Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

# Scalable Nonparametric Tensor Analysis

**Shandian Zhe**
Computer Science Department
Purdue University
szhe@purdue.edu

Multiway data, described by tensors, are common in real-world applications. For example, online advertising click logs can be represented by a three-mode tensor *(user, advertisement, context)*. The analysis of tensors is closely related to many important applications, such as click-through-rate (CTR) prediction, anomaly detection and product recommendation. Despite the success of existing tensor analysis approaches, such as Tucker (Tucker 1966), CANDECOMP/PARAFAC (CP) (Harshman 1970) and infinite Tucker decompositions (Xu, Yan, and Qi 2012), they are either not enough powerful to capture complex hidden relationships in data, or not scalable to handle real-world large data. In addition, they may suffer from the extreme sparsity in real data, i.e., when the portion of nonzero entries is extremely low; they lack of principled ways to discover other patterns—such as an unknown number of latent clusters—which are critical for data mining tasks such as anomaly detection and market targeting .

To address these challenges, I used nonparametric Bayesian techniques, such as Gaussian processes (GP) and Dirichlet processes (DP), to model highly nonlinear interactions and to extract hidden patterns in tensors; I derived tractable variational evidence lower bounds, based on which I developed scalable, distributed or online approximate inference algorithms. The works are summarized as follows.

- **Distributed infinite Tucker decomposition (DinTucker)**. Infinite Tucker decomposition (InfTucker)(Xu, Yan, and Qi 2012) is the first tensor factorization approach that can capture complex, nonlinear relationship between the latent factors: It generalizes the Tucker decomposition into an infinite latent feature space via a Gaussian process model on tensors, namely Tensor-variate Gaussian process (TGP). However, InfTucker has a very limited scalability because of its global GP assumption: It treats the entire vectorized tensor as one single TGP projection and hence leads to huge covariance matrices even for very small tensors. For example, a $100 \times 100 \times 100$ tensor has a $10^6 \times 10^6$ covariance matrix in the InfTucker modeling. To address this issue, I used local GP strategies and proposed a distributed Infinite tucker decomposition model. Specifically, I broke the whole tensor into many, many small subtensors where each subtensor is modeled by a local InfTucker; all the local InfTuckers share the same set of global latent facotors via a hierarchical Bayesian model. Then I developed a distributed stochastic inference algorithm in MAPREDUCE framework, where a set of subtensors are allocated to each MAPPER and a stochastic gradient descent algorithm is used to learn the local latent factors, and the REDUCER aggregates all the local latent factors to update the global latent factors. Experiments demonstrate that the proposed DinTucker algorithm has a similar predictive performance to InfTucker on small data, and is able to process large real tensors will billions of elements. In addition, I finished a few theoretical analysis and found out a close connection between InfTucker and DinTucker in terms of model evidence. The conclusions apply to general global GP and local GP models. The details of this work are in (Zhe et al. 2016b).
- **Simultaneously online factorization and latent cluster discovery.** Continuing with the local GP idea, I further placed Dirchlet process mixture (DPM) (Antoniak 1974) prior over the latent factors in order to discover latent cluster structures in each tensor mode. Therefore, for a *(user, movie, time)* tensor, my second work aims not only to capture the (possible) complex interactions between users, movies and times, but also to discover clusters of users who might have a preference for particular movies types, clusters of movies similar in topics and types, and clusters of watching times. The DPM prior is a nonparametric prior which assumes an unbounded number of clusters. Thereby the posterior inference can automatically infer the appropriate cluster numbers as well as the memberships, given the observed data. For model estimation, I developed an online variational Bayes Expectation-Maximization (VB-EM) algorithm. The algorithm sequentially processes each subtensor (similar to DinTucker, the algorithm first splits the whole tensor into smaller subtensors): In the E step, it updates the variational posteriors for latent cluster memberships. This is efficiently done by caching the global statistics and updating only with local statistics. In the M-step, it uses stochastic gradient descent to optimize the latent factors. Evaluations on real large tensors with billions of elements have shown that the DPM prior not only helps discover la-

tent clusters, but also benefits prediction tasks. The details of this work are given in (Zhe et al. 2015)

- **Distributed flexible nonlinear tensor factorization**. This works aims to tackle the data sparsity issue. As is often the case, real-world tensor data are usually extremely sparse, i.e., a relatively small number of nonzero entries are overwhelmed by by a large amount of zero elements. For example, the portion of nonzero elements in real tensors is often less than one percent. However, many zero elements are actually meaningless—they are either missing or unobserved. Incorporating them all into the factorization can lead to a severe learning bias toward zero. Existing nonlinear factorization approaches are based on TGP formulations, which implicitly assumes that the tensor is dense: The Kronecker product in the covariance structure enforces the model to include all the tensor entries for training. To address this issue, I proposed a new GP factorization model, which gets rid of the Kronecker product and can use any subset of tensor entries for training. One can use a balanced zero and nonzero tensor entries to avoid the learning bias; one can choose the meaningful entries according to their domain knowledge. Specifically, the new model assigns the GP prior over the tensor entries, where the input for each entry is the concatenation of the corresponding latent factors in each tensor mode. Thereby, the Kronecker product structure is eliminated from the model covariance. Further, to handle large data, I derived a tractable and tight variational evidence lower bound (ELBO) using functional derivatives and convex conjugates. The new ELBO subsumes the optimal variational posteriors and thus avoids inefficient, sequential E-M updates and enables efficient, parallel computation and improved inference quality. Based on the new EBLO, I developed a key-value free distributed inference algorithm in MAPREDUCE framework, which greatly reduces the IO cost and fully exploits the memory cache mechanism in fast platforms such as SPARK. Experiments demonstrate that the model not only outperforms DinTucker in terms of predictive accuracy, but also gains much faster running speed. The details of the work are given in (Zhe et al. 2016a)

## References


Antoniak, C. E. 1974. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics* 1152–1174.

Harshman, R. A. 1970. Foundations of the PARAFAC procedure: Model and conditions for an"explanatory"multimode factor analysis. *UCLA Working Papers in Phonetics* 16:1–84.

Tucker, L. 1966. Some mathematical notes on three-mode factor analysis. *Psychometrika* 31:279–311.

Xu, Z.; Yan, F.; and Qi, Y. 2012. Infinite Tucker decomposition: Nonparametric Bayesian models for multiway data analysis. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*.

Zhe, S.; Xu, Z.; Chu, X.; Qi, Y.; and Park, Y. 2015. Scalable nonparametric multiway data analysis. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Zhe, S.; Zhang, K.; Wang, P.; Lee, K.; Xu, Z.; Qi, Y.; and Ghahramani, Z. 2016a. Distributed flexible nonlinear tensor factorization. In *Advances in Neural Information Processing Systems (NIPS)*.

Zhe, S.; Qi, Y.; Park, Y.; Xu, Z.; Molloy, I.; and Chari, S. 2016b. Dintucker: Scaling up gaussian process models on large multidimensional arrays. In *Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*.