TENSOR FACTORIZATION FOR MISSING DATA IMPUTATION IN MEDICAL QUESTIONNAIRES

Justin Dauwels[1], Lalit Garg[1], Arul Earnest[2], Leong Khai Pang[3]

[1]School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[2]Centre for Quantitative Medicine, Duke-NUS Graduate Medical School, Singapore
[3]Dept Rheumatology, Allergy and Immunology, Tan Tock Seng Hospital, Singapore

ABSTRACT

This paper presents innovative collaborative filtering techniques to complete missing data in repeated medical questionnaires. The proposed techniques are based on the canonical polyadic (CP) decomposition (a.k.a. PARAFAC). Besides the standard CP decomposition, also a normalized decomposition is utilized. As an illustration, systemic lupus erythematosus-specific quality-of-life questionnaire is considered. Measures such as normalized root mean square error, bias and variance are used to assess the performance of the proposed tensor-based methods in comparison with other widely used approaches, such as mean substitution, regression imputations and *k*-nearest neighbor estimation. The numerical results demonstrate that the proposed methods provide significant improvement in comparison to popular methods. The best results are obtained for the normalized decomposition.

Index Terms— Medical information systems, Health information management, Public healthcare, Data handling

1. INTRODUCTION

A common problem with questionnaires is missing data [1]. Some level of missing data in repeated questionnaires is frequent and cannot be avoided completely, despite enormous care and effort to prevent it [1]. For instance, patients may elect to leave one or more items unanswered either inadvertently or because they may not wish to respond to questions dealing with a sensitive topic [2]. Missing data may lead to biased parameter estimates and inflated errors [1]. Missing data imputation has been a classical research topic over the last decades. Numerous methods have been proposed including list-wise or case-wise deletion, pair-wise deletion, mean substitution, regression imputations, and weighted K-nearest neighbor (KNN). However, still there is need for better missing data estimation methods, which can predict incomplete data more accurately [3]. In this paper, we use a tensor factorization [4-6] method to complete missing data in repeated questionnaires. Our methods are based on tensor decomposition, i.e., canonical polyadic (CP) decomposition (a.k.a. PARAFAC); a repeated medical questionnaire can naturally be arranged as a three-dimensional questionnaire, where the three dimensions are the questions, respondents, and follow-ups respectively. The proposed CP based methods predict missing responses in repeated questionnaires by effectively learning inherent collaborative relationship structure (from known responses) at different levels (among questions, respondents, and follow-ups). We illustrate our approach by a quality of life questionnaire filled by one hundred systemic lupus erythematosus (SLE) patients from hospitals in Singapore, China, and Vietnam. We use measures such as normalized root mean square error (NRMSE), bias, and variance to assess the performance of our methods and other widely used approaches, such as mean substitution, regression imputations, and *k*-nearest neighbor (KNN) estimation. Our results indicate that the proposed methods provide significant improvement compared to popular methods.

This paper is organized as follows. In the next section, we briefly review some standard imputation methods. In Section 3, we outline our proposed methods, and in Section 4, we describe our data set. In Section 5, we describe our results, and we provide concluding remarks in Section 6.

2. CLASSICAL METHODS FOR MISSING VALUE IMPUTATION

A naïve approach is to replace the missing response of a question by the mean of its known responses, which is referred to as mean substitution (MS). The method is easy to implement and use, however, it decreases the variance of the responses. Another popular family of methods is based on regression, which exploits correlation among known responses to impute missing response. We consider here an advanced regression imputation method called "iterative local least square method for missing value imputation" [7] as benchmark for our proposed methods. Another popular missing value imputation method is weighted *k*-nearest neighbors (KNN), which imputes a missing response of a question as the weighted sum of its known responses from respondents with similar responses to other questions. The performance of KNN is highly dependent on the choice of *k*.

With small k, results may suffer from outliers, while for large k, uncorrelated respondents start to play a role in the predictions [8]. (In our experiments, we implemented KNN method with k=10, as that choice provided the best results.)

3. TENSOR DECOMPOSITION FOR MISSING DATA IMPUTATION

3.1. CANDECOMP/PARAFAC (CP)

Tensors (a.k.a. hypermatrices or multi-way arrays) are multi-dimensional arrays [9]. An order N tensor $\boldsymbol{\chi}_N \in \mathbb{R}^{D_1 \times D_2 \times \cdots \times D_N}$ has $size(\boldsymbol{\chi}_N) = D_1 \times D_2 \times \cdots \times D_N$, where D_i is the size of its i^{th}.

CP is a technique of factorizing a tensor into a minimal sum of rank-one tensors. A rank-one tensor of order M is a tensor which can be written as the outer product of M vectors, i.e.

$$\mathbf{U}_M = \mathbf{V}_1 \circ \mathbf{V}_2 \circ \cdots \circ \mathbf{V}_M = \prod_n^N \left(\circ \mathbf{V}_{nr} \right), \tag{1}$$

where $\mathbf{U}_M \in \mathbb{R}^{I_1, I_2, \ldots, I_M}$ and $\mathbf{V}_m$ is one dimensional tensors or vectors such that $\mathbf{V}_m \in \mathbb{R}^{I_m}$. CP decomposition of tensor $\boldsymbol{\chi}_N$ into R rank-one tensors can be represented as follows:

$$\boldsymbol{\chi}_N = \mathbf{U}_1 + \mathbf{U}_2 + \cdots + \mathbf{U}_R = \sum_{r=1}^{R} \left(\mathbf{U}_r \right), \tag{2}$$

where $\mathbf{U}_r$ is the r^{th} factor of tensor $\boldsymbol{\chi}_N$. From (1) and (2) we can write as follows

$$\boldsymbol{\chi}_N = \sum_{r=1}^{R} \left(\mathbf{V}_{1r} \circ \mathbf{V}_{2r} \circ \cdots \circ \mathbf{V}_{Nr} \right) = \sum_{r=1}^{R} \left(\prod_n^N \left(\circ \mathbf{V}_{nr} \right) \right), \tag{3}$$

where vector $\mathbf{V}_{nr} \in \mathbb{R}^{I_n}$, $\mathbf{V}_i \circ \mathbf{V}_j$ denotes the outer product of tensor $\mathbf{V}_i$ to tensor $\mathbf{V}_j$.

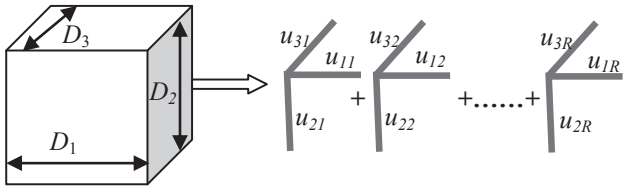

Figure 1. CP decomposition of a three dimensional tensor

Figure 1 is the schematic representation of the CP factorization of a three dimensional tensor. It has been shown that for any tensor there is unique CP factorization [9]. The number of rank-one tensor into which a tensor is factorized is equal to its rank. By omitting some of the rank-one tensors in the decomposition, one can obtain an approximate decomposition. CP factorization is a non-polynomial time complex problem [9]. Approximation methods are used to estimate CP factorization [4-6]. The goal of approximation methods is to minimize the reconstruction error (also called approximation or estimation error), which can be defined as:

$$\varepsilon_{\mathfrak{V}R} = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_N=1}^{I_N} \left(x_{i_1 i_2, \cdots i_N} - \sum_{r=1}^{R} \left(\prod_n^N \left(v_{i_n nr} \right) \right) \right)^2, \tag{4}$$

where $x_{i_1 i_2, \cdots i_N}$ is an element of tensor $\boldsymbol{\chi}_N$ at position $i_1, i_2, \cdots i_N$ and $v_{i_n nr}$ is the i_n^{th} element of vector $\mathbf{V}_{nr}$.

3.2. Missing data imputation using CP

In order to predict missing data, CP learns the latent structure and collaborative relationships among the different dimensions of the tensor (rows, columns, tubes). In medical questionnaire data, the dimensions are questions, respondents, and follow-ups respectively. CP is very effective in capturing dependencies in high-dimensional datasets. Therefore, it can effectively be used for missing data analysis. Acar et al. [10] proposed a tensor factorization model which can be used with incomplete data tensors, i.e. tensors having some of their values missing. An incomplete data tensor $\boldsymbol{\chi}_N$ is multiplied with a tensor $\mathfrak{V}_N$ of size equal to the size of tensor $\boldsymbol{\chi}_N$ of binary elements. Each element of $\mathfrak{V}_N$ defines if the corresponding element of the tensor $\boldsymbol{\chi}_N$ is missing or known. This is solved as an optimization problem minimizing the reconstruction error function defined as follows:

$$\varepsilon_{\mathfrak{V}R} = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_N=1}^{I_N} \left(\mathfrak{V}_{i_1 i_2, \cdots i_N} \left(x_{i_1 i_2, \cdots i_N} - \sum_{r=1}^{R} \left(\prod_n^N \left(v_{i_n nr} \right) \right) \right) \right)^2. \tag{5}$$

The optimization problem in (5) is formulated as a weighted least square problem, which may be solved using Nonlinear Conjugate Gradient based method [4-6].

3.3. CP with column normalized tensor

Since data in a tensor may be unbalanced, it is often recommended to normalize a tensor before decomposing it. In particular, the responses from different patients might vary substantially. We therefore normalize each column (corresponding to individual patients) by subtracting the mean. An element $x'_{i_1 i_2, \cdots i_N}$ of the column normalized tensor $\boldsymbol{\chi}'_N$ of a tensor $\boldsymbol{\chi}_N$ can be defined as follows:

$$x'_{i_1 i_2, \cdots i_N} = x_{i_1 i_2, \cdots i_N} - \bar{x}_{i_1 i_3, \cdots i_N}, \tag{6}$$

where $\bar{x}_{i_1 i_3, \cdots i_N}$ is the mean of column $i_1, i_3, \ldots, i_N$.

In order to impute missing values, we add the corresponding mean value back after performing CP.

3.4. Implementation and cross-validation

We applied the tensor decomposition algorithm of [4-6]. Specifically, we used the Tensor Toolbox [11] for Matlab,

which in turn relies on the Poblano Toolbox [12]. The dataset is stored as a three-dimensional sparse tensor. The three dimensions are respondents, questions and follow-ons respectively. Then we perform CP and reconstructed the factorized tensor. The algorithm imputes missing values minimizing the reconstruction error [4-6]. The column mean is then added back to corresponding elements.

To gauge the generalizability of our proposed method, we used ρ-repeated k-fold cross-validation [12], with 10 repetitions (ρ=10) each. It is repeated ρ=10 times for different proportions of missing values. For each repetition, the SEQOL dataset is randomly partitioned into k (=100/p) subsets based on proportion p of missing value. The size of each dataset is $\left(\frac{size(\chi_N)*p}{100}\right)$. Then k-1 subsets are used as training set. Each training set serves as a missing at random (MAR) dataset with p% missing values. The remaining subset is used as test dataset. We apply the missing data methods to the training set. The estimation error is then calculated as the difference between the imputed value and the original value on the test dataset. The above procedure is repeated k times such that each of the k datasets is used exactly once as the test dataset. For each of the ρ=10 repetitions of k-fold cross validation, the dataset is randomly divided into new k partitions.

We assess the performance of the missing data methods by means of three statistical measures: normalized root mean square error (RMSE), variance, and bias. The normalized root mean square error (RMSE) is calculated as follows:

$$\text{NRMSE} = \frac{1}{\left(x_{\min} - x_{\max}\right)} \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_N=1}^{I_N} \left(x_{i_1, i_2, \cdots i_N} - x'_{i_1, i_2, \cdots i_N} \right)^2 , \quad (7)$$

where $x_{i_1, i_2, \cdots i_N}$ is the actual value and $x'_{i_1, i_2, \cdots i_N}$ is the imputed value. The range of difficulty scores is from 1 to 8, therefore $\left(x_{\min} - x_{\max}\right)$ is 7. Similarly bias is calculated as

$$\text{bias} = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_N=1}^{I_N} \left(x_{i_1, i_2, \cdots i_N} - x'_{i_1, i_2, \cdots i_N} \right). \quad (8)$$

All three measures are also averaged over the r=10 repetitions.

The variance is computed as the mean of the variance for each dataset in the repeated k-fold cross-validation, after replacing the missing values by imputed values. NRMSE, bias, and variance are calculated for different proportions of missing values.

We calculate these measures for the standard methods (mean substitution, KNN, iterative local least square), and both our tensor decomposition methods (i.e., with and without normalization).

4. DATASET

To illustrate the proposed methods and to assess their performance, we consider a SLEQOL questionnaire ("systemic lupus erythematosus-specific quality-of-life instrument") [10]. It contains the responses of hundred systemic lupus erythematosus (SLE) patients from hospitals in Singapore, China and Vietnam. Patients provided difficulty scores (from 1 to 8) to each of 40 questions (total 4000 entries), which are used to determine the presence and burden of disease and treatment related symptoms. Responses were repeatedly recorded for three follow-ons. The dataset has only 40 missing values (0.33%), and therefore, it can serve as clean dataset for our experiments; it allows us to assess missing value methods, since the responses to most questions are known.

Interestingly, the data across the repeated questionnaires are highly correlated. The Pearson correlation coefficient for the baseline with the first (second) follow-up is 0.66 (0.57), and for the first with the second follow-up, it is 0.72.

5. RESULTS

Table 1 and Figure 2 summarize our results. As there is no straightforward algorithm is known to determine the rank of a tensor (i.e. smallest number of components in CP decomposition of the tensor [4]), we used cross-validation to estimate the rank by trying different number of components during training. It estimates rank as 47. The proposed standard-CP based method significantly improves the imputation accuracy in terms of NRMSE compared to existing methods, i.e., by more than 5%. For low proportion (10%) of missing values, it is even up to 8.7%. If we normalize the columns in the tensor before applying CP, the improvement is more than 11% in terms of NRMSE over existing methods; it is even up to 15% for 10% missing values. Imputing missing data should not strongly affect the variance of the data (responses in medical questionnaires). Obviously, mean substitution significantly reduces the variances, since it replaces all missing data by the mean value.

TABLE I. Comparison of classical methods with the proposed CP based methods. Table compares NRMSE, bias and variance of these methods. The variance of the original dataset is 2.05.

Proportion of missing data	**10%**				**20%**				**30%**			
Method	**NRMSE**	**bias**	**variance**	**Standard deviation**	**NRMSE**	**bias**	**variance**	**Standard deviation**	**NRMSE**	**bias**	**variance**	**Standard deviation**
Mean Substitution	0.19	**-0.0006**	1.91	1.38	0.19	**-0.0003**	1.78	1.33	0.19	**-0.0004**	1.64	1.28
weighted K-nearest neighbors Imputation	0.18	**0.0005**	1.94	1.39	0.18	**0.002**	1.84	1.36	0.18	**0.004**	1.73	1.32
Iterative local least square	0.16	-0.056	2.01	**1.42**	0.17	-0.07	**1.96**	**1.40**	0.17	-0.048	**1.89**	**1.37**
CP based method	**0.15**	0.140	1.97	1.40	**0.16**	0.18	1.89	1.37	**0.16**	0.18	1.79	1.34
Column normalized CP based method	**0.14**	-0.024	**2.00**	**1.41**	**0.15**	-0.027	**1.94**	**1.39**	**0.15**	-0.028	**1.88**	**1.37**

Our proposed methods perform similarly in terms of variance. Another important statistical measure is the bias. Imputing missing data should not induce substantial biases. The standard-CP based method leads to a relatively large bias. However, normalization of the tensor columns helps to limit the bias, to a level comparable to standard methods. The bias remains more or less constant with growing percentage of missing data. On the other hand, not surprisingly, the variance decreases with the percentage of missing data; the more missing data to be imputed, the smaller the variability in the data/responses.

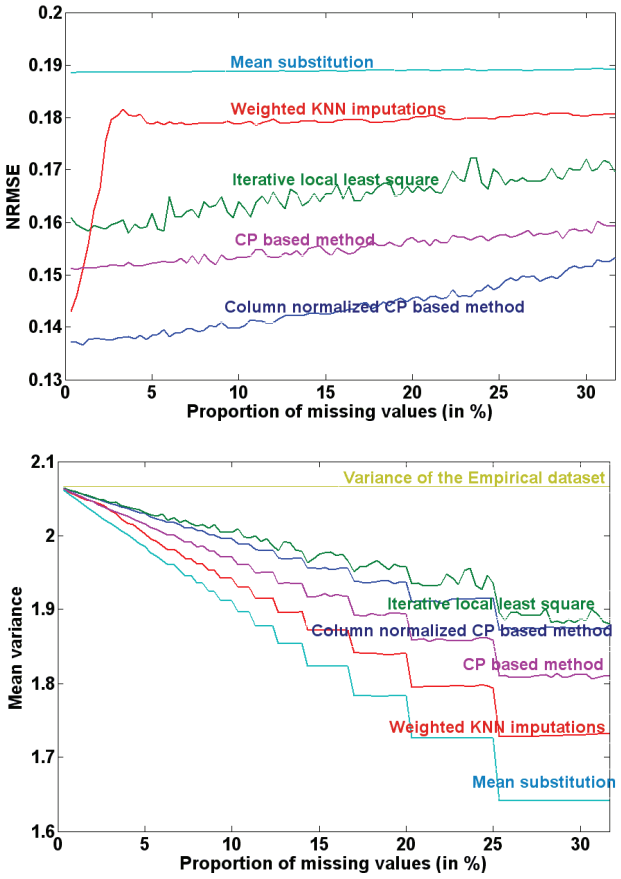

Figure 2. Comparison of Root Mean Square Error (NRMSE; top) and variance (bottom) for different missing data techniques. The variance of the original data set equals 2.05.

6. CONCLUSION

We have used CP decomposition to impute missing data in medical questionnaires. Our numerical results indicate that this approach outperforms standard methods including mean substitution, regression imputations, and weighted *k*-nearest neighbor (KNN) estimation. By normalizing the columns of the tensor before CP decomposition, we can further improve the prediction accuracy and variance. Presently we are exploring alternative advanced machine learning techniques for imputing missing data, such as graphical models. Also, we are working closely together with clinicians to better understand which type of questions are prone to missing data, and for which questions our proposed methods may or may not apply.

6. REFERENCES

[1] U. Müller-Bühl, B. Franke, K. Hermann, P. Engeser, Lowering missing item values in quality-of-life questionnaires: An interventional study, International Journal of Public Health, vol. 56 (1), 2011, pp. 63-69.

[2] S. Fielding, P. M. Fayers, C. R. Ramsay, Investigating the missing data mechanism in quality of life outcomes: a comparison of approaches, Health Qual Life Outcomes, 2009, pp. 22;7:57.

[3] T. Marwala, Introduction to Missing Data. Computational Intelligence for Missing Data Imputation, Estimation, and Management: Knowledge Optimization Techniques. Hershey: IGI Global, 2009, pp. 1-18.

[4] E. Acar, D. M. Dunlavy, T. G. Kolda, M. Mørup, Scalable Tensor Factorizations for Incomplete Data. Chemometrics and Intelligent Laboratory Systems, vol. 106(1), 2011, pp. 41-56.

[5] Dauwels J, Garg L, Earnest A, Pang LK (2011). Handling Missing Data in Medical Questionnaires Using Tensor Factorizations and Decompositions. The Eighth International Conference on Information, Communications, and Signal Processing (ICICS 2011). Singapore 13-16 December, 2011.

[6] G. Tomasi, Practical and computational aspects in chemometric data analysis. Ph.D. The Royal and Agricultural University, Frederiksberg, Denmark, May 2006.

[7] Z. Cai, M. Heydari, G. Lin, Iterated Local Least Squares Imputation for Microarray Missing Values. Journal of Bioinformatics and Computational Biology, vol. 4 (5), 2006, pp. 935-957.

[8] M. M. Subasi, E. Subasi, M. Anthony, P. L. Hammer, A new imputation method for incomplete binary data, Discrete Applied Mathematics, vol. 159 (10), 2011, pp. 1040-1047

[9] M. Mørup, Applications of Tensor (multiway array) factorizations and decompositions in data mining, Data mining and knowledge discovery, 1(1), January/February 2011, pp. 24-40.

[10] K. P. Leong, K. O. Kong, B. Thong, E. T. Koh, T. Y. Lian, C. L. The et al., Development and preliminary validation of a systemic lupus erythematosus-specific quality-of-life instrument (SLEQOL), Rheumatology, vol. 44, 2005;pp. 1267–1276.

[11] B. W. Bader, T. G. Kolda, MATLAB Tensor Toolbox Version 2.4, http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox/, 2010.

[12] U. M. Braga-Neto, E. Dougherty, Exact Performance of Error Estimators for Discrete Classifiers, Pattern Recognition, Vol. 38, No. 11, November 2005, pp. 1799-1814.