Chemometrics and Intelligent Laboratory Systems 106 (2011) 41–56

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Scalable tensor factorizations for incomplete data ☆

Evrim Acar [a], Daniel M. Dunlavy [c], Tamara G. Kolda [b,*], Morten Mørup [d]

[a] TUBITAK-UEKAE, Gebze, Turkey
[b] Sandia National Laboratories, Livermore, CA 94551-9159, United States
[c] Sandia National Laboratories, Albuquerque, NM 87123-1318, United States
[d] Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

ARTICLE INFO

Article history:
Received 1 December 2009
Received in revised form 28 July 2010
Accepted 4 August 2010
Available online 12 August 2010

Keywords:
Missing data
Incomplete data
Tensor factorization
CANDECOMP
PARAFAC
Optimization

ABSTRACT

The problem of incomplete data – i.e., data with missing or unknown values – in multi-way arrays is ubiquitous in biomedical signal processing, network traffic analysis, bibliometrics, social network analysis, chemometrics, computer vision, communication networks, etc. We consider the problem of how to factorize data sets with missing values with the goal of capturing the underlying latent structure of the data and possibly reconstructing missing values (i.e., tensor completion). We focus on one of the most well-known tensor factorizations that captures multi-linear structure, CANDECOMP/PARAFAC (CP). In the presence of missing data, CP can be formulated as a weighted least squares problem that models *only* the known entries. We develop an algorithm called CP-WOPT (CP Weighted OPTimization) that uses a first-order optimization approach to solve the weighted least squares problem. Based on extensive numerical experiments, our algorithm is shown to successfully factorize tensors with noise and up to 99% missing data. A unique aspect of our approach is that it scales to sparse large-scale data, e.g., $1000 \times 1000 \times 1000$ with five million known entries (0.5% dense). We further demonstrate the usefulness of CP-WOPT on two real-world applications: a novel EEG (electroencephalogram) application where missing data is frequently encountered due to disconnections of electrodes and the problem of modeling computer network traffic where data may be absent due to the expense of the data collection process.

1. Introduction

Missing data can arise in a variety of settings due to loss of information, errors in the data collection process, or costly experiments. For instance, in biomedical signal processing, missing data can be encountered during EEG analysis, where multiple electrodes are used to collect the electrical activity along the scalp. If one of the electrodes becomes loose or disconnected, the signal is either lost or discarded due to contamination with high amounts of mechanical noise. We also encounter the missing data problem in other areas of data mining, such as packet losses in network traffic analysis [2] and occlusions in images in computer vision [3]. Many real-world data with missing entries are ignored because they are deemed unsuitable for analysis, but this work contributes to the growing evidence that such data can be analyzed.

Unlike most previous studies on missing data which have only considered matrices, we focus here on the problem of missing data in *tensors* because it has been shown increasingly that data often have more than two modes of variation and are therefore best represented as multi-way arrays (i.e., tensors) [4,5]. For instance, in EEG data each signal from an electrode can be represented as a time-frequency matrix; thus, data from multiple channels is three-dimensional (temporal, spectral, and spatial) and forms a three-way array [6]. Social network data, network traffic data, and bibliometric data are of interest to many applications such as community detection, link mining, and more; these data can have multiple dimensions/modalities, are often extremely large, and generally have at least some missing data. These are just a few of the many data analysis applications where one needs to deal with large multi-way arrays with missing entries. Other examples of multi-way arrays with missing entries from different disciplines have also been studied in the literature [7–9]. For instance, [7] shows that, in spectroscopy, intermittent machine failures or different sampling frequencies may result in tensors with missing fibers (i.e., the higher-order analogues of matrix rows or columns, see Fig. 1). Similarly, missing fibers are encountered in multidimensional NMR (Nuclear Magnetic Resonance) analysis, where sparse sampling is used in order to reduce the experimental time [8].

Our goal is to capture the latent structure of the data via a higher-order factorization, even in the presence of missing data. Handling missing data in the context of matrix factorizations, e.g., the widely-used principal component analysis, has long been studied [10,11] (see

☆ A preliminary conference version of this paper has appeared as [1].

* Corresponding author.
E-mail addresses: evrim.acar@bte.tubitak.gov.tr (E. Acar), dmdunla@sandia.gov (D.M. Dunlavy), tgkolda@sandia.gov (T.G. Kolda), mm@imm.dtu.dk (M. Mørup).

0169-7439/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2010.08.004

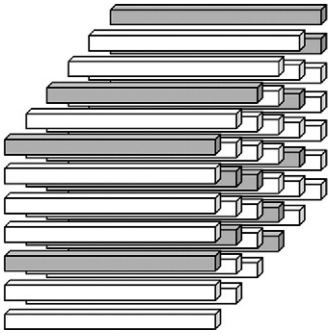

Fig. 1. A 3-way tensor with missing row fibers (in gray).

[3] for a review). It is also closely related to the matrix completion problem, where the goal is to recover the missing entries [12,13] (see Section 3 for more discussion). Higher-order factorizations, i.e., tensor factorizations, have emerged as an important method for information analysis [4,5]. Instead of flattening (unfolding) multi-way arrays into matrices and using matrix factorization techniques, tensor models preserve the multi-way nature of the data and extract the underlying factors in each mode (dimension) of a higher-order array.

We focus here on the CANDECOMP/PARAFAC (CP) tensor decomposition [14,15], which is a tensor model commonly used in various applications [6,16–19]. To illustrate differences between matrix and tensor factorizations, we introduce the CP decomposition for three-way tensors; discussion of the CP decomposition for general N-way tensors can be found in Section 4. Let $\mathcal{X}$ be a three-way tensor of size $I\times J\times K$, and assume its rank is R (see [5] for a detailed discussion on tensor rank). With perfect data, the CP decomposition is defined by *factor matrices* $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ of sizes $I\times R$, $J\times R$, and $K\times R$, respectively, such that

$$x_{ijk} = \sum_{r=1}^{R} a_{ir}b_{jr}c_{kr}, \quad \text{for all } i = 1,\ldots,I,\ j = 1,\ldots,J,\ k = 1,\ldots,K.$$

In the presence of noise, the true $\mathcal{X}$ is not observable and we cannot expect equality. Instead, the CP decomposition should minimize the error function

$$f(\mathbf{A},\mathbf{B},\mathbf{C}) = \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(x_{ijk} - \sum_{r=1}^{R} a_{ir}b_{jr}c_{kr}\right)^2. \tag{1}$$

An illustration of CP for third-order tensors is given in Fig. 2. The CP decomposition is extensible to N-way tensors for $N\geq 3$, and there are numerous methods for computing it [20].

In the case of incomplete data, a standard practice is to impute the missing values in some fashion (e.g., replacing the missing entries using average values along a particular mode). Imputation can be useful as long as the amount of missing data is small; however, performance degrades for large amounts of missing data [1,10]. As a better alternative, factorizations of the data with imputed values for missing entries can be used to re-impute the missing values and the procedure can be repeated to iteratively determine suitable values for the missing entries. Such a procedure is an example of the expectation maximization (EM) algorithm [21]. Computing CP decompositions by combining the alternating least squares method, which computes the factor matrices one at a time, and iterative imputation (denoted EM-ALS in this paper) has been shown to be quite effective and has the advantage of often being simple and fast. Nevertheless, as the amount of missing data increases, the performance of the algorithm may suffer since the initialization and the intermediate models used to impute the missing values will increase the risk of converging to a less than optimal factorization [7]. Also, the poor convergence of alternating

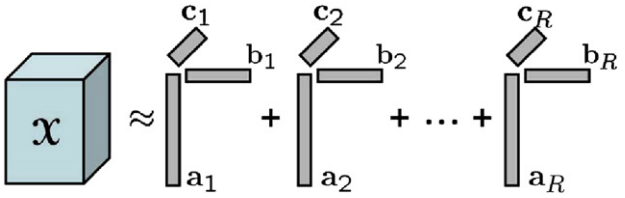

Fig. 2. Illustration of an R-component CP model for a third-order tensor X.

methods due to their vulnerability to flatlining, i.e., stagnation, is noted in [3].

In this paper, though, we focus on using a weighted version of the error function to ignore missing data and model only the known entries. In that case, nonlinear optimization can be used to directly solve the weighted least squares problem for the CP model. The weighted version of Eq. (1) is

$$f_{\mathcal{W}}(\mathbf{A},\mathbf{B},\mathbf{C}) = \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left\{w_{ijk}\left(x_{ijk} - \sum_{r=1}^{R} a_{ir}b_{jr}c_{kr}\right)\right\}^2, \tag{2}$$

where $\mathcal{W}$, which is the same size as $\mathcal{X}$, is a nonnegative weight tensor defined as

$$w_{ijk} = \begin{cases} 1 & \text{if } x_{ijk} \text{ is known,} \\ 0 & \text{if } x_{ijk} \text{ is missing,} \end{cases} \quad \text{for all } i = 1,\ldots,I,\ j = 1,\ldots,J,\ k = 1,\ldots,K.$$

Our contributions in this paper are summarized as follows. (a) We develop a scalable algorithm called CP-WOPT (CP Weighted OPTimization) for tensor factorizations in the presence of missing data. CP-WOPT uses first-order optimization to solve the weighted least squares objective function over all the factor matrices simultaneously. (b) We show that CP-WOPT can scale to sparse, large-scale data using specialized sparse data structures, significantly reducing the storage and computation costs. (c) Using extensive numerical experiments on simulated data sets, we show that CP-WOPT can successfully factor tensors with noise and up to 99% missing data. In many cases, CP-WOPT is significantly faster than the best published direct optimization method in the literature [7]. (d) We demonstrate the applicability of the proposed algorithm on a real data set in a novel EEG application where data is incomplete due to failures of particular electrodes. This is a common occurrence in practice, and our experiments show that even if signals from almost half of the channels are missing, underlying brain activities can still be captured using the CP-WOPT algorithm, illustrating the usefulness of our proposed method. (e) In addition to tensor factorizations, we also show that CP-WOPT can be used to address the tensor completion problem in the context of network traffic analysis. We use the factors captured by the CP-WOPT algorithm to reconstruct the tensor and illustrate that even if there is a large amount of missing data, the algorithm is able to keep the relative error in the missing entries close to the modeling error.

The paper is organized as follows. We introduce the notation used throughout the paper in Section 2. In Section 3, we discuss related work in matrix and tensor factorizations. The computation of the function and gradient values for the general N-way weighted version of the error function and the presentation of the CP-WOPT method are given in Section 4. Numerical results on both simulated and real data are given in Section 5. Conclusions and future work are discussed in Section 6.

2. Notation

Tensors of order $N\geq 3$ are denoted by Euler script letters ($\mathcal{X},\mathcal{Y},\mathcal{Z}$), matrices are denoted by boldface capital letters ($\mathbf{A},\mathbf{B},\mathbf{C}$), vectors are