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

Introduction & Motivation

- **KL-NMF** problem is typically cast as a minimization problem:

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

where X is an observation matrix and W and H are **element-wise nonnegative** factor matrices [1].

- It can be shown that multiplicative update rules of **KL-NMF** is an **EM** algorithm [2] by introducing a latent tensor S having each entry is conditionally **Poisson** distributed with

$$S_{ijk}|W,H \sim \mathcal{PO}(W_{ik}H_{kj}) \qquad X_{ij} = \sum_k S_{ijk} \equiv S_{ij+} \quad (2)$$

- The probabilistic interpretation of KL-NMF leads to a variety of hierarchical models. One obvious choice is independent **Gamma priors** of form $W_{ik} \sim \mathcal{G}(\cdot)$ and $H_{ik} \sim \mathcal{G}(\cdot)$ [2].
- However, these formulations present **scaling redundancy**, i.e. for any positive scalar κ,

$$WH = (\kappa W)(H/\kappa) \quad (3)$$

- The conditional density $p(W,H|S) = p(W|H,S)p(H|S)$ is not available in a convenient closed form.
- NMF is typically used in practice with **normalization**.

Bayesian NMF as an Allocation Model

To avoid the **scaling redundancy**, we propose the following model:

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

$$S_{ijk} \mid W,H,L \sim \mathcal{PO}(W_{ik}H_{kj}L_j) \qquad X_{ij} = \sum_k S_{ijk} \equiv S_{ij+} \quad (5)$$

The **advantage** of the proposed formulation is that the **joint distribution** admits the following factorization:

$$p(S,W,H,L) = \Big(\prod_k \underbrace{p(W_{:,k}|S)}_{\text{Dirichlet}}\Big)\Big(\prod_j \underbrace{p(H_{:,j}|S)}_{\text{Dirichlet}}\Big)\Big(\prod_j \underbrace{p(L_j|S)}_{\text{Gamma}}\Big)p(S) \quad (6)$$

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

$$p(S) = C_X \exp(\ell(S)) \quad (7)$$

$$\ell(S) = \underbrace{\sum_k\sum_i \log\Gamma(\alpha_i + S_{i+k}) + \sum_j\sum_k \log\Gamma(\beta_k + S_{+jk})}_{\text{concave}} \quad (8)$$

$$\underbrace{-\sum_k \log\Gamma(\alpha_+ + S_{++k}) - \sum_i\sum_j\sum_k \log\Gamma(S_{ijk}+1)}_{\text{convex}} \quad (9)$$

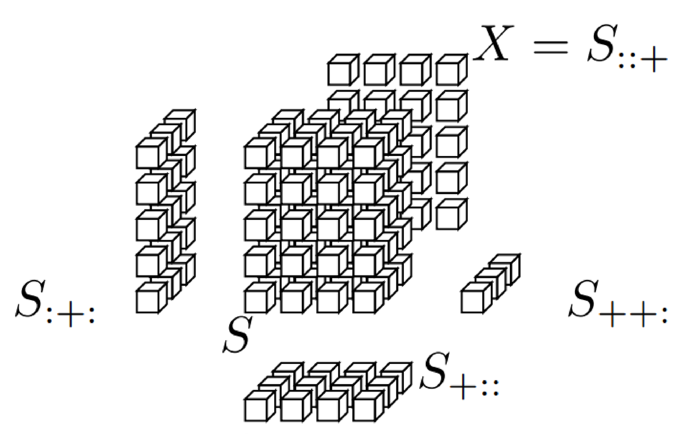
$$C_X = \left(\frac{\Gamma(\alpha_+)}{\prod_i \Gamma(\alpha_i)}\right)^K \left(\frac{\Gamma(\beta_+)}{\prod_k \Gamma(\beta_k)}\frac{b^a}{\Gamma(a)}\right)^J (b+1)^{-(aJ+S_{+++})}\prod_j \frac{\Gamma(a+S_{+j+})}{\Gamma(\beta_+ + S_{+j+})}$$

Interpretation of the Proposed Model

- $\ell(S)$ consists of concave and convex terms all having a simple **entropy** interpretation by **Stirling's approximation**:

$$-\sum_i \log\Gamma(s_i) \approx -\sum_i s_i \log s_i$$

where $\sum_i s_i = const.$



- S is forced by the **concave** terms to have S_{i+k} and S_{+jk} as **concentrated** as possible to a few cells while **convex** terms 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.
- Equivalance with **PLSA** and **LDA** [3].

References

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

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

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[4] 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.

Best Latent Decomposition (BLD)

- **BLD** problem is finding the mode of $p(S \mid X)$:

$$S^* = \arg\max_{S_{::+}=X} p(S) \quad (10)$$

$$W^*, H^*, L^* = \arg\max_{W,H,L} p(W \mid S^*)p(H \mid S^*)p(L \mid S^*) \quad (11)$$

- **Relaxation**: S is a nonnegative integer tensor, but if we extend its domain to $\Omega_X = \{S \in \mathbb{R}_+^{I\times J\times K} \mid S_{::+} = X\}$, it can be shown that by **Lagrange multipliers** method, 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.

Approximate BLD Algorithm

- By the inspiration of the form in (12), seek a solution in the form of

$$S_{ijk} = \lambda_{ij}\nu_{ik}\mu_{kj} \quad (13)$$

- Since we are restricting our attention to a subset of Ω_X, optimal S having this form will give us a **lower bound**.
- Employing the constraint $X = S_{::+}$ simplifies the form of S:

$$S_{ijk} = X_{ij}\frac{\nu_{ik}\mu_{kj}}{\sum_c \nu_{ic}\mu_{cj}} \quad (14)$$

- Gradients of $\ell(S)$ are the difference of two nonnegative matrices:

$$\nabla\ell_{\nu_{ik}} = \nabla\ell^+_{\nu_{ik}} - \nabla\ell^-_{\nu_{ik}} \qquad \nabla\ell_{\mu_{kj}} = \nabla\ell^+_{\mu_{kj}} - \nabla\ell^-_{\mu_{kj}} \quad (15)$$

- **Multiplicative Updates**

$$\nu_{ik} \leftarrow \nu_{ik}\nabla\ell^+_{\nu_{ik}}/\nabla\ell^-_{\nu_{ik}} \qquad \mu_{kj} \leftarrow \mu_{kj}\nabla\ell^+_{\mu_{kj}}/\nabla\ell^-_{\mu_{kj}} \quad (16)$$

Experiments

- **Synthetic Data Experiments**

	Sparseness	$\log p(S)$
KL-NMF	0.45	-239.2
BLD-NMF	**0.50**	**-229.5**

Table 1: Average results on 1000 randomly generated 8×10 matrices. ($K = 2$, $a = 40$, $b = 1$, $\alpha_i = 1$, $\beta_k = 1$)

- **Swimmer Dataset** [4]

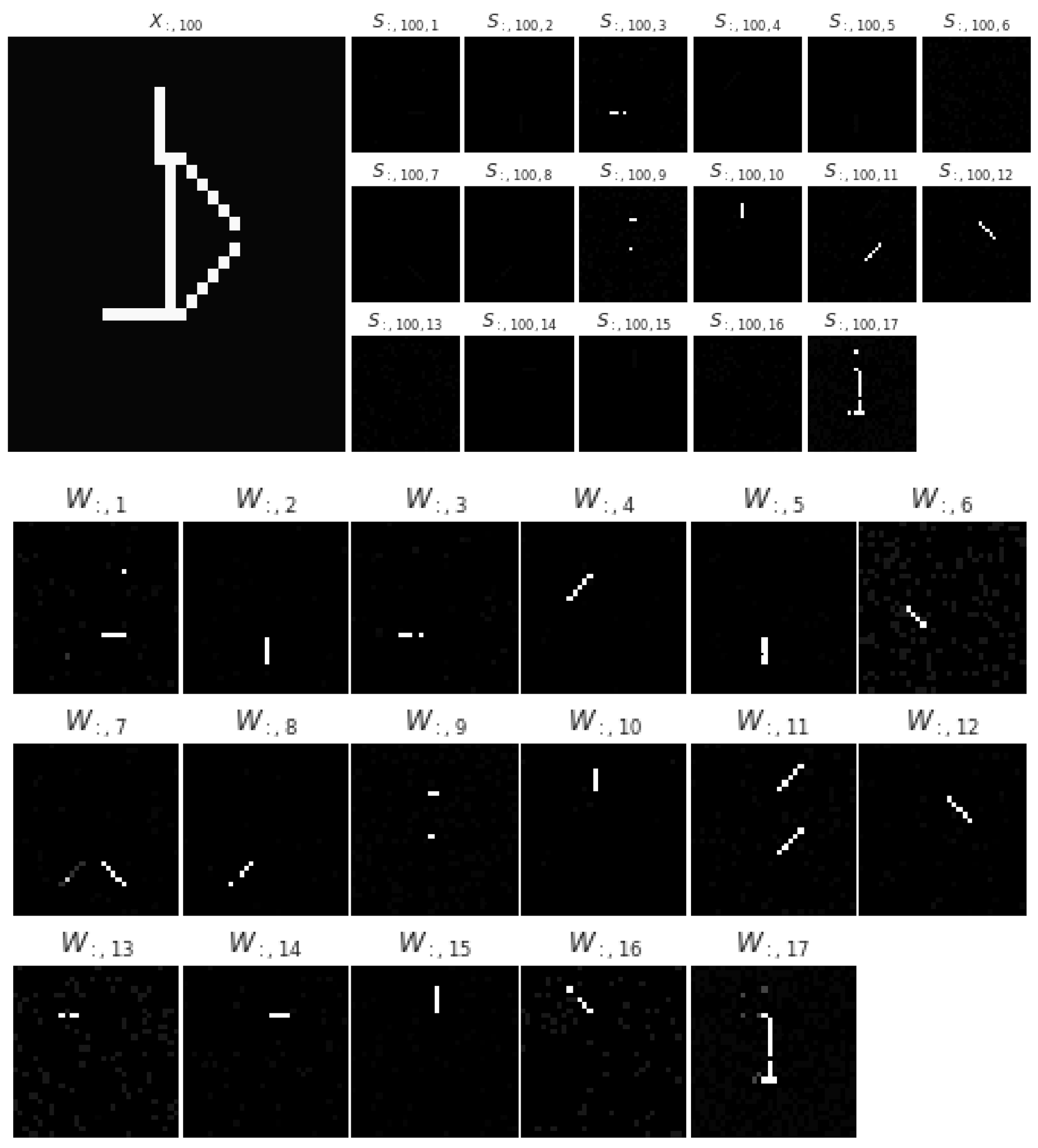


Figure 1: (*Top*) Hidden slices of the estimated S tensor for the 100^{th} sample. (*Bottom*) Columns of estimated W, **hidden representations**. ($K = 17$, $a = 100$, $b = 1$, $\alpha_i = 0.05$, $\beta_k = 10$ for $k = 1, \ldots, K-1$ and $\beta_K = 60$)

- The results show that the proposed approach is able to obtain a **very sparse by-parts** representation.

Conclusion

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.