Diversified Bayesian Nonnegative Matrix Factorization

Maoying Qiao,[1]* Jun Yu,[2] Tongliang Liu,[3] Xinchao Wang,[4] Dacheng Tao[3]
[1]The Commonwealth Scientific and Industrial Research Organisation, Australia
[2]Hangzhou Dianzi University, Hangzhou, China
[3]UBTECH Sydney AI Centre, School of Computer Science,
Faculty of Engineering, The University of Sydney, Darlington, NSW 2008, Australia
[4]Stevens Institute of Technology, Hoboken, New Jersey 07030
maoying.qiao@csiro.au, yujun@hdu.edu.cn, {tongliang.liu, dacheng.tao}@sydney.au.,
xinchao.wang@stevens.edu

Abstract

Nonnegative matrix factorization (NMF) has been widely employed in a variety of scenarios due to its capability of inducing semantic part-based representation. However, because of the non-convexity of its objective, the factorization is generally not unique and may inaccurately discover intrinsic "parts" from the data. In this paper, we approach this issue using a Bayesian framework. We propose to assign a diversity prior to the parts of the factorization to induce correctness based on the assumption that useful parts should be distinct and thus well-spread. A Bayesian framework including this diversity prior is then established. This framework aims at inducing factorizations embracing both good data fitness from maximizing likelihood and large separability from the diversity prior. Specifically, the diversity prior is formulated with determinantal point processes (DPP) and is seamlessly embedded into a Bayesian NMF framework. To carry out the inference, a Monte Carlo Markov Chain (MCMC) based procedure is derived. Experiments conducted on a synthetic dataset and a real-world MULAN dataset for multi-label learning (MLL) task demonstrate the superiority of the proposed method.

Introduction

Nonnegative matrix factorization (NMF) has attracted attention due to its non-negativity constraints. These constraints induce non-subtractive part-based representations to effectively interpret data (Lee and Seung 1999). For example, in the multi-label learning task, NMF factorizes an image dataset X into shared image parts as bases W and the corresponding individual constituent weights as new image representations H. Ideally, when the shared image parts in W are associated with distinct labels, the constituent weights H based on these parts then encode label information and help improve the accuracies of the following classification task. With this promising prospect, many efforts have been dedicated to effectively discovering meaningful parts from the data in a vast number of application scenarios. Examples include document clustering based on topic discovery, hyperspectral unmixing, audio source separating, music analysis, and community detection.

Studies have been conducted to seek unique and exact solutions to NMF, despite the non-convexity nature of the problem. Most of these studies are based on the separability assumption (Chen et al. 2019; Degleris and Gillis 2019). This condition states that the columns of the bases W, which should be a subset of dataset X, i.e., $W \subseteq X$, span a convex hull/simplex/conical hull/cone which includes all data points X (Zhou, Bian, and Tao 2013). However, this condition is too strict to be satisfied in practice. Conditions focusing on relaxing the separability assumption have been developed, such as a near-separable condition, a subset-separable condition (Ge and Zou 2015), and a geometric assumption (Liu, Gong, and Tao 2017). However, all these exact solutions do not consider the low-rank condition of W, which is practically important when NMF plays the role as a dimensionality reduction tool.

In practice, approximate solutions with good generalizability are desired. To seek such solutions to NMF, a variety of regularization penalties, based on either the characteristics of the data or domain-specific prior knowledge, have been imposed. For example, two most widely used penalties in machine learning, i.e., sparseness and smoothness, have been exploited (Tao et al. 2017). Additionally, from the geometric perspective, a large-cone regularizer on the bases W has also been developed (Liu, Gong, and Tao 2017). All these regularized solutions have either empirically or theoretically demonstrated improvement to the original solutions (Lee and Seung 2001) in their generalizability. Furthermore, these penalty-based models account for the low-dimensional requirement, via balancing a trade-off among a regularization penalty, the low-rank requirement of bases W, as well as the model fitness measured by reconstruction error (Liu, Gong, and Tao 2017), KL divergence (Lee and Seung 2001), Itakura-Saito divergence (Ivek 2015), or the earth mover's distance (Sandler and Lindenbaum 2011).

In this paper, we propose a diversified Bayesian NMF model, termed DivBNMF, to enhance the solutions. Our approach can be seen as a Bayesian extension of the large-cone regularized NMF (LCNMF) (Liu, Gong, and Tao 2017) yet exhibits several advantages. First, thanks to the kernel trick

*Part of this work was done when MQ was with Hangzhou Dianzi University.

adopted in DPP, our formulation is more flexible. LCNMF with the two regularizers, i.e., large-volume and large-angle, are two special cases of our model. Second, instead of manually tuning the balance between model fitness and regularization, our method automates this process via Bayesian inference. Third, the proposed model is applied to solve the multi-label learning problem, and the experimental results demonstrate its effectiveness.

Related work

NMF and its extensions Due to the non-convexity of NMF, different kinds of regularizers have been introduced to make the solution unique via enforcing model/application-constrained properties. The L_1-based penalties (Gillis 2012) induce sparse representations to facilitate the interpretation of data, while the total variation penalty produces smooth representations suitable for sequence data (Seichepine et al. 2014). From the geometric point of view, minimizing the cone volume spanned by its bases W gives a unique solution for separable data (Schachtner et al. 2009). By contrast, maximizing the cone volume has been proven to achieve better generalizability (Liu, Gong, and Tao 2017). Two more examples considering the data structure are a graph-based manifold penalty (Cai et al. 2009) and a spatial localization based penalty (Li et al. 2001). To sum up, regularization constrained NMF achieves better performance under their assumed scenarios either empirically or theoretically when compared with the basic model.

NMF has also been extended under the probabilistic scheme (Gillis et al. 2019). The Bayesian counterparts of the penalties listed above have been developed. The connection between the regularized NMF and its Bayesian counterpart has been established. Within the Bayesian framework, noise models (likelihood functions) take over the role of reconstruction error functions while the prior distributions are responsible for encoding regularization (Ivek 2015) during maximization a posterior (MAP) estimation. First, in terms of objectives, the correspondence between the Bayesian version and the basic NMF are: a normally distributed noise likelihood and the Frobenius reconstruction error, a continuous Poissionian noise likelihood and the generalized Kullback-Leibler (KL) divergence, a zero-mean normally distributed noise likelihood or unit-mean Gamma noise likelihood and the Itakura-Saito divergence. Second, the MAP solutions to the Bayesian NMF under various priors are equal to the solutions to the regularized NMF with certain regularizers. The most commonly used prior-regularizer pairs are: The MAP solution under an exponential prior is equivalent to an L_1-based penalty - inducing sparsity; a zero-mean normal prior, which guarantees the uniqueness of the solution (Bayar, Bouaynaya, and Shterenberg 2014), derives an L_2-based penalty; a volume prior proposed in (Arngren, Schmidt, and Larsen 2011) corresponds to the penalty of minimizing the cone volume.

Following the Bayesian development, we extend NMF with the determinantal point processes (DPP) prior which can be geometrically interpreted as a large-volume penalty.

NMF for Multi-Label Learning Multi-label learning (MLL), to handle the situation of assigning an instance multiple labels (Sun et al. 2019; Xing et al. 2019), has recently become a popular tool in many applications, e.g., image processing (Zhang et al. 2019) and text analysis. Many efforts attempt to capture the label-label relationships (Xie et al. 2017; Gong et al. 2019a), and then integrate them into the traditional feature-label learning procedure (Yang et al. 2016). Apart from those, with the intuition that the label-associated parts should correspond to the bases in the decomposition of NMF, we apply the diversity-encouraging prior to enhance the feature learning in MLL. Then a naive K nearest neighbour (K-NN) classifier is employed to finalize MLL. This scheme follows the state-of-the-art developments (Tao et al. 2017; Gong et al. 2019b).

Determinantal Point Processes (DPP) Diversity is a good measure to subsets whenever the property of dissimilarity or repulsiveness is required. Apart from its direct application scenarios (Kulesza and Taskar 2011), diversity has recently become a popular regularizer to enhance model abilities (Qiao et al. 2015). Since DPP was introduced into the machine learning community (Kulesza and Taskar 2012), it has been an effective tool to model the diversity of a subset. It provides a powerful repulsive modeling tool within an easily extended probabilistic framework, and algorithms including self-contained efficient learning and inference have also been developed for it. Following its recent success, we employ DPP here as a prior encoding diversity amongst NMF bases to develop a Bayesian counterpart of the large-cone NMF (LCNMF) (Liu, Gong, and Tao 2017).

Our Model

Background for DPP

DPP is popular for modeling repulsion. In a continuous space, given $S \subseteq \mathcal{R}^D$ and a kernel $L : S \times S \mapsto \mathcal{R}$ with $L(x, y) = \sum_{n=1}^{\infty} \lambda_n \phi_n(x) \bar{\phi}_n(y)$, the probability density of a point configuration $A \subset S$ under a DPP is given by

$$p_L(A) = \frac{\det(L_A)}{\prod_{n=1}^{\infty}(\lambda_n + 1)}, \tag{1}$$

where λ_n and $\phi_n(x)$ are eigenvalues and eigenfunctions, and $\bar{\phi}_n(x)$ are the complex conjugate of $\phi_n(x)$.

When the cardinality of diverse subsets is fixed to K, a K-DPP (Kulesza and Taskar 2011) is then given as

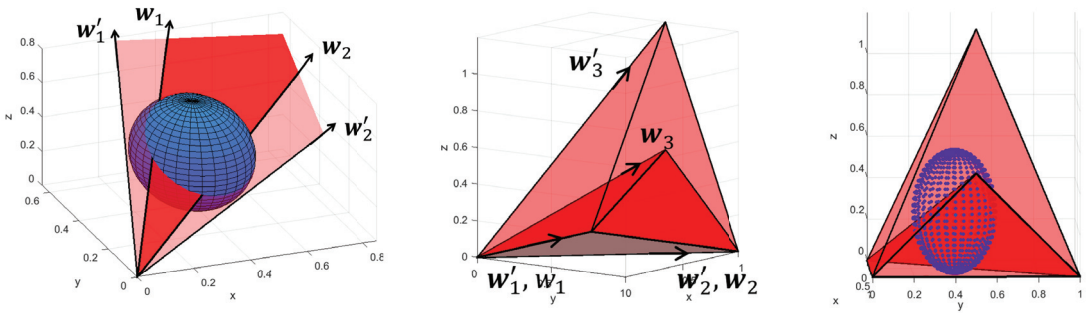
$$p_L^k(A) = \frac{\det(L_A)}{e_k(\lambda_{1:\infty})}, \tag{2}$$

where $e_k(\lambda_{1:\infty})$ is the kth elementary symmetric polynomial of the kernel L.

Building on the work of (Kulesza and Taskar 2012), the kernel L is decomposed as

$$L(\boldsymbol{w}_i, \boldsymbol{w}_j) = q(\boldsymbol{w}_i)\ell(\boldsymbol{w}_i, \boldsymbol{w}_j)q(\boldsymbol{w}_j), \tag{3}$$

where $q(\boldsymbol{w}_i)$ is interpreted as a quality function at a point $\boldsymbol{w}_i$ and $\ell(\boldsymbol{w}_i, \boldsymbol{w}_j)$ as a similarity function between two points $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$. Furthermore, the similarity kernel ℓ can be decomposed as a Gram matrix $\ell(\boldsymbol{w}_i, \boldsymbol{w}_j) = B_i^\top B_j$, where



(a) Large-cone NMF (b) Large-volume of DPP (c) DPP prior for NMF

Figure 1: A geometric interpretation of DPP for NMF. (a) Two simplicial cones in a same $2-D$ plane are shown to approximately encompass the data points in the 3-D space. Only the angles of the two cones are distinct and therefore are the key factor in determining their reconstruction performance within NMF. The cone spanned by $\{\boldsymbol{w}_1^{'}, \boldsymbol{w}_2^{'}\}$, which associates with a larger angle, encompasses more data points than the one spanned by $\{\boldsymbol{w}_1, \boldsymbol{w}_2\}$. In other words, the bases of a large cone achieve better reconstruction performance. (b) Two tetrahedrons sharing the same base but with distinct height values are shown. The one spanned by $\{\boldsymbol{w}_1^{'}, \boldsymbol{w}_2^{'}, \boldsymbol{w}_3^{'}\}$ with a larger height has larger volume than the one spanned by $\{\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3\}$ does. DPP favors the tetrahedron with the larger volume value, because it is built on a determinant operation, whose geometric explanation associates with the volume constructed from the bases. (c) The NMF bases forming a larger volume encompass more data points than the one with a smaller volume as shown in the figure. Therefore, DPP is applied as a prior to enforce large volume constraints to allow NMF achieve better reconstruction results.

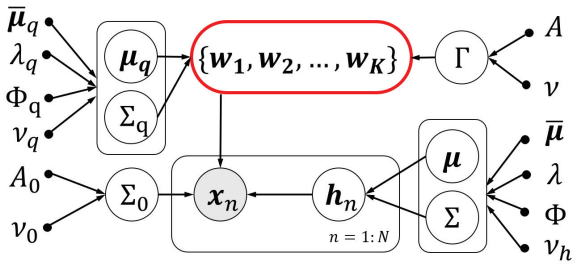


Figure 2: Graphical representation of DivBNMF.

$B_i \in \mathcal{R}^d$ represents a *normalized diversity feature* for point $\boldsymbol{w}_i \in A$ with d can be ∞ as transformed by the similarity kernel ℓ.

With this decomposition, the definition of DPP can be explained from a geometric point of view, as demonstrated in Figure 1b. The probability of a 3-DPP (2) is proportional to the determinant of the similarity kernel associated with each diversity feature set, namely to the determinant of the Gram matrix, which is geometrically equal to the square of the volume of the parallelepiped spanned by those diversity features. Two sets of diversity features $\{\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3\}$ and $\{\boldsymbol{w}_1^{'}, \boldsymbol{w}_2^{'}, \boldsymbol{w}_3^{'}\}$ form two tetrahedrons sharing the same base but with distinct heights. From the figure, the items in the first set are more diverse to each other than the ones in the second set as the third vector $\boldsymbol{w}_3^{'}$ is further from the two shared vectors than $\boldsymbol{w}_3$. Moreover, the first set has a larger height thus has a larger volume. Therefore, it is assigned a higher probability by the DPP/K-DPP. In other words, DPP/K-DPP prefers subsets expanding larger volumes.

The Proposed DivBNMF

NMF approximates a nonnegative matrix $X \in \mathbb{R}_+^{D\times N}$ with two nonnegative matrices $W \in \mathbb{R}_+^{D\times K}$ and $H \in \mathbb{R}_+^{K\times N}$, with $K \ll \max\{D, N\}$.

From its exact algebraic formulation, i.e., $X = WH$, NMF has a geometric interpretation (Donoho and Stodden 2004). Let the K bases $\{\boldsymbol{w}_k \in \mathcal{R}_+^D\}_{k=1}^K$ be the extreme rays for a simplicial cone, i.e.,

$$C_W = \{\boldsymbol{x} : \boldsymbol{x} = \sum_{i=1}^{K} h_i \boldsymbol{w}_i, h_i \in \mathcal{R}_+\}.$$

Geometrically, the exact NMF equation indicates that all D-dimensional data points in the matrix X lie within this simplicial cone. However, when considering based on the inverse, given X, there may exist many different simplicial cones that satisfy the NMF equation. This is caused by the non-convexity nature of NMF. However, a large simplicial cone can be developed by adding extra constraints to the original problem. For example, in (Liu, Gong, and Tao 2017), a large simplicial cone with either a large volume or a large angle constraint results in a good approximate solution. These additional constraints lead to better performance in the reconstruction error and generalizability. This is visually explained in Figure 1a. NMF seeks a simplicial cone to approximately encompass the data points scattering on the surface of a $3-D$ ball. The simplicial cone with a larger angle value can encompass more data points than the one with a smaller angle value, and thus allowes NMF to achieve better reconstruction performance.

The goal of the geometry of NMF is to seek a large simplicial cone generated by the columns of W to encompass as many data points as possible. At the same time, the geometry

of DPP provides a way to favor a large-volume for the set of bases W. Therefore, we employ DPP as a large-volume prior for the columns of W integrated into the probabilistic NMF, as explained in Figure 1c. Based on this idea, we developed a diversified Bayesian NMF model (DivBNMF). Several benefits can be expected from this DPP induced large volume NMF. First, we use the efficient parameter learning and inference algorithms developed for DPP to develop the inference for our model. Second, a principled Bayesian exploitation of the NMF decomposition provides more insights than a single NMF solution. Finally, kernel tricks, on which DPP is built, can be further exploited.

The graphical representation for the proposed model is shown in Figure 2, where shallow circles represent random variables such as parameters W and H and hyper-parameters $\Sigma_0, \boldsymbol{\mu}_q, \Sigma_q, \Gamma, \boldsymbol{\mu}, \boldsymbol{\Sigma}$, shadowed circles represent observations X, and solid dots represent hyper-prior parameters $(\bar{\boldsymbol{\mu}}_q, \lambda_q, \dots)$. The bold red ellipse emphasizes the DPP prior for the bases W of NMF.

The generative process for data and NMF parameters corresponding to the above graphical representation is listed below.

$$\begin{aligned}
\boldsymbol{x}_n &\sim \mathcal{N}(\boldsymbol{x}_n; W\boldsymbol{h}_n, \Sigma_0)u(\boldsymbol{x}_n),\ \boldsymbol{x}_n \in \mathbb{R}_+^D,\ n = 1, \dots, N,\\
\{\boldsymbol{w}_k\} &\sim \text{K-DPP}_L(\boldsymbol{\mu}_q, \Sigma_q, \Gamma),\ \boldsymbol{w}_k \in \mathbb{R}_+^D,\ k = 1, \dots, K,\\
\boldsymbol{h}_n &\sim \mathcal{N}(\boldsymbol{h}_n; \boldsymbol{\mu}, \Sigma)u(\boldsymbol{h}_n),\ \boldsymbol{h}_n \in \mathbb{R}_+^K,\ n = 1, \dots, N.
\end{aligned}$$

Each nonnegative observation $\boldsymbol{x}_n$ is assumed to be sampled from a truncated Gaussian distribution with mean $W\boldsymbol{h}_n$ and covariance Σ_0, with $u(\cdot)$ denoting the unit step function. New low-dimensional representations $\{\boldsymbol{h}_n\} \subset \mathcal{R}_+^K$ are assumed to be independently sampled from a truncated Gaussian prior parameterized with mean $\boldsymbol{\mu}$ and covariance Σ. The bases W of NMF are assumed from a K-DPP prior parameterized with kernel L established in (3) with quality parameters $\boldsymbol{\mu}_q, \Sigma_q$ and diversity parameters Γ. The decomposition of quality and similarity is based on the modeling convenience as in (Kulesza and Taskar 2012). A quality function is

$$q(\boldsymbol{w}_i) = \exp\{-\frac{1}{2}(\boldsymbol{w}_i - \boldsymbol{\mu}_q)^\top \Sigma_q^{-1}(\boldsymbol{w}_i - \boldsymbol{\mu}_q)\}, \quad (4)$$

and a similarity function is

$$\ell(\boldsymbol{w}_i, \boldsymbol{w}_j) = \exp\{-\frac{1}{2}(\boldsymbol{w}_i - \boldsymbol{w}_j)^\top \Gamma^{-1}(\boldsymbol{w}_i - \boldsymbol{w}_j)\}. \quad (5)$$

The quality and similarity functions can be polynomial, Cauchy, RBF, etc. For computational convenience, the RBF kernel is employed throughout this paper.

Additionally, hyper-parameters are assigned hyper-priors which are listed below:

$$\begin{aligned}
\Sigma_0 &\sim \mathcal{W}^{-1}(A_0, \nu_0),\\
(\boldsymbol{\mu}_q, \Sigma_q) &\sim \text{NIW}(\bar{\boldsymbol{\mu}}_q, \lambda_q, \Phi_q, \nu_q),\\
\Gamma &\sim \mathcal{W}^{-1}(A, \nu),\\
(\boldsymbol{\mu}, \Sigma) &\sim \text{NIW}(\bar{\boldsymbol{\mu}}, \lambda, \Phi, \nu_h).
\end{aligned}$$

Most of these hyper-prior forms are chosen from the consideration of computational convenience. For example, Normal-Inverse-Wishart (NIW) distributions are self-conjugate with respect to Gaussian likelihoods, which satisfies H. An inverse Wishart distribution, denoted by $\mathcal{W}^{-1}$, is chosen to define the prior on symmetric, nonnegative definite matrices such as Σ_0 and Γ.

Hyper-parameters are assigned weakly informative priors to allow them to be studied from observations with posterior inference. To implement this, the parameters for the hyper-priors are set below. For inverse Wishart priors $\nu = \nu_0 = D + 1$ and $A = A_0 = I_D$ so that the means for the two variables are $\text{E}(\Gamma) = \frac{A}{\nu - D + 1}$ and $\text{E}(\Sigma_0) = \frac{A_0}{\nu_0 - D + 1}$ respectively. Here I_D is the D-dim identity matrix. For NIW priors, $\nu_q = D$, $\nu_h = K$, and $\Phi_q = I_D$, $\Phi = I_K$ have the same meaning as the inverse Wishart prior. Finally, $\lambda_q = \lambda = 1$ and $\bar{\boldsymbol{\mu}}_q$ and $\bar{\boldsymbol{\mu}}$ are vectors of zero.

Model Remarks

Advantages of a DPP prior First, from the geometric point of view, solutions with less reconstruction error and better generalizability are naturally obtained due to the large volume capability. Second, more succinct solutions can be obtained due to this mutually repulsive prior imposed on bases W. In other words, the solutions obtained by our model require a smaller K when compared with the basic NMF to achieve similar or better performance. Third, the solutions under the scenario of MLL induce more diverse and abstract 'parts' representation to facilitate the labeling task.

Relation to LCNMF When the quality function for all basis $\{\boldsymbol{w}_i\}$ are set equally to 1, i.e., $q(\boldsymbol{w}_i) = 1$ with $i = 1, \dots, K$ and the pairwise similarity functions are set to the inner product of two associated bases in Euclidean space, i.e., $\ell(\boldsymbol{w}_i, \boldsymbol{w}_j) = \boldsymbol{w}_i^\top \boldsymbol{w}_j$, and then the prior encoded with DPP under the MAP inference is degenerated to the large-cone regularizer in LCNMF (Liu, Gong, and Tao 2017). Thus, LCNMF is a special case of the proposed DivBNMFM.

Tuning trade-off parameters In LCNMF, one needs to tune a trade-off parameter manually, i.e., λ, to achieve a balance between the large-cone regularization and model fitness. However, the proposed model, following the principled Bayesian framework, automatically adjusts the balance between the large-volume regularization induced from DPP based hyper-priors and the model fitness given empirical observations.

Inference

Due to the non-conjugacy in our model, the inference involves intractable integrals. Therefore, precisely maintaining full posterior distributions of all hidden random variables including parameters and hyper-parameters is infeasible. Thus, we use approximate solutions. Here, the Gibbs sampling algorithm is adopted, which is an MCMC algorithm and suitable for multivariate probability inference. It obtains samples of all unobserved variables from a joint posterior distribution by alternately and iteratively sampling from the posterior distribution of each variable conditional on all other variables. The overall sampling procedure is summarized in Alg. 1, and the associated conditional posterior distributions for all variables are individually derived and details can be found in supplemental materials.

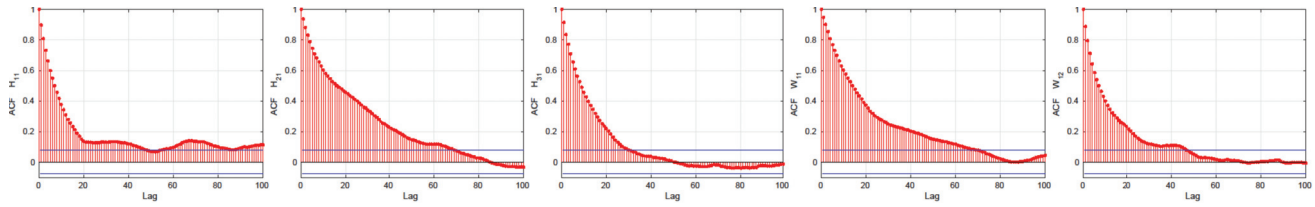


Figure 3: ACF plots of $H_{11}, H_{21}, H_{31}, W_{11}, W_{12}$ for a synthetic dataset with $K = 5$.

Table 1: Reconstruction error comparison on synthetic dataset

Measures	Methods	K				
		2	4	6	8	10
Euclidean	NMF	25.68	21.48	19.81	16.57	0.92
	LVNMF (λ)	25.60 (0.1)	21.34 (0.1)	20.07 (0.01)	16.4 (0.10)	**0.08** (1.0)
	LANMF (λ)	25.53 (1.0)	21.40 (0.1)	19.88 (0.01)	16.41 (1.0)	0.19 (0.01)
	DivBNMF-MAP (Γ)	**24.10** (0.10)	**19.55** (0.1)	**17.74** (1.0)	**14.67** (0.1)	*0.09* (10)
	DivBNMF-Mean (Γ)	*24.18* (0.1)	*19.88* (0.1)	*18.04* (1.0)	*15.23* (0.1)	0.1 (10)
Frobenius	NMF	56.35	45.58	35.57	23.11	1.89
	LVNMF (λ)	56.31 (0.1)	45.61 (0.1)	35.48 (1.0)	23.16 (0.1)	**0.13** (1.0)
	LANMF (λ)	56.28 (1.0)	45.67 (0.1)	35.46 (1.0)	23.22 (0.10)	0.33 (0.01)
	DivBNMF-MAP (Γ)	**52.91** (0.1)	**42.84** (0.1)	**33.6** (1.0)	**21.09** (0.1)	*0.22* (10)
	DivBNMF-Mean (Γ)	*53.11* (0.1)	*43.32* (0.1)	*34.28* (1.0)	*22.39* (0.1)	0.23 (10)

Algorithm 1 Gibbs Sampling for DivBNMF

Data: $t = 1$; sample number: N_S;
initializations $W^t, H^t, \Sigma_0^t, \boldsymbol{\mu}_q^t, \Sigma_q^t, \Gamma^t, \boldsymbol{\mu}^t, \Sigma^t$.
Result: $\{W^t, H^t, \Sigma_0^t, \boldsymbol{\mu}_q^t, \Sigma_q^t, \Gamma^t, \boldsymbol{\mu}^t, \Sigma^t\}_{t=1}^{N_S}$.
while $t < N_S$ **do**
  $t = t + 1$;
  sampling $W^t | X, H^{t-1}, \Sigma_0^{t-1}, \boldsymbol{\mu}_q^{t-1}, \Sigma_q^{t-1}, \Gamma^{t-1}, \boldsymbol{\mu}^{t-1}, \Sigma^{t-1}$;
  sampling $H^t | X, W^t, \Sigma_0^{t-1}, \boldsymbol{\mu}^{t-1}, \Sigma^{t-1}$ from truncated Gaussian distribution;
  sampling $\Sigma_0^t | X, W^t, H^t$ from inverse Wishart distribution;
  sampling $(\boldsymbol{\mu}^t, \Sigma^t) | H^t$ from Normal-Inverse-Wishart distribution;
  sampling $(\boldsymbol{\mu}_q^t, \Sigma_q^t, \Gamma^t) | W^t, \boldsymbol{\mu}_q^{t-1}, \Sigma_q^{t-1}, \Gamma^{t-1}$;
end

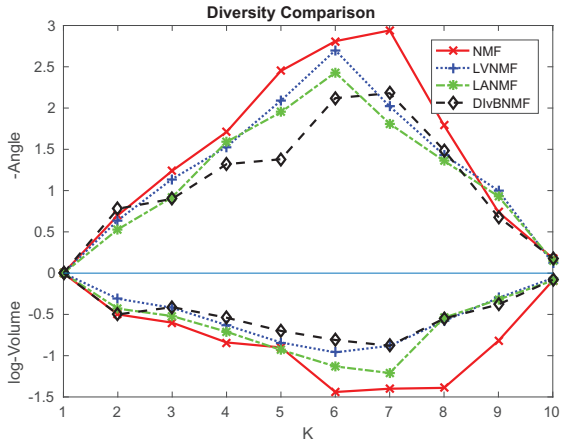


Figure 4: Comparison of diversity measurements on W_is learned from the synthetic dataset.

Experiments

In this section, we present reconstruction results conducted on both a synthetic dataset and a real-world dataset, i.e., the MULAN scene. We also apply the decomposed NMF representation to perform MLL on a real-world dataset to verify the conjecture that the proposed DivBNMF enhances parts-based learning and thus benefits the prediction accuracy of the MLL task. Note that to make fair comparison among different algorithms, the shared parameters W and H were initialized with the same value.

Experiments on synthetic dataset

A synthetic matrix $X = (36 \times 100)$ was randomly generated. The experimental settings are shown below. The results for LANMF and LVNMF were optimized by varying the trade-off parameter λ among $\{0.01, 0.1, 1, 10, 100\}$. All reconstruction errors were averaged over 10 runs.

Sampling analysis The first 2000 iterations were omitted as burn-in period. The ACF plots of five variables are presented in Figure 3. Although the sample autocorrelation values of different variables decay in different speeds in terms of lags, all of them are reasonably small when the lag is beyond 100. Based on this observation, we collected independent samples from the Gibbs chain by sequentially keeping one sample every 100 interval after the burn-in period.

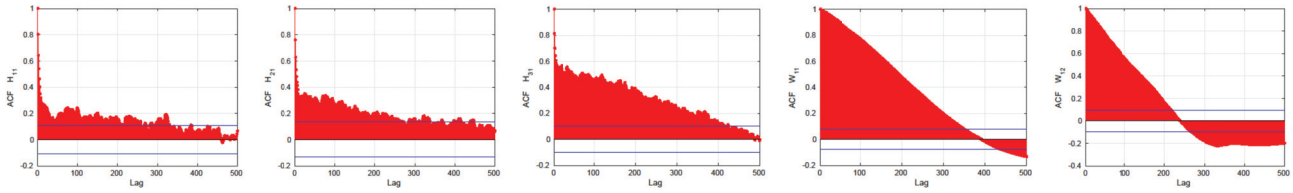


Figure 5: ACF plots of $H_{11}, H_{21}, H_{31}, W_{11}, W_{12}$ for MULAN scene dataset from result with $K = 10$.

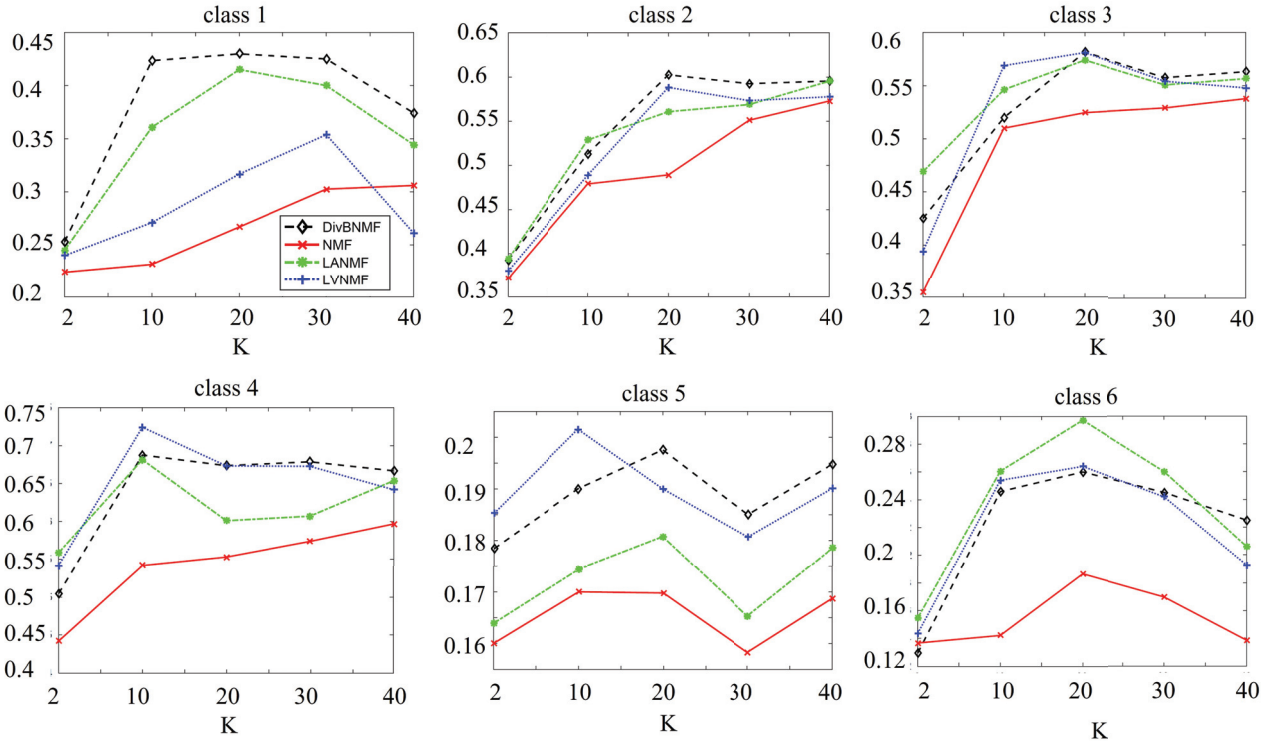


Figure 6: mAP results for 6 classes of MULAN scene testing sets.

Reconstruction error The results of reconstruction errors with varying K for both Mean and MAP estimations of the proposed DivBNMF as well as the point estimations of the baselines are summarized in Table 1. The number in brackets in each cell corresponds to the parameter setting resulting in the reconstruction error. Several trends can be extracted from the table. First, along with the increase of K, more bases were involved, and as a result, better reconstruction was obtained. Second, the diversity-encouraging methods including the proposed DivBNMF, LANMF and LVNMF achieved better performance compared to the unconstrained NMF. This shows the effectiveness of the diversity prior. Furthermore, the proposed DivBNMF with either MAP estimation or mean estimation consistently outperformed the two large cone induced diversity NMFs, namely LANMF and LVNMF. We attribute this superiority to the more flexible diversity modeling ability of our method.

Diversity comparison The diversity measurements of volume and total pairwise angles over the bases Ws corresponding to the above results (MAP estimation for our method) are shown in Figure 4. The W_is learned by NMF achieved the worst diversity measure, while those learned from LANMF and LVNMF demonstrated higher performance of diversity measurements. Comparatively, the proposed DivBNMF achieved the most diverse bases. Comparing this result to the above reconstruction performance, we conclude that the diversity-encouraging priors improve NMF's reconstruction performance. Additionally, the DPP prior within the Bayesian framework encodes more flexible repulsiveness.

Experiments on MULAN scene dataset

We evaluated the performance of the proposed DivBNMF regarding MLL on one nonnegative featured benchmark dataset: the MULAN scene dataset [1]. It contained 2047 images with six labels, each of which was represented by a

[1] http://mulan.sourceforge.net/

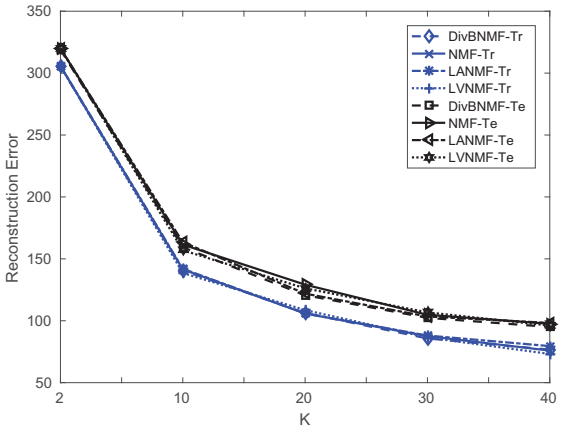


Figure 7: Reconstruction error on both training and testing sets of MULAN scene.

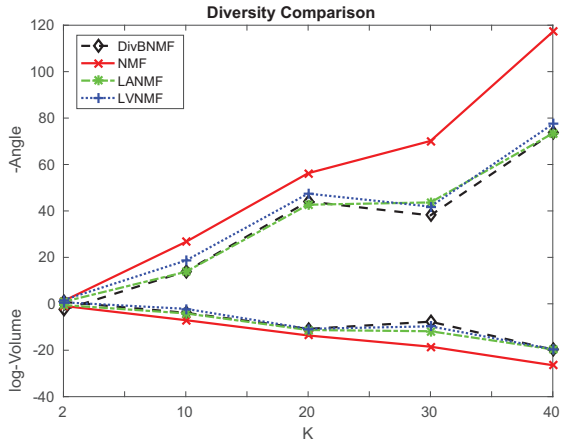


Figure 8: Comparison of diversity measurements on learned W_is for MULAN scene.

294-dimensional nonnegative feature vector. It was split into a training set containing 1211 images and a test set containing 1196 images. Average precision (AP), summarizing over the entire precision-recall curve, was examined as an MLL evaluation criterion.

Sampling analysis The first 10000 iterations were dumped as burn-in period. The thinning interval was set to 500 to collect independent samples as supported by reaching reasonable small ACF values regarding lags for these variables, as evidenced by five ACF plots in Figure 5.

Performance analysis Figure 7 presents the reconstruction error results with varying Ks for both training and testing subsets, and Figure 6 presents the corresponding AP results for the six classes. For the reconstruction error, as shown in the first plot, it was difficult to determine which method among the proposed DivBNMF-MAP and the three baselines achieved the best performance, since these curves for either the training set or testing set were almost superposed with each other. In contrast, in terms of AP, as shown in the second plots, different methods inhibit various ability levels for multi-label prediction. The methods with a diverse regularizer/prior consistently achieved better results over all six classes than the traditional NMF. When zooming in for a close look at each class, the proposed DivBNMF-MAP consistently achieved the best results for classes 1, 2, 3, and 4, while obtaining comparative results for classes 5 and 6.

Diversity analysis The diversity measurements for the four methods with increasing K are shown in Figure 8. The W_is learned from NMF inhibited the lowest diversity in volume and total pairwise mutual angle comparing to the other three diversity-encouraging methods, i.e., DivBNMF, LANMF, and LVNMF. Among those, the proposed DivBNMF obtained a slightly better diversity measurement. To sum up, all these experimental results together verify that diversity-encouraging regularizers/priors facilitate NMF to improve multi-label prediction. Furthermore, a DPP-encoded prior within the Bayesian framework achieves better diversity modeling flexibility.

Conclusion

In this paper, we propose an extended Bayesian NMF approached, termed DivBNMF, with a diversity-encouraging prior for columns of its bases matrix. The merits of DivBNMF include its modeling diversity flexibility inherited from DPP, as well as its labor-reducing ability of automatically adjusting trade-off parameters via hyper-priors benefited from the Bayesian inference. Experiments conducted on both synthetic data reconstruction and real-world multi-label learning tasks verify the effectiveness of the proposed method.

Our future work will focus on two directions. First, due to the non-conjugacy, the posterior inference requires inner loop calculations within each sample sampling. Such operations are considerably time-consuming. Therefore, exploring the possibility of conjugacy would help speed up the whole procedure. The second direction is to extend the current model by enabling it to automatically select the number of bases, namely K. Nonparametric learning (Xuan et al. 2018) could be integrated into the current framework to achieve this task.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 61702145, 61971172 and 61836002, and in part by Australian Research Council Projects FL-170100117, DP-180103424, IH-180100002 and DE190101473.

References

Arngren, M.; Schmidt, M. N.; and Larsen, J. 2011. Unmixing of hyperspectral images using bayesian non-negative matrix factorization with volume prior. *Journal of Signal Processing Systems* 65(3):479–496.

Bayar, B.; Bouaynaya, N.; and Shterenberg, R. 2014. Probabilistic non-negative matrix factorization: Theory and application to microarray data analysis. *Journal of Bioinformatics and Computational Biology* 12(01):1450001.

Cai, D.; He, X.; Wang, X.; Bao, H.; and Han, J. 2009. Locality preserving nonnegative matrix factorization. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 9, 1010–1015.

Chen, Z.; Li, Y.; Sun, X.; Yuan, P.; and Zhang, J. 2019. A quantum-inspired classical algorithm for separable nonnegative matrix factorizations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 4511–4517.

Degleris, A., and Gillis, N. 2019. A provably correct and robust algorithm for convolutive nonnegative matrix factorization. *arXiv preprint arXiv:1906.06899*.

Donoho, D., and Stodden, V. 2004. When does nonnegative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems*, 1141–1148.

Ge, R., and Zou, J. 2015. Intersecting faces: non-negative matrix factorization with new guarantees. In *International Conference on Machine Learning*, 2295–2303.

Gillis, N.; Hien, L. T. K.; Leplat, V.; and Tan, V. Y. 2019. Distributionally robust and multi-objective nonnegative matrix factorization. *arXiv preprint arXiv:1901.10757*.

Gillis, N. 2012. Sparse and unique nonnegative matrix factorization through data preprocessing. *Journal of Machine Learning Research* 13:3349–3386.

Gong, C.; Liu, T.; Yang, J.; and Tao, D. 2019a. Large-margin label-calibrated support vector machines for positive and unlabeled learning. *IEEE Transactions on Neural Networks and Learning Systems* 30(11):3471–3483.

Gong, C.; Shi, H.; Liu, T.; Zhang, C.; Yang, J.; and Tao, D. 2019b. Loss decomposition and centroid estimation for positive and unlabeled learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1–14.

Ivek, I. 2015. Probabilistic formulations of nonnegative matrix factorization.

Kulesza, A., and Taskar, B. 2011. k-dpps: Fixed-size determinantal point processes. In *Proceedings of the International Conference on Machine Learning*, 1193–1200.

Kulesza, A., and Taskar, B. 2012. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning* 5.

Lee, D. D., and Seung, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788–791.

Lee, D. D., and Seung, H. S. 2001. Algorithms for nonnegative matrix factorization. In *Advances in Neural Information Processing Systems*, 556–562.

Li, S. Z.; Hou, X. W.; Zhang, H. J.; and Cheng, Q. S. 2001. Learning spatially localized, parts-based representation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE.

Liu, T.; Gong, M.; and Tao, D. 2017. Large-cone nonnegative matrix factorization. *IEEE Transactions on Neural Networks and Learning Systems*.

Qiao, M.; Bian, W.; Xu, D.; Yi, R.; and Tao, D. 2015. Diversified hidden markov models for sequential labeling. *IEEE Transactions on Knowledge and Data Engineering* 27(11):2947–2960.

Sandler, R., and Lindenbaum, M. 2011. Nonnegative matrix factorization with earth mover's distance metric for image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(8):1590–1602.

Schachtner, R.; Pöppel, G.; Tomé, A. M.; and Lang, E. W. 2009. Minimum determinant constraint for nonnegative matrix factorization. In *ICA*, volume 9, 106–113. Springer.

Seichepine, N.; Essid, S.; Févotte, C.; and Cappe, O. 2014. Piecewise constant nonnegative matrix factorization. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 6721–6725. IEEE.

Sun, L.; Feng, S.; Wang, T.; Lang, C.; and Jin, Y. 2019. Partial multi-label learning by low-rank and sparse decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5016–5023.

Tao, D.; Tao, D.; Li, X.; and Gao, X. 2017. Large sparse cone nonnegative matrix factorization for image annotation. *ACM Transactions on Intelligent Systems and Technology* 8(3):37.

Xie, P.; Salakhutdinov, R.; Mou, L.; and Xing, E. P. 2017. Deep determinantal point process for large-scale multi-label classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 473–482.

Xing, Y.; Yu, G.; Domeniconi, C.; Wang, J.; Zhang, Z.; and Guo, M. 2019. Multi-view multi-instance multi-label learning based on collaborative matrix factorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 5508–5515.

Xuan, J.; Lu, J.; Zhang, G.; Xu, R. Y.; and Luo, X. 2018. Doubly nonparametric sparse nonnegataive matrix factorization based on dependent indian buffet processes. *IEEE Transactions on Neural Network Learning Systems* 29(5):1835–1849.

Yang, H.; Tianyi Zhou, J.; Zhang, Y.; Gao, B.-B.; Wu, J.; and Cai, J. 2016. Exploit bounding box annotations for multi-label object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 280–288.

Zhang, M.; Wang, N.; Li, Y.; and Gao, X. 2019. Neural probabilistic graphical model for face sketch synthesis. *arXiv preprint arXiv:10.1109/TNNLS.2018.2890017*.

Zhou, T.; Bian, W.; and Tao, D. 2013. Divide-and-conquer anchoring for near-separable nonnegative matrix factorization and completion in high dimensions. In *IEEE International Conference on Data Mining (ICDM)*, 917–926. IEEE.