Bayesian Nonnegative Matrix Factorization as an Allocation Model

M. Burak Kurutmaz[1], A. Taylan Cemgil[1], Umut Şimşekli[2], Sinan Yıldırım[3]
1: Boğaziçi University, İstanbul, Turkey
2: LTCI, Télécom ParisTech, Université Paris-Saclay, Paris, France
3: Sabancı University, İstanbul, Turkey

Abstract

We propose a Dirichlet Gamma prior for Bayesian Nonnegative Matrix Factorization (NMF) with the Kullback-Leibler divergence, also known as Bayesian Poisson Factorization and show that both latent factor matrices can be integrated out analytically. For count matrices, this leads to a tensor valued discrete distribution that we call an allocation model. The properties of the allocation model provides an alternative perspective on the empirical behavior of Bayesian NMF such as by-parts representation, and connections to Latent Dirichlet Allocation (LDA). It allows us to formulate alternative novel decompositions for count data beyond low rank approximations, that we name as Best Latent Decomposition (BLD). We develop a multiplicative gradient ascent algorithm for approximately solving the BLD problem and show that these solutions have favorable properties.

1 Introduction

The Nonnegative Matrix Factorization (NMF) problem is typically cast as a minimization problem:

$$W^*, H^* \quad = \quad \arg\min_{W,H\geq 0} D(X||WH) \tag{1}$$

where X is an observed matrix and W and H are element-wise nonnegative factor matrices such that the divergence D is minimized [1]. When D is selected as the information divergence, $D(p||q) = \sum_i p_i \log p_i - p_i \log(q_i) - p_i + q_i$, we obtain the Kullback-Leibler (KL)-NMF. When X is a count matrix, i.e., entries X_{ij} are nonnegative integers, the model may also be referred as the Poisson factorization model as the negative log-likelihood of the Poisson intensity $\Lambda = WH$ is, up to constant terms, equal to the information divergence. More precisely, if

$$X_{ij}|W,H \sim \mathcal{PO}(X_{ij};\Lambda_{ij}) \qquad\qquad \Lambda_{ij} = \sum_k W_{ik}H_{kj}, \tag{2}$$

the log-likelihood is $\mathcal{L}_X(W,H) = -D(X||WH) + \text{const}$. Using the technique of *data augmentation*, we can introduce a new set of latent variables that can be organized as a tensor S where each entry of the tensor is conditionally Poisson distributed with

$$S_{ijk}|W,H \sim \mathcal{PO}(W_{ik}H_{kj}) \qquad\qquad X_{ij} = \sum_k S_{ijk} \equiv S_{ij+}. \tag{3}$$

Thanks to the superposition property of the Poisson distribution [2], the marginal model is exactly equal to (2). The hidden tensor S can be interpreted as a latent decomposition of the observed count matrix X, the sum of slices $S_{::k}$. One advantage of introducing the hidden tensor S is that one can use standard statistical data augmentation procedures such as Expectation-Maximization (EM) algorithm for maximum likelihood parameter estimation or the Gibbs sampler for directly sampling from the posterior distribution $p(W,H|X)$, provided a suitable prior $p(W,H)$ [3]. For example, the popular multiplicative updates of KL-NMF [1] can be shown to be an EM algorithm for maximizing the log-likelihood $\log p(X|W,H)$ in W,H with S as the latent variables [4].

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

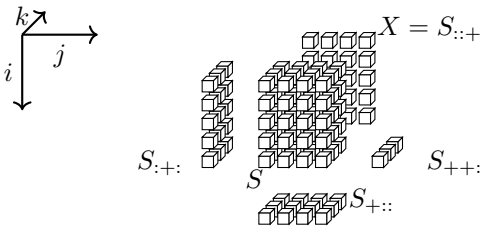

Figure 1: Allocation model of Bayesian KL-NMF. Each cell of the $I \times J \times K$ tensor S contains S_{ijk} balls and each fiber $S_{ij:}$ has a total of $X_{ij} = S_{ij+}$ balls. We view S as a counting measure with well defined marginals and associated entropies. To maximize $p(S)$, the marginals $S_{+::}$ and $S_{:+:}$ (with elements S_{+jk} and S_{i+k} respectively) should have low entropy, whereas $S_{++:}$ and S (with elements S_{++k} and S_{ijk}) should have a high entropy.

2 Bayesian NMF as an Allocation Model

The probabilistic interpretation of KL-NMF leads naturally to a variety hierarchical models, by different prior choices. One obvious choice is independent Gamma priors of form $W_{ik} \sim \mathcal{G}(\cdot)$ and $H_{ik} \sim \mathcal{G}(\cdot)$ [4, 5], as these are conjugate to a Poisson likelihood. However, this symmetric formulation inherits a drawback. In the augmented model, the conditional density $p(W, H|S) = p(W|H, S)p(H|S)$ is not available in a convenient closed form and further approximations are needed. As revealed by the analysis provided in the supplementary [6], this problem is due to the scaling redundancy in the model, also present in the original formulation of NMF [1]: For any scalar $\kappa \neq 0$, $WH = (\kappa W)(H/\kappa)$. To avoid the scaling redundancy, we will parameterize the intensity matrix Λ as

$$\Lambda_{ij} = L_j \sum_k W_{ik} H_{kj} = L_j (WH)_{ij}$$

and will assume that columns of W and H are normalized – in fact NMF is typically used in practice with this normalization. We will assume the observation model in (3) with the following prior model

$$L_j \sim \mathcal{G}(a, b) \qquad W_{:k} \sim \mathcal{D}(\alpha) \qquad H_{:j} \sim \mathcal{D}(\beta) \tag{4}$$

where $\mathcal{D}(\alpha)$ is the Dirichlet distribution with parameter vector α and we use the APL/Matlab notation ":" to denote an entire vector over a given index. In this construction, one could think of WH as a conditional probability table: a matrix where each column is nonnegative and is normalized. To arrive at a nonnegative intensity matrix we scale each column by a positive scalar L_j. With this new formulation, S_{ijk} is conditionally distributed with $\mathcal{PO}(W_{ik}H_{kj}L_j)$.

The advantage of the proposed formulation hinges on a simple observation: since the columns of W and H sum up to 1 by the property of the Dirichlet distribution, $\sum_i \sum_k W_{i,k}H_{k,j} = 1$ for all j.

This simple observation reveals that the cross terms cancel in the Poisson likelihood and that it is possible to obtain a factorization of the posterior, which is not obvious from the corresponding hierarchical model. We can show that the joint distribution admits the following factorization:

$$p(S, W, H, L) = \Big(\prod\nolimits_k p(W_{:,k} \mid S)\Big)\Big(\prod\nolimits_j p(H_{:,j} \mid S)\Big)\Big(\prod\nolimits_j p(L_j \mid S)\Big)p(S), \tag{5}$$

$$p(L_j \mid S) = \mathcal{G}\left(a + S_{+j+}, b + 1\right), p(W_{:,k} \mid S) = \mathcal{D}\left(\alpha + S_{:+k}\right), p(H_{:,j} \mid S) = \mathcal{D}\left(\beta + S_{+j:}\right).$$

More importantly, the marginal $p(S)$, that we call an allocation model, has also a closed form as

$$p(S) \quad = \quad C_X \exp(\ell(S)) \tag{6}$$

$$C_X \quad = \quad \left(\frac{\Gamma(\alpha_+)}{\prod_i \Gamma(\alpha_i)}\right)^K \left(\frac{\Gamma(\beta_+)}{\prod_k \Gamma(\beta_k)} \frac{b^a}{\Gamma(a)}\right)^J (b+1)^{-(aJ+S_{+++})} \prod\nolimits_j \frac{\Gamma(a + S_{+j+})}{\Gamma(\beta_+ + S_{+j+})}$$

$$\ell(S) \quad = \quad \sum\nolimits_k \sum\nolimits_i \log\Gamma(\alpha_i + S_{i+k}) + \sum\nolimits_j \sum\nolimits_k \log\Gamma(\beta_k + S_{+jk}) \tag{7}$$

$$- \sum\nolimits_k \log\Gamma(\alpha_+ + S_{++k}) - \sum\nolimits_i \sum\nolimits_j \sum\nolimits_k \log\Gamma(S_{ijk} + 1) \tag{8}$$

Here $\Gamma(x)$ is the gamma function [7] and $\psi(x) = d \log \Gamma(x)/dx$ is the digamma function [8]. The detailed derivation is given in the supplementary repository [6]. We define various marginals as $X_{ij} = \sum_k S_{ijk} \equiv S_{ij+}$, $S_{+j+} \equiv \sum_{ik} S_{ijk}$, $S_{+++} \equiv \sum_{ijk} S_{ijk}$, $S_{i+k} \equiv \sum_j S_{ijk}$, $S_{+jk} \equiv \sum_i S_{ijk}$, $S_{++k} \equiv \sum_{ij} S_{ijk}$, $\alpha_+ \equiv \sum_i \alpha_i$, $\beta_+ \equiv \sum_k \beta_k$.

2.1 Interpretation of the proposed model

A close look at the allocation model $p(S)$ as a prior distribution on possible latent decompositions of X reveals an alternative perspective about the general nature of the Bayesian KL-NMF model. We imagine a physical system, an allocation model, where a total of S_{+++} balls are placed into $I \times J \times K$ bins. As $S_{ij+} = X_{ij}$, there must be exactly X_{ij} balls in the fiber $(ij1), (ij2) \dots (ijK)$, that we will refer as $(ij:)$.

The C_X term depends only on S_{+++} and S_{+j+}, which are both fixed given X, and therefore C_X is constant when X is given. The second term, $\ell(S)$, consists of two concave and two convex terms, (7) and (8) respectively, all having a simple entropy interpretation by Stirling's formula $\log \Gamma(n+1) = n \log n - n + O(\log n)$. Terms such as $-\sum_{i=1}^N \log \Gamma(s_i) \approx -\sum_{i=1}^N s_i \log s_i$ where $\sum_i s_i = \text{const}$ can be interpreted as an entropy of a mass function defined on N cells where s_i is the mass allocated to cell i. Hence, a high probability decomposition S is forced by the terms in (7) to have its marginal sums S_{i+k} and S_{+jk} as concentrated as possible to a few cells while the terms in (8) force S_{++k} and S_{ijk} to be distributed as evenly as possible. Intuitively, the form of the objective enforces a by-parts representation or sparse factor matrices – this property is hidden in the matrix factorization interpretation arguably, but is more obvious in the allocation model due to the entropy interpretation.

Interestingly, our simple observation will carry forth to other popular topic models, such as probabilistic latent semantic analysis (PLSA) and LDA [9]. In fact, a special case of (6) is the target distribution for the collapsed Gibbs sampler for LDA [10], while the fact that it extends to NMF is in our knowledge not noted in the rich literature that mentions links between various related models [5, 11–13]. Space constraints do not allow us to show the details of the fact that LDA and Bayesian KL-NMF are "almost equivalent" allocation models: while LDA defines the generative model for the individual balls, Bayesian KL-NMF defines a model on their counts.

3 Best Latent Decomposition Problem

The Best Latent Decomposition (BLD) problem is finding the mode of $p(S \mid X)$

$$S^* = \arg\max_{S_{::+}=X} p(S) = \arg\max_{S_{::+}=X} \ell(S) \quad (9)$$

$$W^*, H^*, L^* = \arg\max_{W,H,L} p(W \mid S^*) p(H \mid S^*) p(L \mid S^*) \quad (10)$$

Note that, once S^* is found, the conditional distributions of factor matrices are available in closed form. It may appear first that this formulation has a higher space complexity, as the number of optimization variables is $I \times J \times K$ in contrast to $I \times K + K \times J$ of the original KL-NMF, but it turns out that an efficient algorithm exists. We first relax the condition that S is a nonnegative integer tensor by extending its domain to $\Omega_X = \{S \in {\mathbb{R}_+}^{I \times J \times K} \mid S_{::+} = X\}$. We arrive at the following relaxed optimization problem

$$S^* = \arg\max_{S \in \Omega_X} \ell(S) \quad (11)$$

Then, it can be shown that by the method of Lagrange multipliers [14], the solution of S^* is in the form of

$$\psi(S^*_{ijk} + 1) = \log(\lambda_{ij} \nu_{ik} \mu_{kj}) \quad (12)$$

where $\lambda_{ij}, \nu_{ik}, \mu_{kj}$ are nonnegative Lagrange multipliers. As $\psi(S_{ijk}+1)$ can be well approximated by $\log S_{ijk}$ in (12) for large S_{ijk}, we will now seek a solution in the form of $S_{ijk} = \lambda_{ij} \nu_{ik} \mu_{kj}$. Since we are restricting our attention to a subset of Ω_X, optimal S having this form will give us a lower bound. By employing the constraint $X = S_{::+}$, λ can be explicitly obtained and accordingly we can further simplify the form of S to the following expression:

$$S_{ijk} = X_{ij} \frac{\nu_{ik} \mu_{kj}}{\sum_c \nu_{ic} \mu_{cj}} \quad (13)$$

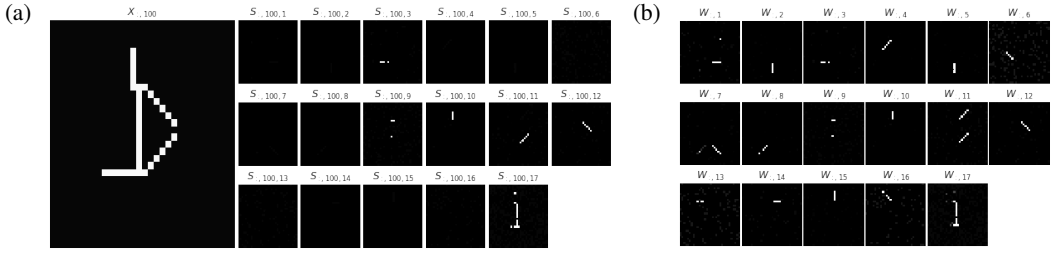

Figure 2: Empirical results on the Swimmer dataset. (a) The observed image $X_{:j}$ and the slices $S_{:jk}$ for $k = 1, \ldots, K$. (b) Columns of the estimated W.

This formulation lets us perform optimization by only using the optimization variables μ and ν, without explicitly storing $I \times J \times K$ tensor S. When the hyper-parameters $\beta_k \geq 1$, gradient of $\ell(S)$ w.r.t. ν and μ can be partitioned into sum of positive and negative terms:

$$\frac{\partial \ell(S)}{\partial \nu_{ik}} = \sum_j \frac{X_{ij}\mu_{kj}}{\sum_c \nu_{ic}\mu_{cj}} \nabla L^+_{ijk} + \sum_j \sum_{k'} X_{ij}\mu_{kj} \frac{\nu_{ik'}\mu_{k'j}}{(\sum_c \nu_{ic}\mu_{cj})^2} \nabla L^-_{ijk'} \tag{14}$$

$$- \Big(\sum_j \frac{X_{ij}\mu_{kj}}{\sum_c \nu_{ic}\mu_{cj}} \nabla L^-_{ijk} + \sum_j \sum_{k'} X_{ij}\mu_{kj} \frac{\nu_{ik'}\mu_{k'j}}{(\sum_c \nu_{ic}\mu_{cj})^2} \nabla L^+_{ijk'} \Big) \tag{15}$$

where $\nabla L^+_{ijk'} = \psi(\beta_{k'} + S_{+jk'}) - \psi(S_{ijk'} + 1)$ and $\nabla L^-_{ijk'} = \psi(\alpha_+ + S_{++k'}) - \psi(\alpha_i + S_{i+k'})$, and they are nonnegative by the monotonicity of ψ. If we define the $\nabla \ell^+_{\nu_{ik}}, \nabla \ell^-_{\nu_{ik}}$ as the terms in (14), (15) respectively, and similarly define $\nabla \ell^+_{\mu_{kj}}, \nabla \ell^-_{\mu_{kj}}$ for the gradient w.r.t. μ_{kj}, they are also nonnegative and the following multiplicative gradient ascent algorithm similar to the one in [15] can be employed: $\nu_{ik} \leftarrow \nu_{ik} \nabla \ell^+_{\nu_{ik}} / \nabla \ell^-_{\nu_{ik}}$ and $\mu_{kj} \leftarrow \mu_{kj} \nabla \ell^+_{\mu_{kj}} / \nabla \ell^-_{\mu_{kj}}$.

4 Preliminary Results and Conclusion

We challenged our model on the `Swimmer` dataset that contains 256 images of size 32×32 [16]. Even though this dataset might seem simple at first sight, it is a specifically constructed case which KL-NMF fails to obtain 'by-parts' representations [16]. We compare the proposed algorithm to the original KL-NMF algorithm [1]. However, here we can only report a subset of our results. The detailed results of all our experiments along with our mathematical derivations are given in our supplementary repository [6].

Various settings for the hyper-parameters result in significantly different semantic representations. In Fig. 2, a particularly interesting case where we set $K = 17$, $a = 100$, $b = 1$, $\alpha_i = 0.05$, $\beta_k = 10$ for $k = 1, \ldots K-1$ and $\beta_K = 60$. This particular choice of α encourages the model to find a sparse representation. Similarly this choice of β_k with $k = 1, \ldots, K-1$ distributes the body parts to different slices evenly, whereas the large value of β_K encourages the model to find a 'common part' (in this case the torso of the swimmer) that appears in all columns of X.

In Fig. 2a, we illustrate the slices $S_{:jk}$, for $k = 1, \ldots, K$, corresponding to the decomposition of fixed column of $X_{:j}$ with $j = 100$ (i.e. 100th image in the dataset). Similarly, Fig. 2b illustrates the optimal W that is obtained by using the estimated S (see (5)). The results show that the proposed approach is able to obtain a very sparse by-parts representation. Here, each slice of $S_{:j:}$ corresponds to a different body part for a given image, whereas each column of W encodes the body-parts for the whole dataset. On the contrary, [16] has shown that the original KL-NMF fails to find a by-parts representation and generates ghosts of the torso that is contaminated to all the slices (see Fig. 2 in [16]).

We believe that the allocation model perspective of Bayesian KL-NMF is fruitful for developing alternative inference algorithms as well as constructing flexible and interpretable models that are needed in many applications. A natural next step in this direction is the computation of the marginal likelihood $p(X)$ via sequential algorithms.

Acknowledgments

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

References

[1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[2] J. F. C. Kingman, *Poisson processes*. Wiley Online Library, 1993.

[3] C. Fevotte and A. T. Cemgil, "Nonnegative matrix factorisations as probabilistic inference in composite models," in *Proc. 17th European Signal Processing Conference (EUSIPCO'09)*, Glasgow, 2009.

[4] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, vol. 2009, 2009.

[5] J. Paisley, D. Blei, and M. I. Jordan, ser. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Chapman and Hall/CRC, Oct 2014, ch. Bayesian Nonnegative Matrix Factorization with Stochastic Variational Inference, pp. 205–224, 0. [Online]. Available: https://doi.org/10.1201/b17520-15

[6] M. B. Kurutmaz, A. T. Cemgil, U. Şimşekli, and S. Yıldırım, "Bayesian nonnegative matrix factorization as an allocation model, supplementary material," https://github.com/mehmetburakkurutmaz/AABI17_allocation_models.

[7] P. J. Davis, "Gamma function and related functions," *Handbook of mathematical functions*, pp. 253–293, 1972.

[8] J. M. Bernardo, "Algorithm as 103: Psi (digamma) function," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 25, no. 3, pp. 315–317, 1976.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944937

[10] Y. W. Teh, D. Newman, and M. Welling, "A collapsed variational Bayesian inference algorithm for Latent Dirichlet Allocation," in *Advances in neural information processing systems*, 2007, pp. 1353–1360.

[11] E. Gaussier and C. Goutte, "Relation between plsa and nmf and implications," in *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2005, pp. 601–602. [Online]. Available: http://portal.acm.org/citation.cfm?id=1076034.1076148

[12] J. Canny, "Gap: A factor model for discrete data," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '04. New York, NY, USA: ACM, 2004, pp. 122–129. [Online]. Available: http://doi.acm.org/10.1145/1008992.1009016

[13] T. P. Faleiros and A. A. Lopes, "On the equivalence between algorithms for non-negative matrix factorization and latent dirichlet allocation," in *ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Bruges (Belgium), 2016, pp. 171–176.

[14] D. P. Bertsekas, *Nonlinear programming*. Athena scientific Belmont, 1999.

[15] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[16] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" in *Advances in neural information processing systems*, 2004, pp. 1141–1148.