Balancing Interpretability and Predictive Accuracy for Unsupervised Tensor Mining

Ishmam Zabir
Dept. of Electrical and Computer Engineering
University of California Riverside
Email: izabi001@ucr.edu

Evangelos E. Papalexakis
Dept. of Computer Science and Engineering
University of California Riverside
Email: epapalex@cs.ucr.edu

***Abstract*—The PARAFAC tensor decomposition has enjoyed an increasing success in exploratory multi-aspect data mining scenarios. A major challenge remains the estimation of the number of latent factors (i.e., the rank) of the decomposition, which is known to yield high-quality, interpretable results. Previously, AutoTen, an automated tensor mining method which leverages a well-known quality heuristic from the field of Chemometrics, the Core Consistency Diagnostic (CORCONDIA), in order to automatically determine the rank for the PARAFAC decomposition, was proposed. In this work, building upon AutoTen, we set out to explore the trade-off between 1) the interpretability of the results (as expressed by CORCONDIA), and 2) the predictive accuracy of the decomposition, towards improving rank estimation quality. Our preliminary results indicate that striking a good balance in that trade-off yields high-quality rank estimation, towards achieving unsupervised tensor mining.**

I. Introduction

Very frequently, tensor mining is done in an entirely unsupervised way, since ground truth and labels are either very expensive or hard to obtain. Our problem, thus, is: given a potentially very large and sparse tensor, and its R-component PARAFAC decomposition [1], compute a quality measure for that decomposition. Subsequently, using that quality metric, we would like to identify a "good" number R of components, and ultimately minimize human intervention and trial-and-error fine tuning. This problem is extremely hard. In fact, even computing the rank of a tensor has been shown to be an NP-hard problem [2], in stark contrast to the matrix rank which can be easily computed in polynomial time (via the Singular Value Decomposition). In this work, we leverage previous work on automated tensor rank estimation [3], and we set out to explore the trade-off between different quality heuristics. We propose a method that successfully balances model interpretability and predictive accuracy for missing data, towards estimating the rank of the PARAFAC tensor decomposition.

II. Background and Related Work

Even though tensor rank is an extremely hard problem, fortunately, there exist heuristics that are able to assist with the above problem and have been shown to work well in practice, in the field of Chemometrics. Such a powerful and intuitive heuristic is the so-called "Core Consistency Diagnostic" [4], [5], which given a tensor and its PARAFAC decomposition, provides a quality measure, which we can in turn use as a proxy of how interpretable our results are. In our previous work [3] we introduce AutoTen, a comprehensive unsupervised and automatic tensor mining method that provides quality assessment of the results by trading off the decomposition quality (measured via the Core Consistency Diagnostic), and the number of latent components one can extract with high-enough quality. AutoTen outperforms state-of-the-art approaches, such as [6] and [7].

III. Proposed Method

The PARAFAC decomposition presents a unique opportunity: it admits a very intuitive interpretation (that of soft-clustering the entities involved in all the modes of the tensor, for a detailed justification, we refer the interested reader to [8]) and albeit heuristic, we have a few means of judging how well our results adhere to the PARAFAC model, and in turn, how interpretable they are. On the other hand, the PARAFAC decomposition has been successfully used for collaborative filtering [9] where the *prediction accuracy* of missing values is the focus. In fact, the notion of judging the quality of the PARAFAC decomposition dates back to the N-Way Toolbox for Matlab [10], which provides a cross-validation method for rank estimation that uses held-out prediction accuracy.

The PARAFAC decomposition is, thus, capable of achieving both high interpretability and high prediction accuracy, but there is a catch. *Unfortunately, a highly predictive decomposition need not be of high quality for interpretability purposes, and vice versa*: for instance, a degenerate PARAFAC decomposition where two or more components are linearly dependent would yield the same prediction accuracy if only one of those components remained in the results after proper scaling, however,

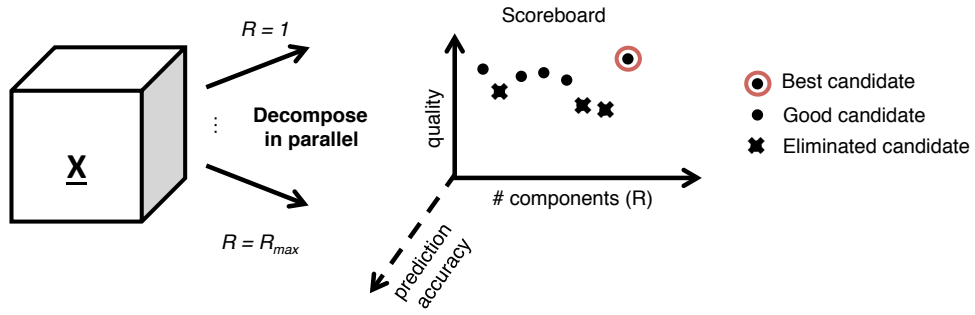

Fig. 1: Trading off interpretability and prediction accuracy.

the redundant components hurt the interpretability of our results.

Here, we propose to augment AUTOTEN in order to identify a decomposition that combines the best of both worlds: high interpretability and high prediction accuracy. At a high level, as shown in Figure 1, our method will seek to find a set of results that give the best trade-off between quality (which might be multi-dimensional) and as prediction accuracy, which is measured by root mean squared error (RMSE) on held-out data. At the same time, our method will aim at maximizing the number of components we can extract that yield high interpretability *and* high prediction accuracy. This is a multi-objective optimization problem and part of our future research revolves around how to best frame it so that we can solve it efficiently without having to exhaustively search the space of solutions. For the sake of our preliminary work here, we use a parameter-free 2-means clustering scheme, similar to [3] which, in a nutshell, separates the different solutions into two clusters, one with high-quality, acceptable solutions, and one with unacceptable solutions, and subsequently identifies a "good" solution from the high-quality cluster.

IV. EXPERIMENTAL EVALUATION

For our preliminary analysis we use the N-way toolbox for Matlab implementations of the Frobenius norm ALS algorithm for PARAFAC, as well as the missing value PARAFAC algorithm. We generated $R \times R \times R$ random tensors with rank R (the generation of those synthetic tensors follows the methodology described in detail in [3]) and we recover their rank using the following methods, and we plot the mean rank estimation error (between the actual rank and the one estimated by each method) in Figure 2:

- AUTOTEN: This is the method proposed in [3] which maximizes CORCONDIA and number of components.
- AUTOTEN-REC: This is the augmented AUTOTEN where both CORCONDIA and reconstruction error are considered as features. In [11] the authors provide guidelines on using Core Consistency in conjunction with the fit for rank selection, however, our work is the first method that does so automatically.
- AUTOTEN-MV: This is the augmented AUTOTEN where both CORCONDIA and missing value prediction are considered as features. *This is the method we ultimately propose in this paper.*
- Baseline-1: This method uses the RMSE in held-out data, and returns the rank of the decomposition for which the reconstruction error stopped decreasing (by a small number, set to 10^{-6}).

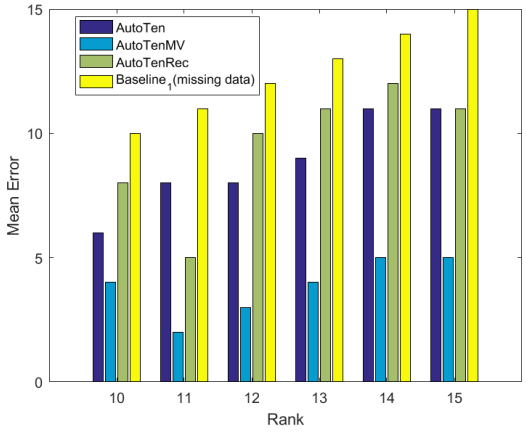

Fig. 2: Balancing interpretability (measured by the Core Consistency) and prediction accuracy on held-out data yields more accurate rank estimation for the PARAFAC decomposition.

In Figure 1 we clearly observe that our proposed method AUTOTEN-MV outperforms the rest of the baselines across the board.

V. Conclusions

In this work we demonstrate the utility of balancing interpretability and prediction accuracy of missing data towards estimating the rank of the PARAFAC decomposition, for unsupervised tensor mining. Our preliminary results indicate that balancing the trade-off between the Core Consistency and prediction accuracy on held-out data (both popular measures of decomposition quality, in isolation), results in more accurate rank estimation than by focusing on one of the two quality measures, or using other state-of-the-art methods.

VI. Acknowledgements

Research was supported by the Bourns College of Engineering at UCR and an Adobe Data Science Research Faculty Award. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding parties.

References

[1] R. Harshman, "Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis," 1970.

[2] J. Håstad, "Tensor rank is np-complete," *Journal of Algorithms*, vol. 11, no. 4, pp. 644–654, 1990.

[3] Papalexakis, Evangelos E, "Automatic unsupervised tensor mining with quality assessment," in *SIAM SDM*, 2016.

[4] R. Bro and H. A. Kiers, "A new efficient method for determining the number of components in parafac models," *Journal of chemometrics*, vol. 17, no. 5, pp. 274–286, 2003.

[5] J. P. C. da Costa, M. Haardt, and F. Romer, "Robust methods based on the hosvd for estimating the model order in parafac models," in *Sensor Array and Multichannel Signal Processing Workshop, 2008. SAM 2008. 5th*. IEEE, 2008, pp. 510–514.

[6] Q. Zhao, L. Zhang, and A. Cichocki, "Bayesian cp factorization of incomplete tensors with automatic rank determination."

[7] M. Mørup and L. K. Hansen, "Automatic relevance determination for multi-way models," *Journal of Chemometrics*, vol. 23, no. 7-8, pp. 352–363, 2009.

[8] E. Papalexakis, C. Faloutsos, and N. Sidiropoulos, "Tensors for data mining and data fusion: Models, applications, and scalable algorithms," *ACM Trans. on Intelligent Systems and Technology*.

[9] L. Xiong, X. Chen, T.-K. Huang, J. G. Schneider, and J. G. Carbonell, "Temporal collaborative filtering with bayesian probabilistic tensor factorization." in *SDM*, vol. 10. SIAM, 2010, pp. 211–222.

[10] C. Andersson and R. Bro, "The n-way toolbox for matlab," *Chemometrics and Intelligent Laboratory Systems*, vol. 52, no. 1, pp. 1–4, 2000.

[11] M. H. Kamstrup-Nielsen, L. G. Johnsen, and R. Bro, "Core consistency diagnostic in parafac2," *Journal of Chemometrics*, vol. 27, no. 5, pp. 99–105, 2013.