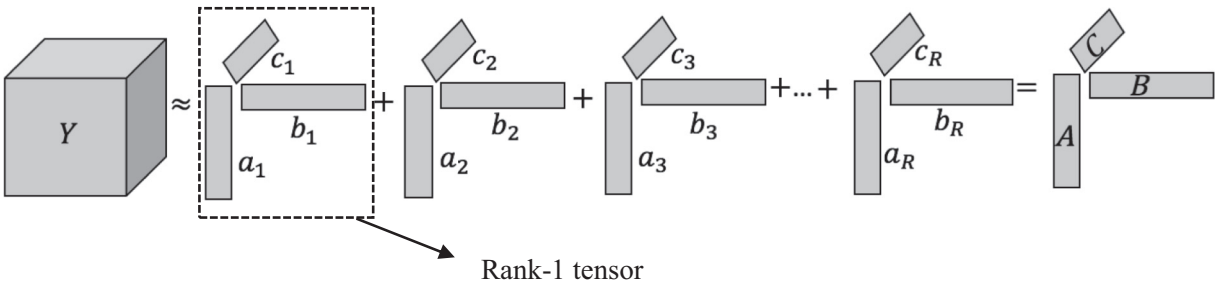

Fig. 1. CPD tensor decomposition.

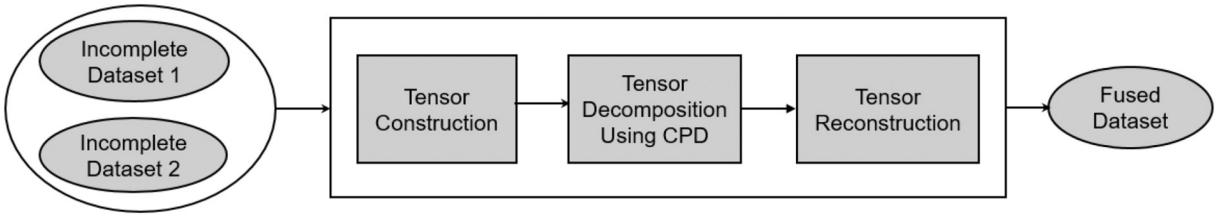

Fig. 2. Overview of proposed method.

dataset is ready to be used for subsequent risk assessments. In the following subsections, the steps of the proposed method are described in detail.

4.1. Incomplete datasets for WMSD risk assessment

For WMSD risk assessment in construction, the effects of various work-related factors on specific risk indicators are often measured. Therefore, a typical WMSD risk-related dataset contains information such as work settings/experimental conditions and specific risk indicators. Often, a WMSD risk may not be characterized with a single type of risk indicator and in that case, multiple types of WMSD risk indicators are measured, e.g., kinematics (flexion, abduction/adduction, internal/external rotation), muscle activity data recorded by EMG, and kinetic data recorded by force plates, and can be available in multiple datasets.

To describe the proposed method, consider two datasets DS1 and DS2 (denoted for *Incomplete Dataset 1* and *Incomplete Dataset 2* in Fig. 2). Each dataset contains two different types of WMSD risk indicators collected from several subjects, investigating various work-induced risks for a number of work settings (e.g., working technique, working posture, working condition). This is common practice of WMSD risk assessment studies in construction, where different types of risk indicators are often measured from the same group of people and the same work settings for an in-depth understanding of a specific type of risk. In this case, a data value in DS1 or DS2 represents a value of a risk indicator measured from a subject at certain work settings. For example, to measure the knee WMSD risk, two common types of risk indicators are measured – awkward knee rotations and maximum EMG of knee postural muscles. Awkward knee rotations may include flexion, abduction/adduction, and internal/external rotations. Maximum EMG of knee postural muscles may include measurements of maximum flexor and extensor muscle activations. The datasets can be represented in a tabular form as shown in Table 1.

In Table 1, each dataset contains a type of risk indicator measurements collected for several subjects at a number of work settings. The column *subject* contains subjects' IDs. The columns $setting_1$, $setting_2$, …, $setting_q$ contain arrangements of q work settings at which measurements of m risk indicators in columns $ind_1{}^1$, $ind_1{}^2$, …, $ind_1{}^m$ are collected in DS1, and n risk indicators are collected in columns $ind_2{}^1$, $ind_2{}^2$, …, $ind_2{}^n$ in DS2. For example, the column $setting_1$ can represent working postures that have two arrangements – static and dynamic. The column $setting_2$ can represent slopes of the working site that have three arrangements – 0°, 15° and 30°. So, in the column $setting_1$, the values are either 'static' or 'dynamic' and in the column $setting_2$, the values are either '0°', '15°' or '30°'. If a study involves only these two settings, then the value of q will be 2. Similarly, the columns $ind_1{}^1$, $ind_1{}^2$, …, $ind_1{}^m$ can contain kinematics measures (flexion, abduction/adduction, internal/external rotation) and $ind_2{}^1$, $ind_2{}^2$, …, $ind_2{}^n$ can contain EMG measures of several muscles. To be more specific with an example, in DS1, a value of any of the m risk indicator columns can be a certain measurement value of a risk indicator (e.g., flexion) captured for certain subject at certain arrangements of q work settings. Note that in both datasets, the values of columns of *subject* and $setting_1$, $setting_2$, …, $setting_q$ are identical; the only difference is the measurements of the risk indicators. These datasets become incomplete when a significant amount of risk indicator measurements are not captured during data collection process.

Table 1
Formats of DS1 and DS2.

Dataset (DS)	Tabular format
DS1	*subject*, $setting_1$, $setting_2$, …, $setting_q$, $ind_1{}^1$, $ind_1{}^2$, …, $ind_1{}^m$
DS2	*subject*, $setting_1$, $setting_2$, …, $setting_q$, $ind_2{}^1$, $ind_2{}^2$, …, $ind_2{}^n$

4.2. Tensor construction

Since the datasets contain multiple risk indicator measurements for the same combination of subjects and settings, the risk indicator measurements from both datasets are concatenated, yielding a new dataset (DS3) as illustrated in Table 2. This process is called early integration data fusion [36].

In Table 2, each risk indicator measurement in DS3 is collected for certain subject at different arrangements of q work settings. Thus, 'subject' and each 'work setting' are considered as dimensions of each risk indicator, resulting in a total of $(1 + q)$ dimensions. For multiple risk indicators, this method allows datasets to be represented as a 'multi-dimensional' or 'multi-modal' tensor, where each cell of the tensor contains a value of a certain risk indicator, measured for a subject at certain arrangements of q work settings. In this study, the 'subject', each 'work setting', and the entire set of 'risk indicators' can be considered as dimensions of the tensor which has a total of $(1 + q + 1)$ dimensions.

To illustrate this tensor representation with a simplified example, consider only one work setting (e.g., working postures) denoted by '*setting*'. In this case, DS3 can be represented as a three-dimensional tensor (*subject* × *risk indicators* × *setting*) $(\boldsymbol{M}_{i,\,j,\,k}) \in \mathbb{R}^{s \times r \times t}$ (Fig. 3), where i, j and k correspond to dimensions *'subject', 'risk indicators'* and *'setting'* respectively; and s, r and t are their sizes where $i \in \{1,2, \ldots s\}$, $j \in \{1,2, \ldots r\}$ and $k \in \{1,2, \ldots t\}$. Each slice along the *setting* axis in Fig. 3 represents a matrix $(\boldsymbol{M}_{i,\,j}) \in \mathbb{R}^{s \times r}$ given a setting arrangement k.

4.3. Tensor decomposition

Following tensor construction, CPD is applied to decompose the tensor into factor matrices. As mentioned earlier, CPD learns the latent structures and collaborative relationships among the dimensions of a tensor. Thus, each factor matrix represents the latent relationships between the corresponding dimension and all the other dimensions captured by latent factor R (previously discussed in Section 2.3.1). Before applying CPD, the latent factor R should be computed properly to ensure unique mapping of the original tensor to the decomposed tensor in the form of factor matrices. In the above example, $(\boldsymbol{M}_{i,\,j,\,k}) \in \mathbb{R}^{s \times r \times t}$ is

Table 2
Format of DS3.

Dataset (DS)	Format
DS3	*subject*, $setting_1$, $setting_2$, …, $setting_q$, $ind_1{}^1$, …, $ind_1{}^m$, $ind_2{}^1$, …, $ind_2{}^n$

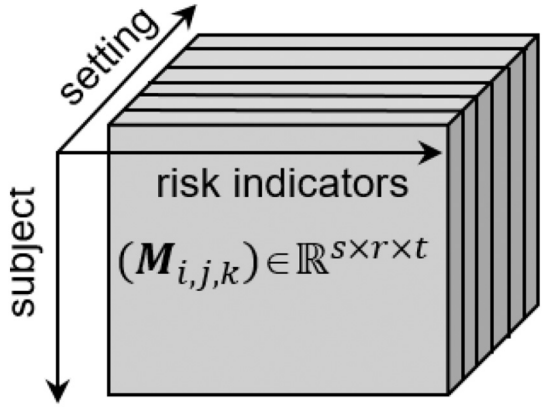

Fig. 3. Schematic representation of DS3 as a 3D tensor $\boldsymbol{M}$.

associated with three dimensions – '*subject*', '*risk indicators*' and '*setting*'; hence, three factor matrices are obtained after decomposition, denoted as $\boldsymbol{FM_1} \in \mathbb{R}^{s \times R}$, $\boldsymbol{FM_2} \in \mathbb{R}^{r \times R}$, $\boldsymbol{FM_3} \in \mathbb{R}^{t \times R}$, as shown in Fig. 4. Here, the factor matrix $\boldsymbol{FM_1}$ obtained for the dimension '*subject*' represents the latent relationships between the subject and the risk indictors collected at different arrangements of the setting. $\boldsymbol{FM_2}$ is obtained for the dimension '*risk indicators*' and represents the latent relationships between the risk indicators and the subjects at different arrangements of the setting for which those risk indicators were collected. $\boldsymbol{FM_3}$ is obtained for the dimension '*setting*' and represents the latent relationships between different arrangements of the setting and the risk indicators collected for the subjects at those arrangements of the setting. CPD leverages all the inter-dimensional correlations among all the dimensions of a tensor and performs data fusion based on those correlations to create a reconstructed tensor. The structures of the factor matrices and the method of computing R will be discussed further in "Implementation and results".

4.4. Tensor reconstruction

After tensor decomposition, the resulting factor matrices are used to reconstruct a new fused tensor where the missing values are filled and values of the WMSD risk indicators are fused based on the inter-dimensional correlations. In the prior example, such tensor can be denoted as $(\boldsymbol{M'}_{i,\ j,\ k}) \in \mathbb{R}^{s \times r \times t}$, which is computed using the following equation:

$$\boldsymbol{M'}_{i,j,k} = \sum_{r=1}^{R} \boldsymbol{FM_{1r}} \circ \boldsymbol{FM_{2r}} \circ \boldsymbol{FM_{3r}} \tag{4}$$

where $\boldsymbol{FM_{1r}}$, $\boldsymbol{FM_{2r}}$, $\boldsymbol{FM_{3r}}$ are the column vectors of $\boldsymbol{FM_1}$, $\boldsymbol{FM_2}$ and $\boldsymbol{FM_3}$ respectively, R is the latent factor and '$\circ$' is the vector outer product. Using the newly reconstructed tensor, values of the WMSD risk indicators for different subjects at different settings are extracted and reorganized into a tabular form, readily available for risk assessment.

5. Implementation and results

5.1. Original datasets collected from previous experiments

The current study considered two risk-related datasets that were collected from the authors' prior human subject laboratory experimental studies. These studies assessed work-related factors for knee WMSDs among residential construction roofers who work on sloped environments. One dataset contains calculated knee rotation (kinematics) data, representing five knee rotational angles – flexion, abduction, adduction, internal and external rotation [42]. The second data set contains EMG data, describing muscle activation of five knee postural muscles – biceps femoris, recus femoris, semitendinosus, vastus lateralis and vastus medialis [43]. Awkward knee rotations and maximum muscle activation have been identified as two quantitative risk indicators of WMSDs in low extremities [44,45]. The five knee rotation angles represent the WMSD risk associated with awkward and extreme kneeling postures, while the EMG data represent the WMSD risk associated with heightened activation of knee postural muscles that might cause knee joint overloading. Both types of data were collected from the experimental study in which 9 subjects simulated the shingle installation roofing task at three roof slopes (0°, 15° and 30°) with two kneeling postures (static and dynamic). Each participant performed the task for five trials. Although the experiment was designed and implemented carefully, both datasets initially collected from the experiment had missing values. About 2% data were missing in the knee rotation dataset and about 20% data were missing in the EMG dataset at random. In this study, observations with missing data were intentionally removed from the initial datasets for the sake of obtaining complete datasets for evaluation. In the knee rotation dataset, each data point represents the maximum angle value of a knee rotation for a certain subject at a specific slope, posture and trial. In the EMG dataset, each data point represents the maximum normalized EMG value of a knee postural muscle of a certain subject at a specific slope, posture and trial. Since only maximum measurements in both datasets were collected as risk indicators, the two datasets could be easily synchronized.

5.2. Imperfect (or incomplete) datasets construction (datasets with missing values)

To evaluate the performance of the proposed method in dealing with missing data, this study was designed to fuse knee rotation and EMG datasets for different proportions of missing values using the CPD method. A proportion of data was randomly removed from both knee kinematics and EMG datasets with percentages of missing data as shown in Table 3. Therefore, fourteen incomplete datasets—seven for knee rotations and seven for EMG—were constructed with different portions of missing values.

5.3. Tensor construction

The incomplete rotation and EMG datasets were then used for tensor construction. As the knee rotation and EMG data were collected from the same subjects at the same work settings (i.e., roof slope, working posture, and trial), they could be combined by the early integration

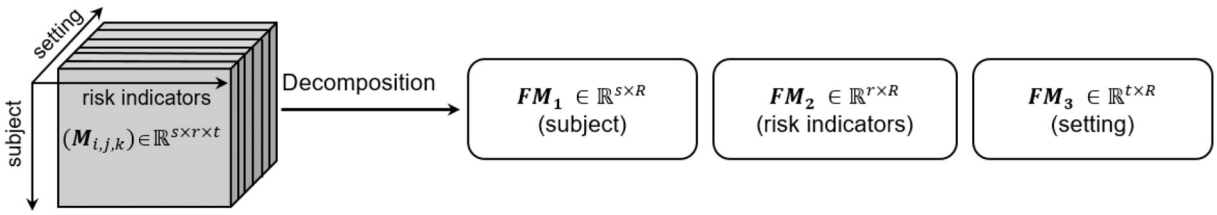

Fig. 4. Schematic illustration of CPD decomposition of the 3D tensor $\boldsymbol{M}$.

Table 3
Proportion of missing data.

Datasets	Missing proportion (%)						
Knee kinematics data	10	20	30	40	50	60	70
Knee EMG data	10	20	30	40	50	60	70

Table 4
Incomplete datasets for tensor construction.

Missing proportion from kinematics data (%)	Missing proportion from EMG data (%)						
10	10	20	30	40	50	60	70
20	10	20	30	40	50	60	70
30	10	20	30	40	50	60	70
40	10	20	30	40	50	60	70
50	10	20	30	40	50	60	70
60	10	20	30	40	50	60	70
70	10	20	30	40	50	60	70

method [46]. In this method, the columns of these two types of risk indicators were concatenated to form a single set of length ten vectors which was referred as 'risk indicators' in this study. For tensor construction, the following forty-nine incomplete datasets as shown in Table 4 were considered.

An N dimensional tensor $\boldsymbol{T}_N \in \mathbb{R}^{I_1 \times I_2 \times \ldots\ldots \times I_N}$ has size $(\boldsymbol{T}_N) = I_1 \times I_2 \times \ldots\ldots \times I_N$, where, I_i is the size of its i^{th} dimension [31]. As the risk indicators were measured from multiple subjects at different working postures, roof slopes and multiple trials, the collected data could be represented as a five-dimensional sparse tensor $\boldsymbol{T} \in \mathbb{R}^{I_1 \times I_2 \times I_3 \times I_4 \times I_5}$, where I_1 = No. of subjects (total 9), I_2 = No. of risk indicators [EMG and kinematics; total 10 (5 knee rotation angles and 5 knee muscle EMG)], I_3 = No. of trials (total 5), I_4 = No. of postures [total 2 (static and dynamic)], I_5 = No. of slopes [total 3 (0°,15°,30°)]. In short, the tensor $\boldsymbol{T}$ maps the knee rotations and EMG measurements of nine subjects, obtained from five trials, performed on three roof slopes, using two postures. Tensor $\boldsymbol{T}$ was referred to as 'incomplete tensor' in this study.

For a certain posture and slope, the sparse tensor $\boldsymbol{T}$ is denoted as $\boldsymbol{T}_{i,j}$ which represents the values of the risk indicators obtained from five trials performed by each subject at that roof slope and posture. In $\boldsymbol{T}_{i,j}$, $i \in \{1,2\}$,with 1 and 2 representing the static and dynamic postures respectively; $j \in \{1,2,3\}$,with 1, 2 and 3 representing the 0°, 15° and 30° slopes respectively. Then, the five-dimensional tensor $\boldsymbol{T}$ can be illustrated in Fig. 5.

Similarly, the original datasets collected from the previous experiments were constructed to a five-dimensional tensor $\boldsymbol{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3 \times I_4 \times I_5}$ as the benchmark data, which was referred to as the "original tensor" in this study and was used for validation of performance of the fusion method.

5.4. Tensor decomposition

Once a tensor $\boldsymbol{T}$ was constructed, the CPD tensor decomposition was applied to decompose the tensor $\boldsymbol{T}$ into five factor matrices, representing information in each dimension. The factor matrices were represented as matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$, and $\boldsymbol{E}$, providing latent information in dimensions of subjects, risk indicators, trials, postures, and slopes, respectively.

CPD factorizes the five-dimensional sparse tensor $\boldsymbol{T} \in \mathbb{R}^{I_1 \times I_2 \times I_3 \times I_4 \times I_5}$ as:

$$\boldsymbol{T} \approx \sum_{r=1}^{R} \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r \circ \boldsymbol{d}_r \circ \boldsymbol{e}_r \equiv [\![\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{E}]\!] \tag{5}$$

where $\boldsymbol{a}_r \in \mathbb{R}^{I_1}$, $\boldsymbol{b}_r \in \mathbb{R}^{I_2}$, $\boldsymbol{c}_r \in \mathbb{R}^{I_3}$, $\boldsymbol{d}_r \in \mathbb{R}^{I_4}$, $\boldsymbol{e}_r \in \mathbb{R}^{I_5}$, for $r = 1, \ldots, R$. Factor matrix $\mathbf{A} = [\, \boldsymbol{a}_1, \boldsymbol{a}_2 \ldots\ldots\ldots \boldsymbol{a}_R] \in \mathbb{R}^{I_1 \times R}$, $\mathbf{B} = [\, \boldsymbol{b}_1, \boldsymbol{b}_2 \ldots\ldots\ldots \boldsymbol{b}_R] \in \mathbb{R}^{I_2 \times R}$, $\mathbf{C} = [\, \boldsymbol{c}_1, \boldsymbol{c}_2 \ldots\ldots\ldots \boldsymbol{c}_R] \in \mathbb{R}^{I_3 \times R}$, $\mathbf{D} = [\, \boldsymbol{d}_1, \boldsymbol{d}_2 \ldots\ldots\ldots\ldots \boldsymbol{d}_R] \in \mathbb{R}^{I_4 \times R}$, and $\mathbf{E} = [\, \boldsymbol{e}_1, \boldsymbol{e}_2 \ldots \boldsymbol{e}_R] \in \mathbb{R}^{I_5 \times R}$ are collections of all R vectors of $\boldsymbol{a}_r$, $\boldsymbol{b}_r$, $\boldsymbol{c}_r$, $\boldsymbol{d}_r$, $\boldsymbol{e}_r$ respectively.

Fig. 6 illustrates the process of the CPD tensor decomposition.

Eq. (5) can be represented as:

$$\boldsymbol{T} \approx \sum_{R} \boldsymbol{A}_r \circ \boldsymbol{B}_r \circ \boldsymbol{C}_r \circ \boldsymbol{D}_r \circ \boldsymbol{E}_r \tag{6}$$

Factor matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$ and $\boldsymbol{E}$ are computed by solving the following optimization equation:

$$\arg\min \| \boldsymbol{T} - \sum_{R} \boldsymbol{A}_r \circ \boldsymbol{B}_r \circ \boldsymbol{C}_r \circ \boldsymbol{D}_r \circ \boldsymbol{E}_r \| \tag{7}$$

To solve Eq. (7), the alternating least squares algorithm was applied using an optimization-based program *cpd_als* in Tensor Toolbox from MATLAB [47].

5.4.1. Selection of R (number of rank-1 tensors) for CPD

Before applying CPD, a proper R needs to be selected to guarantee a unique mapping of the original tensor to the decomposed tensor. To determine the smallest number of rank-1 tensors required for CPD, the core consistency diagnostic method (CORCONDIA) described in [48] was applied with the N-way toolbox of MATLAB [49]. In this method, a Tucker core is used for assessing the appropriateness of a CPD model. Tucker core array, when presenting the coordinates of the parts of the multi-dimensional array within the subspaces, is considered as the regression of a multi-dimensional array onto the subspaces characterized by the factor matrices [48]. In the CONCORDIA method, to signify an optimal representation of the multi-dimensional array with respect to the subspaces defined by the factors, a parameter called "core consistency" is measured. A core consistency value close to 100% indicates an appropriate CPD model that can properly represent the variability of the factors. With the increase of the number of components (R), the core consistency value typically increases up to a certain point and then decreases. To apply CPD in this study, one component model ($R = 1$) was first attempted and then different values of R were tried by gradually increasing the number of components until the core consistency value reached close to 100%. Those values of R with the highest core consistencies were considered for CPD application.

5.5. Tensor reconstruction (fused dataset for risk assessment)

Once the tensor was decomposed into factor matrices, these factor matrices were used to reconstruct a new tensor from its decomposed state. This reconstructed tensor is an approximate mapping of the sparse tensor $\boldsymbol{T}$, which was referred to as 'fused tensor' and represented as $\boldsymbol{T}' \in \mathbb{R}^{I_1 \times I_2 \times I_3 \times I_4 \times I_5}$, calculated by:

$$\boldsymbol{T}' = \sum_{r=1}^{R} \boldsymbol{A}_r \circ \boldsymbol{B}_r \circ \boldsymbol{C}_r \circ \boldsymbol{D}_r \circ \boldsymbol{E}_r \tag{8}$$

In this tensor, all the missing values were imputed, and the risk indicator values were reconstructed. Tensorlab (a MATLAB package for tensor computation) was used for the imputation and reconstruction [50]. Note that the factor matrices computed by Eq. (7) ensured the minimal reconstruction error during the imputation and reconstruction of the risk indicators. The reconstructed risk indicators were then extracted to a spread sheet, readily for use in WMSD risk assessment.

To allow for reliable observations, the procedures in Subsections 5.2 to 5.5 (except for construction of the five-dimensional tensor $\boldsymbol{X}$ from the original datasets) were repeated three times. In each repeat, the random removal of certain portion of data as designed earlier led to different preservation in each dataset. Three repeats ensured a two-thirds (67%) probability that the averaged results were more accurate than a single experiment while maintaining time-efficiency in implementation.

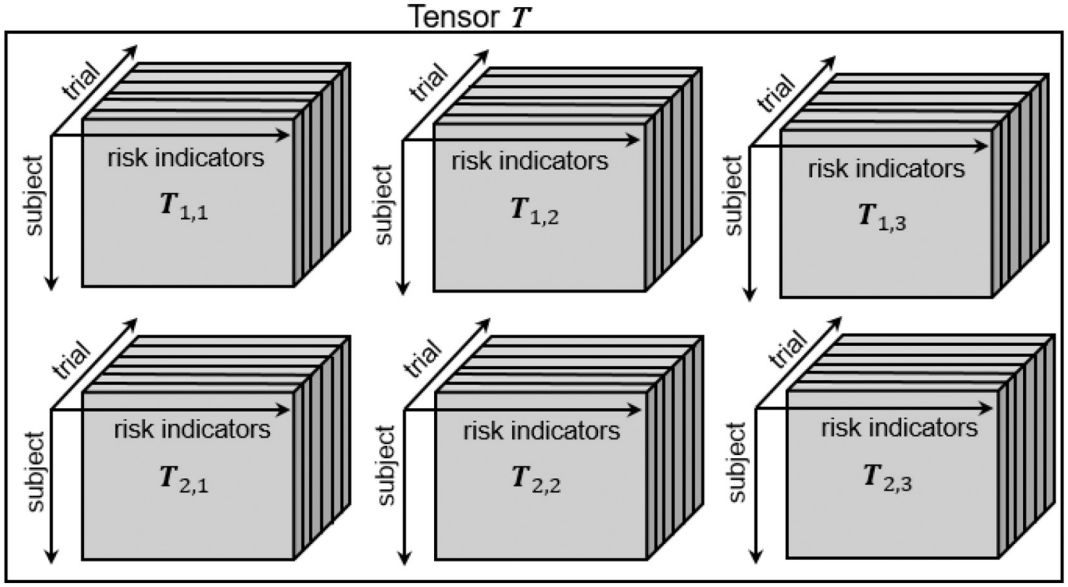

Fig. 5. Representation of the five-dimensional tensor $\mathcal{T}$.

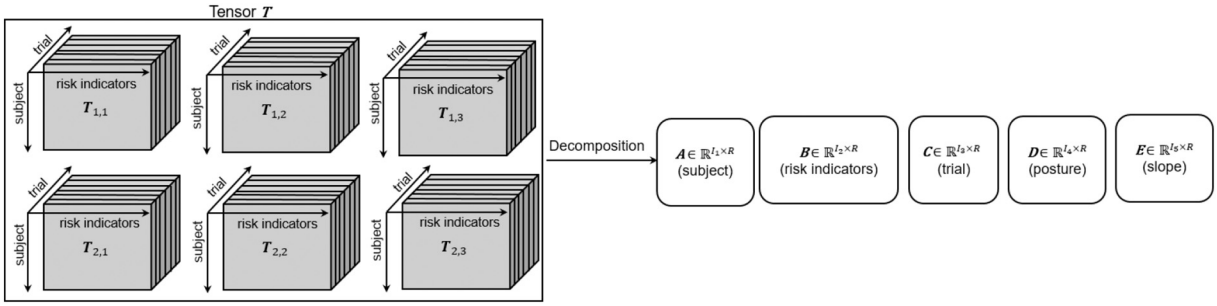

Fig. 6. CPD tensor decomposition of tensor $\mathcal{T}$.

5.6. Performance analysis of the data fusion

The performance of the data fusion method was evaluated by comparing the reconstructed data to the original data using two statistical measures—root mean square error (RMSE) and mean absolute error (MAE). RMSE and MAE were used to measure the standard error and average magnitude of the error of the fused tensor, respectively [51,52].

The *RMSE* was calculated using the following equation:

$$RMSE = \sqrt{\frac{1}{N}\left[\sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2}\sum_{i_3=1}^{I_3}\sum_{i_4=1}^{I_4}\sum_{i_5=1}^{I_5}(X_{i_1,i_2,i_3,i_4,i_5} - T'_{i_1,i_2,i_3,i_4,i_5})^2\right]} \quad (9)$$

Here, $X_{i_1,\ i_2,\ i_3,\ i_4,\ i_5}$ is an element (a value of a risk indicator of a certain subject at a certain slope, posture and trial) in the original tensor X and $T'_{i_1,\ i_2,\ i_3,\ i_4,\ i_5}$is an element in the re-constructed tensor (fused dataset). N is the total number of cells of the tensor which was computed as $N = I_1 \times I_2 \times I_3 \times I_4 \times I_5$.

The *MAE* was calculated using the following equation:

$$MAE = \frac{1}{N}\left[\sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2}\sum_{i_3=1}^{I_3}\sum_{i_4=1}^{I_4}\sum_{i_5=1}^{I_5}|(X_{i_1,i_2,i_3,i_4,i_5} - T'_{i_1,i_2,i_3,i_4,i_5})|\right] \quad (10)$$

The performance of the data fusion method was also evaluated by comparing the risk assessment results obtained from tensor T' to those obtained from tensor X collected in previous experimental studies. In this comparison, consistency of the effects of roof slope and working posture on five knee rotation angles and five knee postural muscles' EMG between the original and fused datasets were considered as a performance metric. In previous studies, the authors explored if the work-related factors (i.e., roof slope and working posture) are associated with knee WMSDs among construction roofers [42,43]. Those studies considered the original tensor (X) for risk assessments, whereas in this study, the fused tensor (T') was considered for assessing the similar risk to see if the risk assessment results were consistent. This way, the performance of the fusion method could be evaluated.

The consistency (Con) was computed using the following equation:

$$\text{Con} = \frac{\sum_{i=1}^{P} f(m_i, m'_i)}{P} \times 100. \quad (11)$$

where, P = Total number of possible effects of roof slope, working posture and their interactions on five knee rotation angles and EMG of five knee postural muscles [total 30 = 3 factors (slope, posture, slope-posture interaction) × 10 response variables (5 knee rotations and 5 EMG measurements)].

m_i denotes the effects of the factors (slope and posture) on the response variables obtained from the original tensor (X).

m'_i denotes the effects of the factors (slope and posture) on the response variables obtained from the fused tensor (T').

$f(m_i, m'_i) = \begin{pmatrix} 1, if\ m_i = m'_i \\ 0, if\ m_i \neq m'_i \end{pmatrix}$.

All the resulting measurements of the performance metrics were averaged over the three trial repetitions.

5.7. Results

Table 5 summarizes the results, including the RMSE, MAE, consistency (Con), number of rank-1 tensors R (latent factor) and core

Table 5
Summary of results.

Missing proportion (%)		RMSE	MAE	Con (%)	*R*	Core consistency (%)
Knee rotation	EMG					
10	10	1.39	0.69	80	17	98
10	20	1.38	0.8192	80	14	96.5
10	30	1.39	0.818	80	15	97
10	40	1.4	0.8412	80	14	96
10	50	1.44	0.805	77	14	96.5
10	60	1.57	0.9597	77	12	95
10	70	2.34	1.47	74	10	94.67
20	10	1.67	0.9791	80	12	97
20	20	1.73	1.04	80	11	97
20	30	1.78	1.04	80	11	95.6
20	40	1.75	1.06	77	12	94.8
20	50	1.79	1.08	74	13	94
20	60	1.83	1.09	74	13	93
20	70	2.45	1.25	70	10	92.7
30	10	1.82	1	77	13	97
30	20	2	1.14	74	11	95.6
30	30	2.14	1.17	74	11	96
30	40	2.19	1.25	74	10	95
30	50	2.26	1.29	74	10	95.7
30	60	2.32	1.24	74	10	93.5
30	70	5.6	3.4	57	11	94
40	10	2.11	1.11	87	12	96.5
40	20	2.95	1.32	74	10	94
40	30	2.98	1.31	74	10	95
40	40	2.52	1.24	74	12	93.7
40	50	2.83	1.3	74	12	92.76
40	60	2.7	1.39	74	11	93
40	70	2.77	1.51	70	13	92
50	10	4.9	1.69	80	22	93
50	20	5.11	1.81	80	31	93
50	30	5.28	1.99	80	29	92.6
50	40	6.19	2.24	77	34	92
50	50	6	2.4	77	43	91.5
50	60	6.5	2.7	77	40	92
50	70	6.8	3.8	60	10	91
60	10	3.38	1.45	84	12	93
60	20	4.4	1.76	77	15	93.6
60	30	4.63	1.8	77	13	92
60	40	5	2	77	13	92.8
60	50	5.32	2.08	74	10	93
60	60	5.39	2.16	70	16	92
60	70	5.6	2.2	70	11	91
70	10	6.7	2.7	77	20	90
70	20	7.7	3.2	74	17	89
70	30	7.5	3	74	12	91.5
70	40	7.54	3.38	70	10	90
70	50	8	3.3	70	11	92.4
70	60	8.29	3.4	70	11	90
70	70	8.8	3.67	67	12	88

consistency values computed for the knee rotation and EMG datasets with different proportions of missing values.

5.7.1. RMSE curves for datasets with different proportions of missing values

Fig. 7 displays the RMSE curves for different proportions of missing data from the knee rotation and EMG datasets. In Fig. 7, rot-DS and EMG-DS means the rotation dataset and EMD dataset respectively.

Generally, the RMSE curve demonstrated an increasing trend with the increase of missing values in both rot-DS and EMG-DS. It was graphically observed that, for up to 40% missing data from rot-DS and up to 60% missing data from EMG-DS, the RMSE values displayed a small variation from 1 to 3. When the proportion of the missing data from rot-DS increased to 50%, the RMSE value increased drastically for different proportions of missing values in EMG-DS. The RMSE curves for 50%, 60% and 70% missing data from the rot-DS shifted slightly upward, indicating that with the increase of missing values in rot-DS, the RMSE values generally increased. From 50% to up to 70% missing values from rot-DS and up to 70% missing values from EMG-DS, the RMSE values

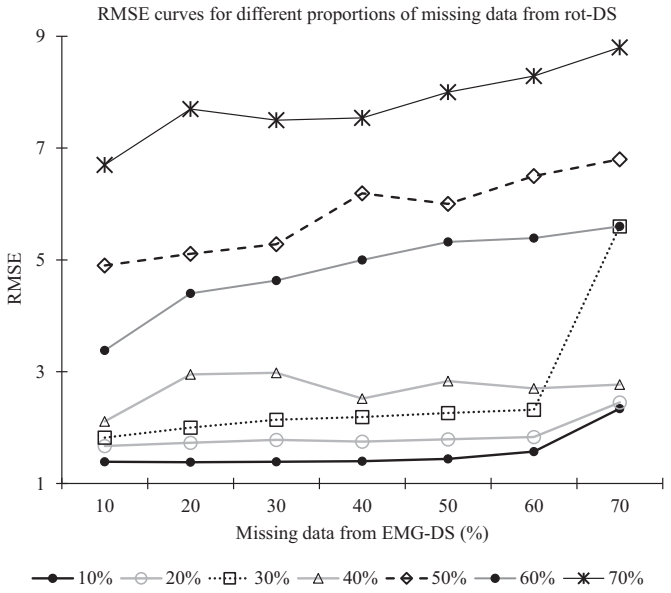

Fig. 7. RMSE curves for 10%, 20%, 30%, 40%, 50%, 60% and 70% missing data from rotation dataset.

varied considerably from 3.5 to 9. These results suggested that with the increase in missing values in both datasets, the performance of the fusion method decreased. The variation in RMSE values was small for up to 40% missing values in rot-DS and up to 60% missing values in EMG-DS. However, with further increase of missing values in both datasets, RMSE values varied considerably.

5.7.2. MAE curves for datasets with different proportions of missing values

Fig. 8 shows the MAE curves for different proportions of missing data from the knee rotation and EMG datasets. In Fig. 8, rot-DS and EMG-DS means the rotation dataset and EMD dataset respectively.

The MAE curves demonstrated similar patterns to the RMSE curves. In general, the MAE curves demonstrated an increasing trend with the increase of missing values in both datasets. It was observed that, for up to 40% missing values from rot-DS and up to 60% missing values from EMG-DS, the variation of the MAE values was between 0.6 and 1.5.

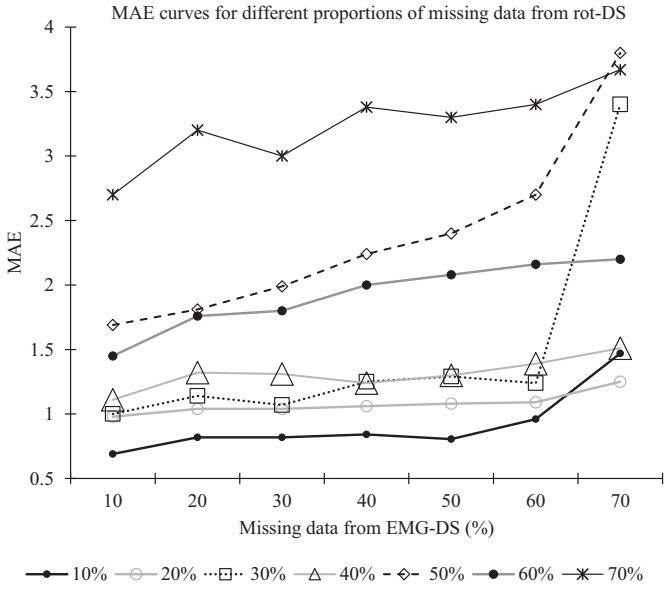

Fig. 8. MAE curves for 10%, 20%, 30%, 40%, 50%, 60% and 70% missing data from rotation dataset.

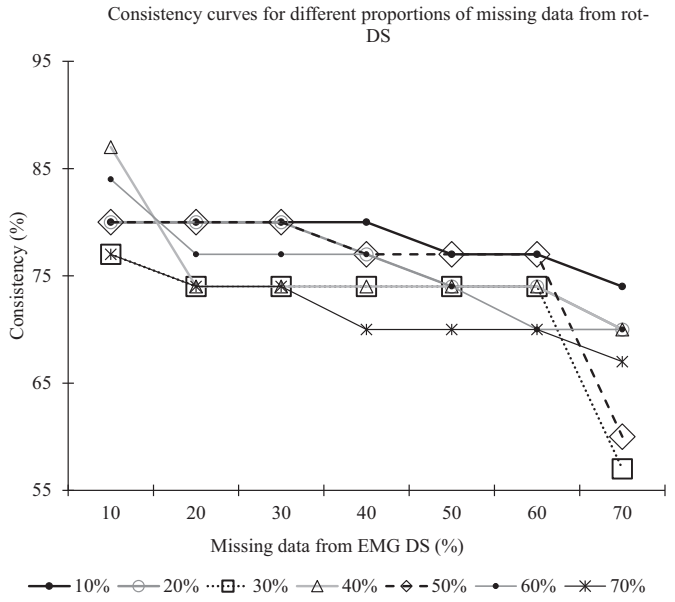

Fig. 9. Consistency curves for 10%, 20%, 30%, 40%, 50%, 60% and 70% missing data from rotation DS.

When missing values in EMG-DS were further increased (up to 70%), a sharp increase in MAE curve was observed when 30% data were missing in rot-DS. When the proportion of missing values in rot-DS was further increased to 50%, MAE values also tended to increase. Similar to the RMSE curves, for 50%, 60% and 70% missing data in rot-DS, the MAE curves shifted slightly upward, indicating that with the increase of missing values in rot-DS, MAE values also increased. In general, from 50%–70% missing values in rotation dataset and up to 70% missing values in EMG dataset, the values of MAE varied considerably from 1.5–4. These results indicated that with the increase in missing values in both datasets, the performance of the fusion method tended to decease as MAE values increased. The variation in MAE values was small for up to 40% missing values in rot-DS and up to 60% missing values in EMG-DS. However, with further increase of missing values in both datasets, the MAE values varied considerably.

5.7.3. Consistency curves for datasets with different proportions of missing values

Fig. 9 shows the consistency curves for different proportions of missing data from the knee rotation and EMG datasets. In Fig. 9, rot-DS and EMG-DS means the rotation dataset and EMD dataset respectively.

The consistency curves generally demonstrated a decreasing trend with the increase of missing values in both datasets. It was observed that for up to 70% missing values from rot-DS and up to 60% missing values from EMG-DS, the consistency values varied to a small extent from 87% to 70%. However, for 70% missing values in EMG-DS, the consistency value decreased considerably for several configurations, particularly for 30%, 50% and 70% missing values in rot-DS. These results indicated that with the increase in missing values in both datasets, the performance of the fusion method decreased in general.

6. Discussion and study limitations

To ensure reliable risk assessment of WMSD, proper data collection is crucial. However, human-based data can potentially have missing data points due to dropout from the data collection technology. Moreover, risks sometimes cannot be fully quantified with a single risk indicator and thus multiple heterogeneous risk indicators are often collected for risk assessment. As a result, a viable method is needed that reserves the interrelation among multiple risk indicators and potential risk factors during rectification of these multiple imperfect and incomplete datasets.

This research proposed the use of a CPD tensor decomposition-based data fusion method, which fuses multiple imperfect datasets in order to replace missing data. The proposed method generates a fused dataset by learning the latent structures and collaborative relationships among the potential risk factors and the risk indicators captured in multiple work settings. By learning their correlations, this proposed method replaces the missing values in the multiple incomplete datasets with minimum artifacts. The sample analysis results demonstrated that for up to a significant portion of missing values from both datasets, the proposed method can generate a fused dataset, by which reliable risk assessment results can be obtained consistent to those obtained from the original datasets to a great extent (87% to 70%). The possible reason for some inconsistencies in the risk assessment results may be attributed to the presence of significant amount of missing values in the individual datasets. In spite of this, for the specific datasets used in this study, the proposed method performed consistently up to 40% of missing data in the original datasets and could be reasonably applied to those datasets. The RMSE and MAE values demonstrated that the reconstruction errors in fusion were very small up to this range of missing data. This suggests that the reconstructed fused values of the risk indicators in the new dataset are close to those in the original, unaltered experimental datasets and hence, a good fusion performance. Although there is no standard threshold for good or bad RMSE and MAE values, low values of RMSE and MAE (within 10% of the range of the reference values) are preferable. RMSE should be less than 10% of the range of target property value [53]. In the current study, for different portions of missing values, the obtained RMSE values ranged from 0.74% to 5% relative to the difference between the minimum and maximum values of the risk indicators in the original dataset. Additionally, for up to 40% missing values from the knee rotation dataset and 60% missing values from the EMG dataset, the RMSE values were within 15% of the average value of the risk indicators in the original dataset. Similarly, the MAE values were within 10% of the average value of the risk indicators in the original dataset. However, for the specific datasets used in this study, the RMSE and MAE values varied considerably as the proportion of missing data increased to 50% and 70% in the rotation and EMG datasets respectively. The possible reason might be the presence of significant amount of missing values in the individual datasets that drastically deteriorated the performance of the fusion method in reconstruction of the data and therefore negatively affected the RMSE and MAE values.

Apart from the previous three performance metrics, the performance of the present method can be further validated by the ‘core consistency’ values obtained when measuring number of rank-1 tensors R (latent factor) for CPD. For different R values presented in Table 5, the ‘core consistency’ values ranged from 88%–98%, indicating that the CPD tensor decomposition is able to represent the variability of the multi-dimensional EMG and rotation data in the subspace substantially, thus providing a good fusion result. Overall, the findings suggest that the proposed method can handle up to 40% missing data and can reasonably be applied in WMSD risk assessment studies. The performance of this method was evaluated based on: a) how close a reconstructed fused dataset was to the original dataset, and b) how consistent the risk assessment results from a fused dataset were to those from the original dataset.

Although the proposed method has been proven successful in generating a fused dataset for used in risk assessment studies, there are still some limitations that are worth further investigation. First, in this study, only the maximum knee rotation angles and maximum normalized muscle activity measurements were considered for fusion. It would be interesting to explore a detailed time series dataset instead of an aggregated metric such as maximum. Although data fusion is known to provide deeper and more accurate insights than those obtained from individual single data source [54], another limitation of this work is the

performance of the data fusion was only assessed by comparing the risk assessment results obtained from the fused datasets to the original dataset. The fused datasets were proven to be statistically similar to the original dataset, indicating that the fusion method was successful in reconstruction of a dataset when a significant amount of data are missing. Yet, how a fused dataset outperforms the individual incomplete datasets by which the fused dataset is constructed in terms of performance of risk assessment has not been measured.

7. Conclusion and future extension

The current research proposed a method that applies the CPD tensor decomposition technique to fuse multiple imperfect and incomplete datasets, as well as replacing missing data for assessing WMSD risks among construction workers. The proposed method helps not only in replacing missing values, but also holds the correlation among the potential risk factors and the risk indicators during replacement. The method was validated by comparing the risk assessment results obtained from the fused datasets to those from the original experimental dataset. RMSE and MAE values were also computed to compare the performance of the fusion method with the increase of missing values. The validation results suggest that the reconstructed fused datasets were statistically similar to the original experimental datasets and therefore able to provide consistent risk assessment results up to a significant amount of missing values. This method can be used as an alternative risk assessment method when data collected from experimental studies or real construction sites are incomplete due to missing data points. The findings will help enable accurate assessment of work-related risk factors for WMSDs among construction workers. Although the proposed method only used two sparse datasets for fusion, it can be extended to fuse more datasets. In future, this method will be explored in fusing imperfect and incomplete multiple time series datasets. The performance of the fusion will also be further tested by comparing consistencies of the risk assessment results to those obtained from datasets with missing values.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors acknowledge the support of the National Institute for Occupational Safety and Health (NIOSH), who funded this research. The findings and conclusions in this research are those of the authors and do not necessarily represent the opinion of the National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention.

References

[1] BLS, Nonfatal occupational injuries and illnesses: cases with days away from work, < https://www.bls.gov/news.release/pdf/osh2.pdf> (4 March, 2019), 2019.
[2] G. David, Ergonomic methods for assessing exposure to risk factors for work-related musculoskeletal disorders, Occup. Med. 55 (3) (2005) 190–199, https://doi.org/10.1093/occmed/kqi082.
[3] M.C. Data, Secondary Analysis of Electronic Health Records, Springer International Publishing, 2016.
[4] W. Young, G. Weckman, W. Holland, A survey of methodologies for the treatment of missing values within datasets: limitations and benefits, Theor. Issues Ergon. Sci. 12 (1) (2011) 15–43, https://doi.org/10.1080/14639220903470205.
[5] S. Rabanser, O. Shchur, S. Günnemann, Introduction to tensor decompositions and their applications in machine learning, arXiv Preprint, 2017 https://arxiv.org/abs/1711.10781v1.
[6] J.Y. Lin, Y. Lu, X. Tu, How to avoid missing data and the problems they pose: design considerations, Shanghai Arch. Psychiatry 24 (3) (2012) 181–184, https://doi.org/10.3969/j.issn.1002-0829.2012.03.010.
[7] H. Kang, The prevention and handling of the missing data, Korean J. Anesthesiol. 64 (5) (2013) 402–406, https://doi.org/10.4097/kjae.2013.64.5.402.
[8] K.S. Button, J.P. Ioannidis, C. Mokrysz, B.A. Nosek, J. Flint, E.S. Robinson, M.R. Munafò, Power failure: why small sample size undermines the reliability of neuroscience, Nat. Rev. Neurosci. 14 (5) (2013) 365–376, https://doi.org/10.1038/nrn3475.
[9] N.K. Malhotra, Analyzing marketing research data with incomplete information on the dependent variable, J. Mark. Res. 24 (1) (1987) 74–84, https://doi.org/10.2307/3151755.
[10] R.M. Hamer, P.M. Simpson, Last observation carried forward versus mixed models in the analysis of psychiatric clinical trials, Am. J. Psychiatr. 166 (6) (2009) 639–641, https://doi.org/10.1176/appi.ajp.2009.09040458.
[11] A. Gelman, T.E. Raghunathan, Using conditional distributions for missing-data imputation, Stat. Sci. 15 (2001) 268–269 http://www.stat.columbia.edu/~gelman/research/published/arnold2.pdf.
[12] D. Lahat, T. Adali, C. Jutten, Multimodal data fusion: an overview of methods, challenges, and prospects, Proc. IEEE 103 (9) (2015) 1449–1477, https://doi.org/10.1109/JPROC.2015.2460697.
[13] M. Amiri, R. Jensen, Missing data imputation using fuzzy-rough methods, Neurocomputing 205 (2016) 152–164, https://doi.org/10.1016/j.neucom.2016.04.015.
[14] J. Luengo, J.A. Sáez, F. Herrera, Missing data imputation for fuzzy rule-based classification systems, Soft. Comput. 16 (5) (2012) 863–881, https://doi.org/10.1007/s00500-011-0774-4.
[15] R. Winkler, F. Klawonn, R. Kruse, Fuzzy c-means in high dimensional spaces, Int. J. Fuzzy Syst. Appl. 1 (1) (2011) 1–16, https://doi.org/10.4018/IJFSA.2011010101.
[16] F. Castanedo, A review of data fusion techniques, Sci. World J. 2013 (2013), https://doi.org/10.1155/2013/704504.
[17] S.N. Razavi, C.T. Haas, Multisensor data fusion for on-site materials tracking in construction, Autom. Constr. 19 (8) (2010) 1037–1046, https://doi.org/10.1016/j.autcon.2010.07.017.
[18] S.N. Razavi, C.T. Haas, Reliability-based hybrid data fusion method for adaptive location estimation in construction, J. Comput. Civ. Eng. 26 (1) (2011) 1–10, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000101.
[19] A. Shahi, J.M. Cardona, C.T. Haas, J.S. West, G.L. Caldwell, Activity-based data fusion for automated progress tracking of construction projects, Construction Research Congress 2012: Construction Challenges in a Flat World, 2012, pp. 838–847, , https://doi.org/10.1061/9780784412329.085.
[20] R. Soman, M. Kyriakides, T. Onoufriou, W. Ostachowicz, Numerical evaluation of multi-metric data fusion based structural health monitoring of long span bridge structures, Struct. Infrastruct. Eng. 14 (6) (2018) 673–684, https://doi.org/10.1080/15732479.2017.1350984.
[21] A. Anaissi, M. Makki Alamdari, T. Rakotoarivelo, N. Khoa, A tensor-based structural damage identification and severity assessment, Sensors 18 (1) (2018) 111, https://doi.org/10.3390/s18010111.
[22] J. Chen, J. Qiu, C. Ahn, Construction worker's awkward posture recognition through supervised motion tensor decomposition, Autom. Constr. 77 (2017) 67–81, https://doi.org/10.1016/j.autcon.2017.01.020.
[23] T. Cheng, G.C. Migliaccio, J. Teizer, U.C. Gatti, Data fusion of real-time location sensing and physiological status monitoring for ergonomics analysis of construction workers, J. Comput. Civ. Eng. 27 (3) (2012) 320–335, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000222.
[24] T. Cheng, J. Teizer, G.C. Migliaccio, U.C. Gatti, Automated task-level activity analysis through fusion of real time location sensors and worker's thoracic posture data, Autom. Constr. 29 (2013) 24–39, https://doi.org/10.1016/j.autcon.2012.08.003.
[25] H. Chen, X. Luo, Z. Zheng, J. Ke, A proactive workers' safety risk evaluation framework based on position and posture data fusion, Autom. Constr. 98 (2019) 275–288, https://doi.org/10.1016/j.autcon.2018.11.026.
[26] Y. Zheng, Methodologies for cross-domain data fusion: an overview, IEEE Trans. Big Data 1 (1) (2015) 16–34, https://doi.org/10.1109/TBDATA.2015.2465959.
[27] Y. Zheng, Y. Liu, J. Yuan, X. Xie, Urban computing with taxicabs, Proceedings of the 13th International Conference on Ubiquitous Computing, 2011, pp. 89–98, , https://doi.org/10.1145/2030112.2030126.
[28] W. Wen, C. Wu, Y. Wang, Y. Chen, H. Li, Learning structured sparsity in deep neural networks, Advances in Neural Information Processing Systems, 2016, pp. 2074–2082 https://papers.nips.cc/paper/6504-learning-structured-sparsity-in-deep-neural-networks.pdf.
[29] L. Sorber, M. Van Barel, L. De Lathauwer, Structured data fusion, IEEE J. Sel. Top. Signal Process. 9 (4) (2015) 586–600, https://doi.org/10.1109/JSTSP.2015.2400415.
[30] E.E. Papalexakis, C. Faloutsos, N.D. Sidiropoulos, Tensors for data mining and data fusion: models, applications, and scalable algorithms, ACM Trans. Intel. Syst. Technol. 8 (2) (2017) 16, https://doi.org/10.1145/2915921.
[31] J. Dauwels, L. Garg, A. Earnest, L.K. Pang, Tensor factorization for missing data imputation in medical questionnaires, 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 2109–2112, , https://doi.org/10.1109/ICASSP.2012.6288327.
[32] K. Xie, X. Li, X. Wang, G. Xie, J. Wen, J. Cao, D. Zhang, Fast tensor factorization for accurate internet anomaly detection, IEEE/ACM Trans. Network. (TON) 25 (6) (2017) 3794–3807, https://doi.org/10.1109/TNET.2017.2761704.
[33] J. Charlier, R. State, J. Hilger, Non-negative paratuck2 tensor decomposition combined to lstm network for smart contracts profiling, 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), 2018, pp. 74–81, , https://doi.org/10.1109/BigComp.2018.00020.
[34] L. Xiong, X. Chen, T.-K. Huang, J. Schneider, J.G. Carbonell, Temporal collaborative

filtering with bayesian probabilistic tensor factorization, Proceedings of the 2010 SIAM International Conference on Data Mining, 2010, pp. 211–222, , https://doi.org/10.1137/1.9781611972801.19.
[35] D.M. Dunlavy, T.G. Kolda, E. Acar, et al., ACM Trans. Knowl. Discov. Data (TKDD) 5 (2) (2011) 10, https://doi.org/10.1145/1921632.1921636.
[36] M. Žitnik, B. Zupan, Data fusion by matrix factorization, IEEE Trans. Pattern Anal. Mach. Intell. 37 (1) (2014) 41–53, https://doi.org/10.1109/TPAMI.2014.2343973.
[37] T.G. Kolda, B.W. Bader, Tensor decompositions and applications, SIAM Rev. 51 (3) (2009) 455–500, https://doi.org/10.1137/07070111X.
[38] N. Vervliet, O. Debals, L. Sorber, L. De Lathauwer, Breaking the curse of dimensionality using decompositions of incomplete tensors: tensor-based scientific computing in big data analysis, IEEE Signal Process. Mag. 31 (5) (2014) 71–79, https://doi.org/10.1109/MSP.2014.2329429.
[39] E. Acar, D.M. Dunlavy, T.G. Kolda, M. Mørup, Scalable tensor factorizations for incomplete data, Chemom. Intell. Lab. Syst. 106 (1) (2011) 41–56, https://doi.org/10.1016/j.chemolab.2010.08.004.
[40] M.T. Asif, N. Mitrovic, J. Dauwels, P. Jaillet, Matrix and tensor based methods for missing data estimation in large traffic networks, IEEE Trans. Intell. Transp. Syst. 17 (7) (2016) 1816–1825, https://doi.org/10.1109/TITS.2015.2507259.
[41] T.G. Kolda, J. Sun, Scalable tensor decompositions for multi-aspect data mining, 2008 Eighth IEEE International Conference on Data Mining, IEEE, 2008, pp. 363–372, , https://doi.org/10.1109/ICDM.2008.89.
[42] S.P. Breloff, A. Dutta, F. Dai, E.W. Sinsel, C.M. Warren, X. Ning, J.Z. Wu, Assessing work-related risk factors for musculoskeletal knee disorders in construction roofing tasks, Appl. Ergon. 81 (2019) 102901, https://doi.org/10.1016/j.apergo.2019.102901.
[43] A. Dutta, S.P. Breloff, F. Dai, E.W. Sinsel, C.M. Warren, R.E. Carey, J.Z. Wu, Effects of Working Posture and Roof Slope on Activation of Lower Limb Muscles during Shingle Installation, Ergonomics, 1-12 (2020), pp. 1–7, https://doi.org/10.1080/00140139.2020.1772378 https://www.tandfonline.com/doi/abs/10.1080/00140139.2020.1772378?journalCode=terg20.
[44] D.C. Kingston, L.M. Tennant, H.C. Chong, S.M. Acker, Peak activation of lower limb musculature during high flexion kneeling and transitional movements, Ergonomics 59 (9) (2016) 1215–1223, https://doi.org/10.1080/00140139.2015.1130861.
[45] J.K. Hofer, R. Gejo, M.H. McGarry, T.Q. Lee, Effects on tibiofemoral biomechanics from kneeling, Clin. Biomech. 26 (6) (2011) 605–611, https://doi.org/10.1016/j.clinbiomech.2011.01.016.
[46] P. Pavlidis, J. Weston, J. Cai, W.S. Noble, Learning gene functional classifications from multiple data types, J. Comput. Biol. 9 (2) (2002) 401–411, https://doi.org/10.1089/10665270252935539.
[47] B. Bader, T. Kolda, MATLAB tensor toolbox version 2.4, < https://www.sandia.gov/~tgkolda/TensorToolbox/index-2.6.html > (May 14, 2019).
[48] R. Bro, H.A. Kiers, A new efficient method for determining the number of components in PARAFAC models, J. Chemom. 17 (5) (2003) 274–286, https://doi.org/10.1002/cem.801.
[49] C.A. Andersson, R. Bro, The N-way toolbox for MATLAB, Chemom. Intell. Lab. Syst. 52 (1) (2000) 1–4, https://doi.org/10.1016/S0169-7439(00)00071-X.
[50] O. Debals, F. Van Eeghem, N. Vervliet, L. De Lathauwer, Tensor computations using Tensorlab, https://www.tensorlab.net/demos/tutorial.pdf>, (May 19, 2019).
[51] I. Alimuddin, J.T.S. Sumantyo, H. Kuze, Assessment of pan-sharpening methods applied to image fusion of remotely sensed multi-band data, Int. J. Appl. Earth Obs. Geoinf. 18 (2012) 165–175, https://doi.org/10.1016/j.jag.2012.01.013.
[52] K.B. Ng, P.B. Kantor, Predicting the effectiveness of naive data fusion on the basis of system characteristics, J. Am. Soc. Inf. Sci. 51 (13) (2000) 1177–1189, https://doi.org/10.1002/1097-4571(2000)9999:9999%3C::AID-ASI1030%3E3.0.CO;2-E.
[53] D.L. Alexander, A. Tropsha, D.A. Winkler, Beware of R 2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models, J. Chem. Inf. Model. 55 (7) (2015) 1316–1322, https://doi.org/10.1021/acs.jcim.5b00206.
[54] B. Khaleghi, A. Khamis, F.O. Karray, S.N. Razavi, Multisensor data fusion: a review of the state-of-the-art, Inform. Fusion 14 (1) (2013) 28–44, https://doi.org/10.1016/j.inffus.2011.08.001.