Bayesian Learning of Non-Negative Matrix/Tensor Factorizations by Simulating Pólya Urns

M. Burak Kurutmaz, A. Taylan Cemgil, Melih Barsbey
Boğaziçi University, İstanbul, Turkey
Umut Şimşekli
LTCI, Télécom ParisTech, Université Paris-Saclay, Paris, France
Sinan Yıldırım
Sabancı University, İstanbul, Turkey

Abstract

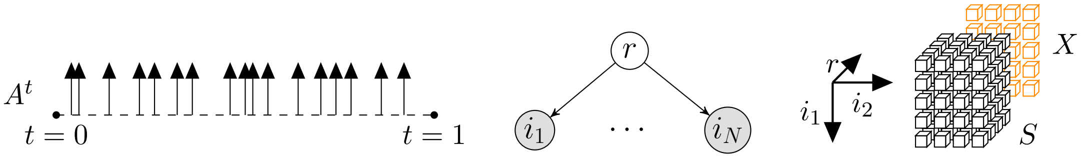
We introduce a dynamic generative model, Bayesian Allocation Model (BAM) that makes the connection among nonnegative tensor decompositions, topic models, and Bayesian networks explicit. Our construction is based on marking the events of a Poisson process using a Bayesian network and integrating out the parameters analytically where the resulting marginal process is a Pólya urn. The marginal likelihood can then be characterized as the probability that an urn process hits a certain small set. This alternative random process perspective allows us to develop sequential Monte Carlo (SMC) algorithms for computing nonnegative tensor decompositions and the marginal likelihood whose computational complexity is independent from the tensor dimensions. Our analysis and simulation results suggest that the SMC algorithms have favorable properties in the sparse data regime, but our construction also provides additional justification for the popular variational algorithms in the dense/large data limits.

1. Introduction

In the last decades, several approaches for the analysis of relational data have been proposed. Topic models (Blei, 2012) and nonnegative matrix/tensor decompositions (Cichocki et al., 2009; Kolda and Bader, 2009) are some of the basic tools that have been widely adopted in various application domains as surveyed in (Sidiropoulos et al., 2017; Şimşekli et al., 2015) knowledge bases (Nickel et al., 2016), graph analysis, social networks (Sun et al., 2006; Jeon et al., 2016), music, audio, source separation (Smaragdis and Brown, 2003; Mørup et al., 2006; Şimşekli et al., 2015), link prediction (Ermiş et al., 2015).

In this paper, we introduce a model that we name as the *Bayesian Allocation Model* (BAM) constructed as a dynamical model that enables us to exploit close connections between *topic models*, *nonnegative tensor factorization models* with Kullback-Leibler cost as well as discrete probability models that can be expressed compactly as *Bayesian networks*. Due to space limitations and for keeping the notation simpler, we restrict ourselves here to a particular model structure: the nonnegative PARAFAC (Cichocki et al., 2009), or equivalently, naive Bayes or latent Dirichlet allocation (Blei et al., 2003) for multiway data, but the framework is applicable to more general model topologies.

The key contribution of our approach is a surprisingly simple generic *sequential Monte Carlo algorithm* (Doucet and Johansen, 2009) that exploits and respects the sparsity of observed tensors. The computational complexity of the algorithm scales with the total sum



(a) Allocation Process (b) Naive Bayes (c) Allocation Tensor

of the elements in the observed tensor to be decomposed and does not depend on the size of the observed tensor, unlike optimization based factorization methods. Besides, our discrete construction allows us to use efficient sparse data structures for implementation.

An additional property of our method is that we can calculate the marginal likelihood of any nonnegative tensor model, which is useful for Bayesian model selection. Our method can also be viewed as a model scoring method for Bayesian networks with hidden variables. We illustrate with examples that our algorithm gives promising results as a practical algorithm in the sparse data regime.

2. Bayesian Allocation Model

Let A^t for $t \in [0, 1)$ be a Poisson process defined on the unit interval with a constant intensity λ. An *allocation process* is a collection of Poisson processes that are thinned from the *base process* A^t and indexed as $A^t(r, i_1, \ldots, i_N)$ which we will refer compactly as $A^t(r, i_{1:N})$. The index r has the cardinality of R and each index i_n for $n \in \{1, \ldots, N\}$ has the cardinality of I_n, so we have $R \times \prod_n I_n$ total *subprocesses*. The thinning is done by assigning each *token* (event) to a subprocess $A^t(r, i_{1:N})$ independently with probability $\theta(r, i_{1:N})$. We also define $T \equiv A^{t=1}$ as the total number of tokens generated by the base process and for each $\tau \in \{1, \ldots, T\}$ we define the following *allocation sequence tensors*

$$s^\tau(r, i_{1:N}) = \mathbf{1}_{[\text{token } \tau \text{ is assigned to } A^t(r,i_{1:N})]} \qquad S^\tau(r, i_{1:N}) = \sum\nolimits_{\tau'=1}^{\tau} s^{\tau'}(r, i_{1:N}) \tag{1}$$

where s^τ indicates the assignment of the token τ and S^τ counts the total number assignments to each subprocess up to token τ. At the end, i.e. when T^{th} token is generated, we obtain the *allocation tensor* $S(r, i_{1:N}) \equiv S^T(r, i_{1:N})$. Since the assignments to indices are random, we could view each index as a random variable and θ as a probability distribution on them. Moreover, we could assume conditional independence relations on the indices which are implied by a *Bayesian network* so that θ is factorized with respect to that network. Although any choice of network is equally valid, we will restrict our attention to *Naive Bayes network* (Figure 1(b)) which implies the factorization $\theta(r, i_{1:N}) = \theta_0(r) \prod_{n=1}^{N} \theta_n(r, i_n)$. Naive Bayes also indicates that the index r is latent and only the indices $i_{1:N}$ are observable. Hence, we cannot observe the full S, but only its marginal $X(i_{1:N}) \equiv \sum_r S(r, i_{1:N})$.

To complete the probabilistic description of *Bayesian Allocation Model* (BAM), we define priors on λ and θ_n's in terms of an arbitrary nonnegative tensor $\alpha(r, i_{1:N})$

$$\alpha_+ \equiv \sum\nolimits_{r,i_{1:N}} \alpha(r, i_{1:N}) \quad \alpha_0(r) \equiv \sum\nolimits_{i_{1:N}} \alpha(r, i_{1:N}) \quad \alpha_n(r, i_n) \equiv \sum\nolimits_{i'_{1:N}: i'_n = i_n} \alpha(r, i'_{1:N})$$
$$\lambda \sim \mathcal{G}(\alpha_+, b) \qquad \theta_0 \sim \mathcal{D}(\alpha_0) \qquad \theta_n(r, :) \sim \mathcal{D}(\alpha_n(r, :))$$

This formulation of BAM is closely related to a *Bayesian Poisson CP/PARAFAC decomposition*. To see the relationship, let us define $W_1(r, i_1) \equiv \lambda \theta_0(r) \theta_1(r, i_1)$ and for $n \in \{2, \ldots, N\}$ define $W_n \equiv \theta_n$, then we obtain the following generative model for X:

$$W_1(r, i_1) \sim \mathcal{G}(\alpha_1(r, i_1), b) \quad W_n(r, :) \sim \mathcal{D}(\alpha_n(r, i_n)) \quad X(i_{1:N}) \sim \mathcal{PO}\big(\sum\nolimits_r \prod\nolimits_n W_n(r, i_n)\big)$$

Apart from the selection of the priors, maximum likelihood solution of W_n's corresponds to approximating X by a CP decomposition under KL (Kullback-Leibler) divergence (Cemgil, 2009; Paisley et al., 2014). Likewise, different networks correspond to other factorizations, and to equivalent topic models: while topic models are the models for individual tokens where the nodes indicate relational entities, tensor factorizations view the nodes as indices and model counts of tokens. Moreover, let us define the various marginals of S^τ as follows:

$$S^\tau_+ \equiv \sum\nolimits_{r, i_{1:N}} S^\tau(r, i_{1:N}) \quad S^\tau_0(r) \equiv \sum\nolimits_{i_{1:N}} S^\tau(r, i_{1:N}) \quad S^\tau_n(r, i_n) \equiv \sum\nolimits_{i'_{1:N}: i'_n = i_n} S^\tau(r, i'_{1:N})$$

By integrating out the parameters, we end up with the *marginal allocation model*:

$$p(s^\tau(r, i_{1:N}) = 1 \mid s^{1:\tau-1}) = \frac{\alpha_0(r) + S^{\tau-1}_0(r)}{\alpha_+ + \tau - 1} \prod\nolimits_{n=1}^{N} \frac{\alpha_n(r, i_n) + S^{\tau-1}_n(r, i_n)}{\alpha_0(r) + S^{\tau-1}_0(r)} \tag{2}$$

that describes the marginal distribution $p(s^{1:T})$ as a *Pólya urn* process (Mahmoud, 2008) which we substitute θ_0 with an urn U_0 containing R colors of balls, and each $\theta_n(r, :)$ with the urn U_{nr} containing I_n colors where the initial balls distributed in urns are α_0 and $\alpha_n(r, :)$ respectively. At each step τ, we draw a ball from U_0 and observe its color as r. Then, we draw a ball from the urn U_{nr} to determine i_n for each $n \in \{1, \ldots, N\}$. At the end of the step, each drawn ball is returned in the urn with an additional ball of the same color and $S^\tau(r, i_{1:N})$ is incremented by 1. Since the counts of the additional balls corresponds to marginals of S^τ, sampling is possible by storing only those marginals.

3. Variational Bayes

Although $p(X)$ is intractable, it can be approximated by the *evidence lower bound* (ELBO) $\mathcal{B}(q, X)$ as

$$\log p(X) \geq \langle \log p(X, S, \lambda, \Theta) \rangle_q - \langle \log q(S, \lambda, \Theta) \rangle_q \equiv \mathcal{B}\big(q, X\big)$$

For convenience, we choose q in a form that admits mean-field factorization $q(S, \lambda, \Theta) = q(S)q(\lambda, \Theta)$. Then, maximizing $\mathcal{B}(q, X)$ with respect to $q(S)$ and $q(\lambda, \Theta)$ alternatingly leads to the following set of marginal variational distributions

$$q(\theta_0) = \mathcal{D}(\beta_0) \quad q(\theta_n(r, :)) = \mathcal{D}(\beta_n(r, :)) \quad q(\lambda) = \mathcal{G}(c, d) \quad q(S(:, i_{1:N})) = \mathcal{M}(X(i_{1:N}), \pi(: \mid i_{1:N}))$$

where the variational parameters π, c, d and β_n's are the functions of the sufficient statistics. Then, we can derive a closed form expression for ELBO in terms of variational parameters:

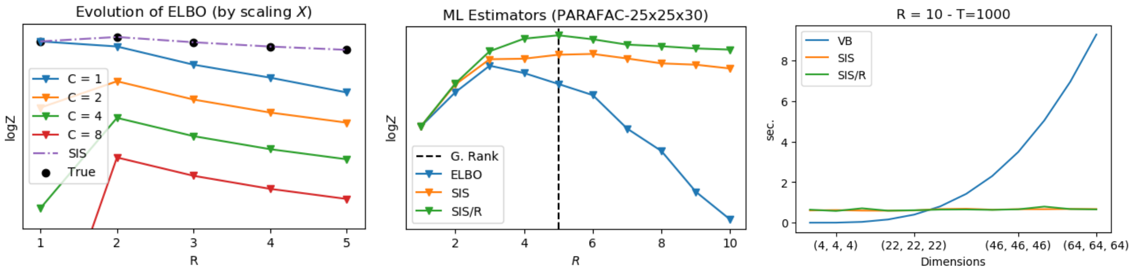
$$e^{\mathcal{B}} = \frac{b^\alpha_+}{d^c} \left(\frac{\prod_r \Gamma(\alpha_0(r))}{\prod_r \Gamma(\beta_0(r))} \right)^{N-1} \left(\prod_{n=1}^{N} \frac{\prod_{r, i_n} \Gamma(\beta_n(r, i_n))}{\prod_{r, i_n} \Gamma(\alpha_n(r, i_n))} \right) \frac{\prod_{r, i_{1:N}} \pi(r \mid i_{1:N})^{-X(i_{1:N})\pi(r \mid i_{1:N})}}{\prod_{i_{1:N}} \Gamma(X(i_{1:N}) + 1)}$$

4. Sequential Importance Sampling

As an alternative to lower bounding the intractable marginal likelihood, we can approximate it via importance sampling (Doucet and Johansen, 2009). Our process interpretation naturally forms a sequence of target distributions $\big\{\varphi_\tau(s^{1:\tau})\big\}_{\tau=1}^{T}$ where each φ_τ is known up to a normalization constant Z_τ with the condition that $Z_T = p(X)$, as follows:

$$X^\tau(i_{1:N}) \equiv \sum\nolimits_r S^\tau(r, i_{1:N}) \qquad \varphi_\tau(s^{1:\tau}) \equiv p(s^{1:\tau} \mid X^\tau) = p(s^{1:\tau})/p(X^\tau) \equiv \gamma(s^{1:\tau})/Z_\tau$$

Then, we select a sequence of importance distributions $\big\{q_\tau(s^{1:\tau})\big\}_{\tau=1}^{T}$ which can be evaluated point-wise and admits the factorization $q_\tau(s^{1:\tau}) = q_\tau(s^\tau \mid s^{1:\tau-1})\, q_{\tau-1}(s^{1:\tau-1})$. Due to this



(a) NMF $(N = 2)$ (b) PARAFAC $(N = 3)$ (c) Comparison of runtimes

factorization, sampling $s^{1:T} \sim q_T$ can be done sequentially via sampling s^1 from $q_1(s^1)$ and remaining s^τ's from the distributions $q_\tau(s^\tau \mid s^{1:\tau-1})$. Our particular choice of importance distributions and corresponding unnormalized weights are

$$\begin{aligned} q_\tau(s^\tau \mid s^{1:\tau-1}) &\equiv p(s^\tau(r, i_{1:N}) = 1 \mid s^{1:\tau-1}, x^\tau(i_{1:N}) = 1) \propto \frac{\alpha_0(r) + S_0^{\tau-1}(r)}{\alpha_+ + \tau - 1} \prod_{n=1}^N \frac{\alpha_n(r, i_n) + S_n^{\tau-1}(r, i_n)}{\alpha_0(r) + S_0^{\tau-1}(r)} \\ w_\tau(s^{1:\tau}) &\equiv \frac{\gamma_\tau(s^{1:\tau})}{q_\tau(s^{1:\tau})} = \frac{\gamma_{\tau-1}(s^{1:\tau-1})}{q_{\tau-1}(s^{1:\tau-1})} \frac{\gamma_\tau(s^{1:\tau})}{\gamma_{\tau-1}(s^{1:\tau-1})\, q_\tau(s^\tau \mid s^{1:\tau-1})} = w_{\tau-1}(s^{1:\tau-1})\, \nu_\tau(s^{1:\tau}) \end{aligned}$$

where ν_τ's are the *incremental importance weights*. Notice that $p(X)$ is the expectation of w_T under density q_T, hence we can sample $s^{1:T}$ paths from q_T, calculate their weights, and estimate $p(X)$ by averaging them. To calculate weights of samples $s^{1:T}$ dynamically which is also crucial for resampling, we derive an expression for incremental importance weights:

$$\nu_\tau(s^{1:\tau}) = \left(\sum_r \frac{\alpha_0(r) + S_0^{\tau-1}(r)}{\alpha_+ + \tau - 1} \prod_{n=1}^N \frac{\alpha_n(r, i_n) + S_n^{\tau-1}(r, i_n)}{\alpha_0(r) + S_0^{\tau-1}(r)} \right) \prod_{i_{1:N}} \left(\frac{\tau}{X^\tau(i_{1:N})} \right)^{x^\tau(i_{1:N})}$$

Note that SMC calculates an unbiased estimator of $p(X)$ for any sequence $X^{1:T}$, but the chosen $X^{1:T}$ is the parameter of q and affects the variance of the estimators. So, it is advisable to also sample $X^{1:T}$ from backward Pólya urn, *i.e.* sampling $X^{T:1}$ without replacement.

5. Preliminary Results and Conclusion

To illustrate their contrasting nature, we tested VB and SMC on three synthetic examples: (i) On a 3×4 matrix with $T = 9$, we enumerated all possible S's exhaustively to calculate true $p(X \mid R)$ by (Eq. 2) and found the "rank" $R = 2$ is the most likely. To show the transformation of ELBO as T increases, we scaled X with integer C. Figure 2(a) shows that ELBO is not accurate for original X, but as we double T its trend starts to resemble true $p(X)$ whereas SIS estimates $p(X)$ with negligible errors in sparse data regime.
(ii) On a $25 \times 25 \times 30$ tensor generated with $R = 5$ and $T = 500$. Figure 2(b) shows that SIS with adaptive resampling (SIS/R) estimates maximum $p(X)$ at true R whereas naive SIS and VB seem deviated below at higher R.
(iii) Figure 2(c) demonstrates that the runtime of VB scales with the size of X whereas runtime of SMC algorithms (without parallelization) is independent from it, making SMC algorithms a promising choice in the sparse data regime.

Bayesian Allocation Model highlights the connections of nonnegative tensor factorizations and discrete Bayes networks, justifies VB in dense data regime by bringing an alternative perspective to sample size (Wang and Blei, 2018), explains the "by parts representation" nature (Donoho and Stodden, 2004) of NMF/NTF via self-reinforcing Pólya urns. It also offers an SMC algorithm that scales with sum but not the observed size of the tensor, hence practical for sparse tensors with large dimensions, and provides a model scoring method.

Acknowledgments

The work was supported by the French National Research Agency grant ANR-16-CE23-0014 (FBIMATRIX).

References

David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, April 2012. ISSN 0001-0782. doi: 10.1145/2133806.2133826. URL http://doi.acm.org/10.1145/2133806.2133826.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=944919.944937.

Ali Taylan Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009, 2009.

Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun-ichi Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in neural information processing systems*, pages 1141–1148, 2004.

Arnaud Doucet and Adam M Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704):3, 2009.

Beyza Ermiş, Evrim Acar, and A Taylan Cemgil. Link prediction in heterogeneous data via generalized coupled tensor factorization. *Data Mining and Knowledge Discovery*, 29 (1):203–236, 2015.

Inah Jeon, Evangelos E Papalexakis, Christos Faloutsos, Lee Sael, and U Kang. Mining billion-scale tensors: algorithms and discoveries. *The VLDB JournalThe International Journal on Very Large Data Bases*, 25(4):519–544, 2016.

Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

Hosam Mahmoud. *Polya Urn Models*. Chapman & Hall/CRC, 1 edition, 2008. ISBN 1420059831, 9781420059830.

Morten Mørup, Lars Kai Hansen, Josef Parnas, and Sidse M Arnfred. Decomposing the time-frequency representation of eeg using non-negative matrix and multi-way factorization. *Technical University of Denmark Technical Report*, 2006.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2016.

J. Paisley, D. Blei, and M. I. Jordan. chapter Bayesian Nonnegative Matrix Factorization with Stochastic Variational Inference, pages 205–224. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Chapman and Hall/CRC, Oct 2014. ISBN 978-1-4665-0408-0. doi: 10.1201/b17520-15. URL https://doi.org/10.1201/b17520-15. 0.

Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, 2017.

Umut Şimşekli, Tuomas Virtanen, and Ali Taylan Cemgil. Non-negative tensor factorization models for bayesian audio processing. *Digital Signal Processing*, 47:178–191, 2015.

Paris Smaragdis and Judith C Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE workshop on applications of signal processing to audio and acoustics*, volume 3, pages 177–180. New York, 2003.

Jimeng Sun, Dacheng Tao, and Christos Faloutsos. Beyond streams and graphs: dynamic tensor analysis. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 374–383. ACM, 2006.

Yixin Wang and David M Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, (just-accepted):1–85, 2018.