Some Examples of Big Data Analysis Using Tensors

Martijn Boussé[*] MARTIJN.BOUSSE@KULEUVEN.BE
Otto Debals[*†] OTTO.DEBALS@KULEUVEN.BE
Nico Vervliet[*] NICO.VERVLIET@KULEUVEN.BE
Frederik Van Eeghem[*†] FREDERIK.VANEEGHEM@KULEUVEN.BE
Lieven De Lathauwer[*†] LIEVEN.DELATHAUWER@KULEUVEN.BE

[*]Department of Electrical Engineering (ESAT) STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium.

[†]Group Science, Engineering and Technology, KU Leuven Kulak, E. Sabbelaan 53, 8500 Kortrijk, Belgium.

Keywords: big data, curse of dimensionality, higher-order tensor, tensor decomposition

Abstract

Tensors tools provide some remarkable possibilities for the compact representation of big data and to break the curse of dimensionality (CoD). We give a short overview by means of some examples. First, we explain how compact tensor models can be used to solve large-scale blind source separation problems. Next, we discuss how incomplete tensors can break the CoD and how randomization can be used to compute models for large-scale tensors.

1. Tensors and tensor decompositions

Higher-order tensors are higher-dimensional arrays of numerical values and are natural generalizations of vectors (first order) and matrices (second order) (Kolda & Bader, 2009; Cichocki et al., 2015; Hackbusch, 2012). Recently, multilinear algebra has received increased attention, mostly because of the powerful tensor decompositions and their mild uniqueness properties, see Figure 1 for an example. These tensor models are more powerful than simple linear models, but are less complex to deal with than nonlinear models. They have several properties that are not present in conventional matrix techniques, making them proper tools for countless applications in signal processing and data analysis (Cichocki et al., 2015). The mild uniqueness, mentioned above, is a key property for blind source separation (BSS). Moreover, tensor tools provide a powerful framework for the compact representation and analysis of large and possibly high-dimensional datasets, as well as alleviating or even breaking the curse of dimensionality (CoD) (Cichocki, 2013; Vervliet et al., 2014), see, e.g., Figure 2.

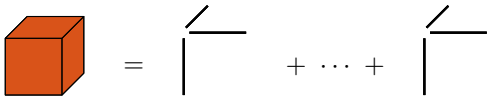


Figure 1. A canonical polyadic decomposition (CPD) writes a tensor as a sum of rank-1 terms, i.e., a sum of outer products of vectors, and is unique under mild conditions. The model is often used in blind source separation.

Tensorlab is MATLAB toolbox that provides state-of-the-art optimization-based algorithms to compute tensor decompositions and other tensor operations, supporting full, sparse, structured, and even incomplete tensors (Vervliet et al., 2016). Our toolbox allows us to couple the models and optionally impose structure, enabling structured data fusion. These tensor tools are not only limited to higher-dimensional data as a lot of matrix data can also be tensorized as in, e.g., tensor-based BSS, see (Debals & De Lathauwer, 2015).

2. Large-scale blind source separation

Many real-life signals are compressible, meaning that they can be represented by a number of parameters that is much smaller than the total number of samples. Often some tensor representation of such a signal allows an approximate low-rank model, as shown in Figure 2. This approach is known from tensor-based scientific computing and has successfully allowed one

Appearing in *Proceedings of Benelearn 2016*. Copyright 2016 by the author(s)/owner(s).

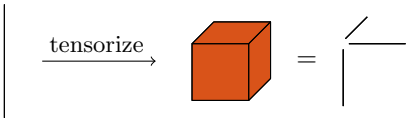


Figure 2. Some vectors admit a low-rank tensor model after reshaping it into a tensor. Such a representation is very compact because one needs only a few terms to describe the original vector. Several well-known signals such as (exponential) polynomials, trigonometric, and rational functions (Boussé et al., 2015b) admit such a tensor model.

to solve problems in a number of unknowns that exceed the number of atoms in the universe (Beylkin & Mohlenkamp, 2002; Grasedyck et al., 2013; Khoromskij, 2012; Oseledets, 2011; Vervliet et al., 2014).

We introduced this signal model in BSS to tackle large-scale problems (Boussé et al., 2015a; Boussé et al., 2015b). We showed that our method reformulates BSS as a tensor decomposition, assuming the sources admit the hypothesized low-rank structure. The tensor is obtained by applying a tensorization method called segmentation on the observed data matrix. It is the careful exploitation of the low-rank structure that enables a unique separation of the sources.

In applications with many sensors and high sensor density often the mixture is compressible as well, which allows us to apply the same strategy as above. We extended our method to handle the low-rank structure on *both* levels of the BSS problem, enabling a very compact representation that allows one to solve large-scale BSS problems. Finally, we applied the same strategy for large-scale blind system identification (BSI) by assuming the system coefficients have a low-rank structure (Boussé et al., 2016a; Boussé et al., 2016b).

3. Designing new materials using incomplete tensor decompositions

Designing new materials with certain physical properties is an important problem in material sciences. In the case of alloys, e.g., one can vary the concentrations of the different metals to obtain materials with specific properties. Measuring for example the melting point of an alloy as a function of the concentrations leads to a higher-order tensor: assume we measure the melting point as a function of K concentration values of N different constituent materials, resulting in a tensor containing K^{N-1} values. (The last concentration is not needed because concentrations add up to 100%.) Clearly, the number of values explodes for large N, even when K is small. Typically, this tensor can only be partially filled empirically due to the CoD.

Tensorlab's built-in support for incomplete tensors allows us to model the melting point in the above example using the decomposition of only the incomplete tensor (Vervliet et al., 2014; Vervliet et al., 2016), enabling a simple, fast, and cost-effective approach for exploration of new materials. A first successful case study was performed in (Vervliet et al., 2014; Sorber et al., 2015), where we considered an alloy consisting of $N = 10$ different materials. The complete tensor would contain approximately $100^{N-1} = 10^{18}$ elements. The measured dataset contained only $130,000$ random measurements of the melting temperature because measuring all the elements is clearly infeasible.

4. Decomposing large-scale tensors

The canonical polyadic decompositions (CPD), as depicted in Figure 1, is a powerful and compact tensor model for many applications in fields such as signal processing, biomedical sciences, and machine learning. Decomposing a large-scale tensor, however, poses some computational challenges due to the CoD: the memory and computational complexity depends linearly on the number of entries which increase exponentially with the order of the tensor.

Over the years several strategies have been proposed to tackle this problem, see (Vervliet & De Lathauwer, 2016) and references therein for a short overview. As a matter of fact, the approach described in the previous section is such a strategy to break the CoD. The complexity of our algorithms for incomplete tensors is linear in the number of *known* entries instead of the total number of entries. Clearly, the former can be far lower than the latter, as was exemplified by the application from material sciences. One way to deal with large-scale tensors is then to deliberately undersample the tensor.

In (Vervliet & De Lathauwer, 2016), we proposed a randomized block sampling approach to compute the CPD of a tensor. In each iteration, the algorithm first samples a new random block of entries which are not necessarily adjacent. Next, only a limited number of variables are updated using the previous results as initialization because of the 'locality property' of the CPD, see (Vervliet & De Lathauwer, 2016) for more details. This approach allowed us to compute a decomposition of a 12.5 GB real life and 8 TB synthetic tensor in only a few minutes on a standard PC.

References

Beylkin, G., & Mohlenkamp, M. J. (2002). Numerical operator calculus in higher dimensions. *Proceedings of the National Academy of Sciences of the United States of America* (pp. 10246–10251).

Boussé, M., Debals, O., & De Lathauwer, L. (2015a). A novel deterministic method for large-scale blind source separation. *Proceedings of the 23rd European Signal Processing Conference (EUSIPCO 2015, Nice, France)* (pp. 1935–1939).

Boussé, M., Debals, O., & De Lathauwer, L. (2015b). A tensor-based method for large-scale blind source separation using segmentation. *Technical Report 15-59, ESAT-STADIUS, KU Leuven, Leuven, Belgium.*

Boussé, M., Debals, O., & De Lathauwer, L. (2016a). Tensor-based large-scale blind system identification using segmentation. *Technical Report 16-109, ESAT-STADIUS, KU Leuven, Leuven, Belgium.*

Boussé, M., Debals, O., & De Lathauwer, L. (2016b). A tensor-based method for large-scale blind system identification using segmentation. *Technical Report 16-38, ESAT-STADIUS, KU Leuven, Leuven, Belgium.* (Accepted for publication in the Proceedings of the 24th European Signal Processing Conference (EUSIPCO 2016, Budapest, Hungary).

Cichocki, A. (2013). Era of big data processing: A new approach via tensor networks and tensor decompositions. *International Workshop on Smart Info-Media Systems in Asia (Nagoya, Japan).*

Cichocki, A., Mandic, D. P., De Lathauwer, L., Zhou, G., Zhao, Q., Caiafa, C. F., & Phan, A. H. (2015). Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE Signal Processing Magazine*, *32*, 145–163.

Debals, O., & De Lathauwer, L. (2015). Stochastic and deterministic tensorization for blind signal separation. In *Latent variable analysis and signal separation*, vol. 9237 of *Lecture Notes in Computer Science*, 3–13. Springer Berlin / Heidelberg.

Grasedyck, L., Kressner, D., & Tobler, C. (2013). A literature survey of low-rank tensor approximation techniques. *GAMM-Mitteilungen*, *36*, 53–78.

Hackbusch, W. (2012). *Tensor spaces and numerical tensor calculus*, vol. 42. Springer.

Khoromskij, B. N. (2012). Tensors-structured numerical methods in scientific computing: Survey on recent advances. *Chemometrics and Intelligent Laboratory Systems*, *110*, 1–19.

Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, *51*, 455–500.

Oseledets, I. (2011). Tensor-train decomposition. *SIAM Journal of Scientific Computing*, *33*, 2295–2317.

Sorber, L., Van Barel, M., & De Lathauwer, L. (2015). Stuctured data fusion. *IEEE Journal of Selected Topics in Signal Processing*, *9*, 586–600.

Vervliet, N., & De Lathauwer, L. (2016). A randomized block sampling approach to canonical polyadic decomposition of large-scale tensors. *IEEE Journal on Selected Topics in Signal Processing*, *10*, 284–295.

Vervliet, N., Debals, O., Sorber, L., & De Lathauwer, L. (2014). Breaking the curse of dimensionality using decompositions of incomplete tensors: Tensor-based scientific computing in big data analysis. *IEEE Signal Processing Magazine*, *31*, 71–79.

Vervliet, N., Debals, O., Sorber, L., Van Barel, M., & De Lathauwer, L. (2016). Tensorlab 3.0.