Automatic Unsupervised Tensor Mining with Quality Assessment - Supplementary Material

Evangelos E. Papalexakis
Carnegie Mellon University
epapalex@cs.cmu.edu

1 Introduction

This is the supplementary material of the SDM 2016 submission. In Section 2 we discuss details on AUTOTEN. Section 3 shows results on a data mining case study of AUTOTEN on the `Amazon co-purchase` dataset, and finally Section 4 is an overview of tensor applications in data mining, highlighting the importance of the topic for data mining researchers and practitioners.

2 Algorithmic Details & Further Discussion

As we describe on the main text, the data-driven algorithm for choosing the "best" point (F^*, c^*) is the following:

- **Max c step**: Given vector $\mathbf{c}$, run 2-means clustering on its values. This will essentially divide the vector into a set of good/high values and a set of low/bad ones. If we call m_1, m_2 the means of the two clusters, then we select the cluster index that corresponds to the maximum between m_1 and m_2.
- **Max F step**: Given the cluster of points with maximum mean, we select the point that maximizes the value of F. We call this point (F^*, c^*).

Figure 1 shows pictorially the output of this two-step algorithm for a set of points taken from a real dataset. Note that the choice of the above algorithm, *intuitively*, is a good compromise between the quality of the decomposition, as indicated by the CORCONDIA value, as well as the number of latent patterns that we uncover.

Another alternative is to formally define a function of c, F that we wish to maximize, and select the maximum via enumeration. Coming up with the particular function to maximize, considering the intuitive objective of maximizing the number of components that we can extract with reasonably high quality (c), is a hard problem, and we risk biasing the selection with a specific choice of a function. Nevertheless, an example such function can be $g(c, F) = logclogF$ for $c > 0$, and $g(0, F) = 0$; this function essentially measures the area of the rectangle formed by the lines connecting (F, c) with the axes (in the log-log space) and intuitively seeks to find a good compromise between maximizing F and c. This function performs closely to the proposed data-driven approach and we defer a detailed discussion and investigation to future work.

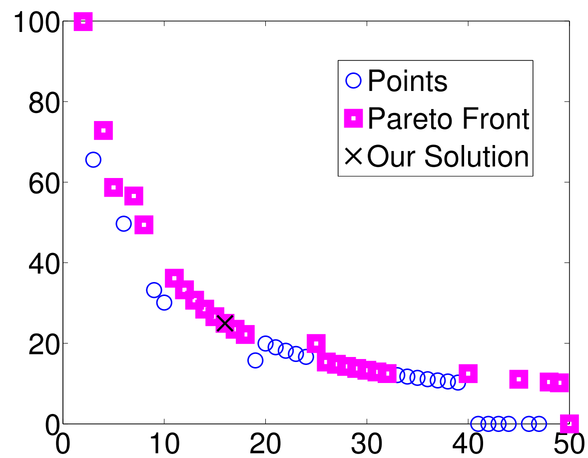
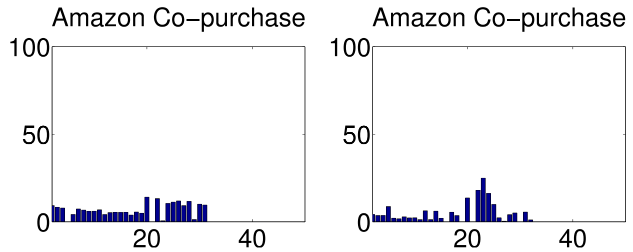


Figure 1: Example of choosing a good point

After choosing the "best" points (F^*_{Fro}, c^*_{Fro}) and (F^*_{KL}, c^*_{KL}), at the final step of AUTOTEN, we have to select between the results of `CP_ALS` and `CP_APR`. In order do so, we can use the following strategies:

1. Calculate
$$s_{Fro} = \sum_f \mathbf{c}_{Fro}(f)$$
and
$$s_{KL} = \sum_f \mathbf{c}_{KL}(f),$$
and select the method that gives the largest sum. The intuition behind this data-driven strategy is choosing the loss function that is able to discover results with higher quality on aggregate, for more potential ranks.
2. Select the results that produce the maximum value between c^*_{Fro} and c^*_{KL}. This strategy is conservative and aims for the highest quality of results, possibly to the expense of components of lesser quality that could still be acceptable for exploratory analysis.



(a) Frobenius norm (b) KL-Divergence

Figure 2: Core Consistency for the `Amazon co-purchase` dataset for $F = 2 \cdots 50$.

3. Select the results that produce the maximum value between F^*_{Fro} and F^*_{KL}. Contrary to the previous strategy, this one is more aggressive, aiming for the highest number of components that can be extracted with acceptable quality.

Empirically, the last strategy seems to give better results, however they all perform very closely in synthetic data. Particular choice of strategy depends on the application needs, e.g. if quality of the components is imperative to be high, then strategy 2 should be preferred over strategy 3.

3 Data Mining Case Study

3.0.1 Analyzing `Amazon co-purchase` This dataset records pairs of products that were purchased together by the same customer on Amazon, as well as the category of the first product in the pair. This dataset, as shown in Figures 2(a) and 2(b) does not have perfect trilinear structure, however a low rank trilinear approximation still offers reasonably good insights for product recommendation and market basket analysis.

By analyzing this dataset, we seek to find coherent groups of products that people tend to purchase together, aiming for better product recommendations and suggestions. For the purposes of this study, we extracted a small piece of the co-purchase network of 256 products. AUTOTEN was able to extract 24 components by choosing KL-Divergence as a loss.

On Table 1 we show a representative subset of our resulting components (which were remarkably sparse, due to the KL-Divergence fitting by `CP_APR`). We observe that products of similar genre and themes tend to naturally cluster together. For instance, cluster #1 contains mostly self improvement books. We also observe a few topical outliers, such as the book `How to Kill a Monster (Goosebumps)` in cluster #1, and CD `Desde Que Samba E Samba` in cluster #3 that contains Technical / Software Development books.

4 Tensors and their data mining applications

We have elaborated on relevant prior work throughout the text, however here, we first show that there is a vast number of tensor applications in data mining, that can potentially benefit from our work.

One of the first applications was on web mining, extending the popular HITS algorithm [10]. There has been work on analyzing citation networks (such as DBLP) [11], detecting anomalies in computer networks[11, 14, 16], extracting patterns from and completing Knowledge Bases [4] and analyzing time-evolving or multi-view social networks. [1, 11, 12, 9], The long list of application continues, with extensions of Latent Semantic Analysis [2, 3], extensions of Subspace Clustering to higher orders [8], Crime Forecasting [15], Image Processing [13], mining Brain data [5, 6], trajectory and mobility data [18, 17], and medical data [7].

References

[1] Brett W Bader, Richard A Harshman, and Tamara G Kolda. Temporal analysis of semantic graphs using asalsan. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 33–42. IEEE, 2007.

[2] Deng Cai, Xiaofei He, and Jiawei Han. Tensor space model for document analysis. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 625–626. ACM, 2006.

[3] Kai-Wei Chang, Wen-tau Yih, and Christopher Meek. Multi-relational latent semantic analysis. In *EMNLP*, pages 1602–1612, 2013.

[4] Kai-Wei Chang, Wen-tau Yih, Bishan Yang, and Christopher Meek. Typed tensor decomposition of knowledge bases for relation extraction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1568–1579, 2014.

[5] Ian Davidson, Sean Gilpin, Owen Carmichael, and Peter Walker. Network discovery via constrained tensor analysis of fmri data. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 194–202. ACM, 2013.

[6] Lifang He, Xiangnan Kong, S Yu Philip, Ann B Ragin, Zhifeng Hao, and Xiaowei Yang. Dusk: A dual structure-preserving kernel for supervised tensor learning with applications to neuroimages. *matrix*, 3(1):2, 2014.

[7] Joyce C Ho, Joydeep Ghosh, and Jimeng Sun. Marble: high-throughput phenotyping from electronic health records via sparse nonnegative tensor factorization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 115–124. ACM, 2014.

[8] Heng Huang, Chris Ding, Dijun Luo, and Tao Li. Simultaneous tensor subspace selection and clustering: the equivalence of high order svd and k-means clustering. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pages 327–335. ACM, 2008.

[9] Meng Jiang, Peng Cui, Fei Wang, Xinran Xu, Wenwu Zhu, and Shiqiang Yang. Fema: flexible evolutionary multi-faceted analysis for dynamic behavioral pattern discovery. In

Cluster type	Products	Product Types
#1 Self Improvement	Resolving Conflicts At Work : A Complete Guide for Everyone on the Job	Book
	How to Kill a Monster (Goosebumps)	Book
	Mensa Visual Brainteasers	Book
	Learning in Overdrive: Designing Curriculum, Instruction, and Assessment from Standards : A Manual for Teachers	Book
#2 Psychology, Self Improvement	Physicians of the Soul: The Psychologies of the World's Greatest Spiritual Leaders	Book
	The Rise of the Creative Class: And How It's Transforming Work, Leisure, Community and Everyday Life	Book
#3 Technical Books	Beginning ASP.NET Databases using C#	Book
	BizPricer Business Valuation Manual wSoftware	Book
	Desde Que Samba E Samba	Music
#4 History	War at Sea: A Naval History of World War II	Book
	Jailed for Freedom: American Women Win the Vote	Book
	The Perfect Plan (7th Heaven)	Book

Table 1: Latent components of the `Amazon co-purchase` dataset, as extracted using AUTOTEN

Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1186–1195. ACM, 2014.

[10] Tamara G Kolda, Brett W Bader, and Joseph P Kenny. Higher-order web link analysis using multilinear algebra. In *Data Mining, Fifth IEEE International Conference on*, pages 8–pp. IEEE, 2005.

[11] Tamara G Kolda and Jimeng Sun. Scalable tensor decompositions for multi-aspect data mining. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 363–372. IEEE, 2008.

[12] Yu-Ru Lin, Jimeng Sun, Paul Castro, Ravi Konuru, Hari Sundaram, and Aisling Kelliher. Metafac: community discovery via relational hypergraph factorization. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 527–536. ACM, 2009.

[13] Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):208–220, 2013.

[14] K. Maruhashi, F. Guo, and C. Faloutsos. Multiaspectforensics: Pattern mining on large-scale heterogeneous networks with tensor analysis. In *Proceedings of the Third International Conference on Advances in Social Network Analysis and Mining*, 2011.

[15] Yang Mu, Wei Ding, Melissa Morabito, and Dacheng Tao. Empirical discriminative tensor analysis for crime forecasting. In *Knowledge Science, Engineering and Management*, pages 293–304. Springer, 2011.

[16] Evangelos E Papalexakis, Christos Faloutsos, and Nicholas D Sidiropoulos. Parcube: Sparse parallelizable tensor decompositions. In *Machine Learning and Knowledge Discovery in Databases*, pages 521–536. Springer, 2012.

[17] Yilun Wang, Yu Zheng, and Yexiang Xue. Travel time estimation of a path using sparse trajectories. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 25–34, New York, NY, USA, 2014. ACM.

[18] Vincent Wenchen Zheng, Bin Cao, Yu Zheng, Xing Xie, and Qiang Yang. Collaborative filtering meets mobile recommendation: A user-centered approach. In *AAAI*, volume 10, pages 236–241, 2010.