Variational Bayesian Matrix Factorization for Bounded Support Data

Project Report to be submitted in Partial Fulfillment of the Requirements for the Award of the Degree of

Bachelor of Technology
in
Instrumentation Engineering

by

Hasanth Palanchu

Under the supervision of

Prof. Anirban Mukherjee

Department of Electrical Engineering
Indian Institute of Technology Kharagpur
December 2015

Abstract

A Novel Bayesian matrix factorization method for bounded support data is presented. Each entry in the observation matrix is assumed to be beta distributed. As the beta distribution has two parameters, two parameter matrices can be obtained, which matrices contain only nonnegative values. In order to provide low-rank matrix factorization, the nonnegative matrix factorization (NMF) technique is applied. Furthermore, each entry in the factorized matrices, i.e., the basis and excitation matrices, is assigned with gamma prior. Therefore, we name this method as beta-gamma NMF (BG-NMF). Due to the integral expression of the gamma function, estimation of the posterior distribution in the BG-NMF model can not be presented by an analytically tractable solution. With the variational inference framework and the relative convexity property of the log-inverse-beta function, we propose a new lower-bound to approximate the objective function. With this new lower-bound, we derive an analytically tractable solution to approximately calculate the posterior distributions. Each of the approximated posterior distributions is also gamma distributed, which retains the conjugacy of the Bayesian estimation. In addition, a sparse BG-NMF can be obtained by including a sparseness constraint to the gamma prior. Evaluations with synthetic data and real life data demonstrate the good performance of the proposed method.

Keywords: Non-negative matrix factorization, Bayesian estimation, Bounded support data, Variational inference, Extended factorized approximation, Relative convexity.

Contents

1 Introduction

NON-negative matrix factorization (NMF) was introduced as an alternative way for reducing the dimensionality of the data. Unlike the principal component analysis (PCA) or the independent component analysis (ICA) which has no constraint on the data, the NMF factorizes a non-negative matrix into a product of two non-negative matrices (a basis matrix and an excitation matrix). It is a fundamental technique for low rank non-negative matrix approximation and has been widely used in information retrieval, image analysis, source separation, speech de-noising, collaborative filtering and other applications.

In the previous research, a lot of algorithms were proposed to realize the matrix factorization efficiently. By minimizing the Formula norm of the reconstruction error and the Kullback-Leibler (KL) divergence between the original matrix and the reconstructed matrix respectively, Lee et al. proposed two algorithms for NMF. To emphasize the effect of presenting the local features in the face images, the optimization with sparseness constraints was introduced. Also, assigning different weights to the vectors in the basis matrix could also improve the local representation. Bayesian estimation, in general, can provide robust solution to parameter estimation. Authors proposed an solution for ICA with mean-field approach. This is an early solution for Bayesian treatment of matrix factorization. To extend the NMF into a probabilistic framework, Schmidt et al. considered the reconstruction error as Gaussian distributed and presented a Bayesian treatment to NMF where the reconstruction error Formula is assumed to be Gaussian distributed and exponential prior is assigned to the entries in the basis and excitation matrices. The Gibbs sampler was utilized to simulate the posterior distribution and an efficient iterated conditional mode (ICM) algorithm was proposed. To infer different optimization criteria, the relation between the Itakura-Saito (IS) divergence and some other cost function of the NMF (e.g., the Euclidean distance, the generalized KL divergence) was studied. Furthermore, the NMF with beta divergence, which is a general form of distance measure, was introduced. These two works were applied successfully on audio/musical data source separation. Cemgil assumed a Poisson distribution to the entries in the observation matrix and assign a gamma prior to the entries in the basis and the excitation matrices. Even though this method described the KL divergence measure in a statistical framework, there are two disadvantages. Firstly, the assumption that the entry in the original matrix is Poisson distributed violates the statistical interpretation of the KL-NMF on continuous data. Secondly, this paper applied the Poisson NMF (P-NMF) for gray image processing. The Poisson assumption was suitable only for integer data, so that the continuous property of the data was violated. Furthermore, the author applied the P-NMF to analyze image data, which ignored the bounded property of image data. One possible way of justifying the likelihood of KL-NMF for real data is writing Formula, where Formula are Poisson distributed and Formula is uniformly distributed on Formula. A recent paper for the Bayesian treatment of the NMF derived a nonparametric Bayesian NMF with gamma process for music record processing. In that paper, each entry in the observation matrix was assumed exponential distributed and the NMF was applied to the parameter matrix, rather than to the observation matrix directly. The gamma process was used to control the channel gain (weighting factor for each basis) such that the model size could be decided automatically based on the data. However, for the purpose of choosing the distribution appropriate for the spectrogram data, the

authors did not consider conjugate pairs of distribution.

2 Non-Negative Matrix Factorization

The conventional NMF problem is presented as

$$\mathbf{X}_{P\times T} \approx \mathbf{W}_{P\times K}\mathbf{V}_{K\times T}, \tag{1}$$

where $\mathbf{X}_{P\times T}$, $\mathbf{W}_{P\times K}$, and $\mathbf{V}_{K\times T}$ contains non-negative values X_{pt}, W_{pk}, and V_{kt} respectively and $p = 1, ..., P$, $t = 1, ..., T$, $k = 1, ..., K$. Usually, we choose KT such that the NMF is a low rank matrix approximation. W and V are usually named as the basis matrix and the excitation matrix, respectively. Denoting the tth column in $\mathbf{X}$ as X_t, we have that X_t is a linear combination of all the columns in $\mathbf{W}$, with weighting coefficients from the tth column in $\mathbf{V}$. In addition to the conventional NMF method, the NMF can also be treated in a probabilistic way so that we estimate the parameters of the underlying model, instead of estimating the basis and the excitation matrices directly.

In several practical applications, the data we are processing have bounded support property. In the following paragraph, we propose a Bayesian matrix factorization approach for bounded support continuous data and derive an analytically tractable solution for calculation convenience (with conjugate pairs of prior and posterior distributions).

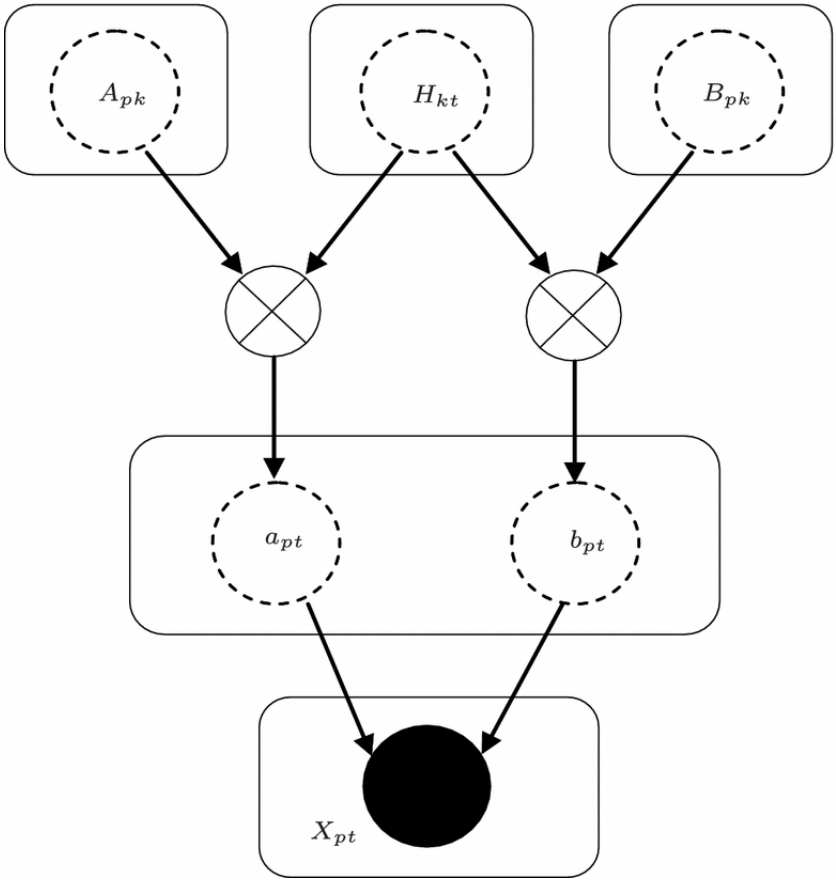
3 Bayesian Matrix Factorization For Bounded Support Data

The bounded support data (usually defined in the interval $[x_t, x_h]$, which can be linearly compressed to $[0, 1)$ can be modeled more efficiently with the beta distribution. So far as we know, there is no matrix factorization strategy proposed for the bounded support data. We believe this is due to the difficulty in explicitly placing the bounded support constraint in the factorized matrices. Instead of introducing the bounded support constraint directly to the factorized matrices, we propose a generative model for the observation matrix where the bounded support property is implicitly utilized.

3.1 Generative Model

We assume that each bounded support element X_{pt} is generated from a beta distribution with parameters a_{pt} and b_{pt}. Thus with an observation matrix $\mathbf{X}_{P\times T}$, we have two parameter matrices $\mathbf{a}$ and $\mathbf{b}$ of size $P\times T$, respectively. Similar to the GaP-NMF, we jointly factorize each parameter matrix, rather than the observation matrix, into a product of a basis matrix and an excitation matrix respectively as

$$\begin{aligned}\mathbf{a}_{P\times T} &\approx \mathbf{A}_{P\times K}\mathbf{H}_{K\times T},\\ \mathbf{b}_{P\times T} &\approx \mathbf{B}_{P\times K}\mathbf{H}_{K\times T}.\end{aligned} \tag{2}$$



Since all the entries in $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{H}$ are non-negative, we $\mathbf{X}$ assign a gamma prior to each entry. With the above description, we assume that the matrix (with element $X_{pt} \in [0, 1]$) is drawn according to the following generative model 1

$$\begin{aligned} A_{pk} &\sim \mathrm{Gamma}(\mathrm{A_{pk}}; \mu_0, \alpha_0), \\ B_{pk} &\sim \mathrm{Gamma}(\mathrm{B_{pk}}; \nu_0, \beta_0), \\ H_{kt} &\sim \mathrm{Gamma}(\mathrm{H_{kt}}; \rho_0, \zeta_0), \end{aligned} \tag{3}$$

$$X_{pt} \sim \mathrm{Beta}\Big(\mathrm{X_{pt}}; \sum_{\mathrm{k}} \mathrm{A_{pk}H_{kt}}, \sum_{\mathrm{k}} \mathrm{B_{pk}H_{kt}}\Big), \tag{4}$$

where $\mathrm{Gamma}(\mathrm{x}; \mathrm{k}, \theta)$ is the gamma density with parameters k, θ defined as

$$\mathrm{Gamma}(\mathrm{x}; \mathrm{k}, \theta) = \frac{\theta^{\mathrm{k}}}{\Gamma(\mathrm{k})} \mathrm{x}^{\mathrm{k}-1} \mathrm{e}^{-\theta \mathrm{x}}, \ \mathrm{k}, \theta > 0, \tag{5}$$

and $\mathrm{Beta}(\mathrm{x}; \mathrm{u}, \mathrm{v})$ is the beta density with parameter u, v defined as

$$\mathrm{Beta}(\mathrm{x}; \mathrm{u}, \mathrm{v}) = \frac{1}{\mathcal{B}(\mathrm{u}, \mathrm{v})} \mathrm{x}^{\mathrm{u}-1} (1 - \mathrm{x})^{\mathrm{v}-1}, \ \mathrm{u}, \mathrm{v} > 0, \tag{6}$$

where $\mathcal{B}(u, v) = \frac{\Gamma(u)\Gamma(v}{\Gamma(u+v)}$ and $\Gamma(.)$ is the gamma function. Fig. 1 shows the details of this generative model.

3.2 Variational Inference

If we consider the conjugate match between the prior distribution and the posterior distribution, the forms of the prior distribution and the posterior distribution are required to be the same.

Given the prior distribution, the inference to the posterior distribution is the central problem in the Bayesian analysis, which is also important in our BG-NMF model. The exact Bayesian inference for BG-NMF is not analytically tractable. With the principle of variational inference, we have already divided the latent variables $\mathbf{Z} = \mathbf{A}, \mathbf{B}, \mathbf{H}$ into disjoint groups A, B, H and assigned a gamma prior to each entry in those matrices (see Section 3.1). Thus the prior distributions of the latent variables are

$$\begin{aligned} p(\mathbf{A}) &= \prod_{p,k} p(A_{pk}), \\ p(\mathbf{B}) &= \prod_{p,k} p(B_{pk}), \\ p(\mathbf{H}) &= \prod_{k,t} p(H_{kt}), \\ p(\mathbf{Z}) &= p(\mathbf{A})p(\mathbf{B})p(\mathbf{H}). \end{aligned} \tag{7}$$

If we treat each element in X as conditionally independent from each other given the latent variable Z, the probability density function of the observation X is

$$\begin{aligned} p(\mathbf{X} \mid \mathbf{Z}) = \prod_{p,t} & \frac{1}{\mathcal{B}\left(\sum_k A_{pk}H_{kt}, \sum_k B_{pk}H_{kt}\right)} \\ & \times (X_{pt})^{\sum_k A_{pk}H_{kt}-1} \left(1 - X_{pt}\right)^{\sum_k B_{pk}H_{kt}-1}. \end{aligned} \tag{8}$$

Denoting the posterior distribution of A_{pk}, B_{pk}, and H_{kt} as $q(A_{pk})$, $q(B_{pk})$, and $q(H_{kt})$, respectively, we can decompose the log marginal likelihood of X as

$$\begin{aligned} \ln\ p(\mathbf{X}) &= \mathbf{E}_{q(\mathbf{Z})}\left[\ln \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})}\right] - \mathbf{E}_{q(\mathbf{Z})}\left[\ln \frac{p(\mathbf{Z} \mid \mathbf{X})}{q(\mathbf{Z})}\right] \\ &= \mathbf{E}_{q(\mathbf{Z})}\left[\ln \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})}\right] + \mathrm{KL}\left(\mathrm{q}(\mathbf{Z}) \| \mathrm{p}(\mathbf{Z} \mid \mathbf{X})\right). \end{aligned} \tag{9}$$

In the above equation, we approximate the true posterior distribution $p(\mathbf{Z} \mid \mathbf{X})$ by

$$q(\mathbf{Z}) \approx q(\mathbf{A})q(\mathbf{B})q(\mathbf{H}) = \prod_{p,k} q(A_{pk})q(B_{pk}) \prod_{t} q(H_{kt}). \tag{10}$$

To minimize the KL divergence from $q(\mathbf{Z})$ to $p(\mathbf{Z} \mid \mathbf{X})$ is equivalent to maximizing $\mathbf{E}_{q(\mathbf{Z})}\left[\ln \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})}\right]$, which is the objective function in the variational inference. If we consider A_{pk} as the only variable and fix the remaining variables in $\mathbf{Z}$ for a moment, the optimal solutions to $q^*(A_{pk})$ can be obtained as

$$
\begin{aligned}
\ln q^*(A_{pk}) &= \mathbf{E}_{\setminus q(A_{pk})}\left[\ln p(\mathbf{X},\mathbf{Z})\right] + \text{const} \\
&= \sum_t \mathbf{E}_{\setminus q(A_{pk})} \underbrace{\left[-\ln \mathcal{B}\left(\sum_k A_{pk}H_{kt}, \sum_k B_{pk}H_{kt}\right)\right]}_{\mathrm{F}(\mathrm{A_{p,:}},\mathrm{B_{p,:}},\mathrm{H_{:,t}})^2} \\
&\qquad + \left(\sum_t \bar{H}_{kt}\ln X_{pt}\right) A_{pk} \\
&\qquad + (\mu_0 - 1)\ln A_{pk} - \alpha_0 A_{pk} + \text{const},
\end{aligned}
\tag{11}
$$

where $\bar{x}$ denoted the expected value of x.2

The optimal solutions to $q^*(B_{pk})$ and $q^*(H_{pk})$ can be obtained in a similar way, by following the the variational inference principles. Details about these optimal solutions can be found in Appendix A.

In order to get conjugate pairs and an analytically tractable solution, we need to approximate $\ln q^*(B_{pk})$ to have the logarithmic form of the gamma distribution.

If we assume that $\mathbf{F}(\mathbf{A_{p,:}}, \mathbf{B_{p,:}}, \mathbf{H_{:,t}})$ in (1) can be approximated by an expression expressed only in terms of $\ln A_{pk}$, the inverse-scale parameter in the gamma distribution can be updated analytically as

$$
\alpha^*_{pk} = \alpha_0 - \sum_t \bar{H}_{kt}\ln X_{pt}. \tag{12}
$$

Also we have

$$
\beta^*_{pk} = \beta_0 - \sum_t \bar{H}_{kt}\ln(1 - X_{pt}). \tag{13}
$$

Similarly, the inverse-scale parameter for H_{kt} has an analytical solution as

$$
\zeta^*_{kt} = \zeta_0 - \sum_p [\bar{A}_{pk}\ln X_{pt} + \bar{B}_{pk}\ln(1 - X_{pt})]. \tag{14}
$$

One way to have an analytical tractable solution to the shape parameters is that the sum-expectation parts in above only contain $\ln A_{pk}$, $\ln B_{pk}$, and $\ln H_{pk}$, respectively. However, due to the integral expression of the gamma function $\Gamma(.)$, the expectation of $\ln\Gamma(.)$ is not analytically tractable. Thus, an analytically tractable solution can not be obtained directly.

3.3 An Analytically Tractable Solution via Extended Factorized Approximation

According to the extended factorized approximation [6], [26], [32], [34], even though we can not express the sum-expectation parts in (10), (55), and (56) directly in the form we need, we could still find an auxiliary function $\mathbf{E}_{q(\mathbf{Z})}\left[\ln \widetilde{p}(\mathbf{X},\mathbf{Z})\right]$, which satisfies

$$
\mathbf{E}_{q(\mathbf{Z})}\left[\ln p(\mathbf{X},\mathbf{Z})\right] \geq \mathbf{E}_{q(\mathbf{Z})}\left[\ln \widetilde{p}(\mathbf{X},\mathbf{Z})\right]. \tag{15}
$$

Then a lower bound to the objective function $\mathbf{E}_{q(\mathbf{Z})}\left[\ln \frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\right]$ in (9) can be obtained as

$$\mathbf{E}_{q(\mathbf{Z})}\left[\ln \frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\right] \geq \mathbf{E}_{q(\mathbf{Z})}\left[\ln \widetilde{p}(\mathbf{X},\mathbf{Z})\right] - \mathbf{E}_{q(\mathbf{Z})}\left[\ln q(\mathbf{Z})\right]. \tag{16}$$

Maximizing this lower bound is asymptotically equivalent to maximizing the objective function in (9). In this paper, we will take the EFA method to derive an analytically tractable solution to the Bayesian estimation of BG-NMF.

3.3.1 Relative Convexity

Before going through the details, we study some properties of $\mathbf{F_{pt}} \equiv \mathbf{F}(\mathbf{A}_{\mathbf{p},:}, \mathbf{F}(\mathbf{B}_{\mathbf{p},:}, \mathbf{F}(\mathbf{H}_{:,\mathbf{t}})$.

Property 3.1

The log-inverse-beta function

$$\mathrm{F_{pt}} = -\ln \mathcal{B}\left(\sum_{\mathrm{k}} \mathrm{x_k}, \sum_{\mathrm{k}} \mathrm{y_k}\right) \tag{17}$$

is convex relative to $\ln X$ for arbitrary Formula, if and only if $\sum_k y_k \geq 1$. In the above property, we used x_k and y_k to denote $A_{pk}H_{kt}$ and $B_{pk}H_{kt}$ respectively and $\mathbf{x} = [x_1, ..., x_K]^T, \mathbf{y} = [y_1, ..., y_K]^T$.

Proof

The elements of the Hessian matrix of F_{pt} in (16) with respect to $\ln X$ are

$$\mathcal{H}_{mn} = \frac{\partial^2 \mathrm{F}_{pt}}{\partial \ln x_m \ln x_n} = \begin{cases} cx_m^2 + ex_m m = n \\ cx_m x_n m \neq n, \end{cases} \tag{18}$$

where

$$\begin{aligned} c &= \psi'\left(\sum_{k=1}^{K}(x_k + y_k)\right) - \psi'\left(\sum_{k=1}^{K} x_k\right), \\ e &= \psi\left(\sum_{k=1}^{K}(x_k + y_k)\right) - \psi\left(\sum_{k=1}^{K} x_k\right). \end{aligned} \tag{19}$$

The upper-left $k \times k$ $(k = 1, ..., K)$ sub-matrix of the Hessian matrix $\mathcal{H}_{k\times k}$ is

$$\mathcal{H}_{k\times k} = c \times \mathrm{dd^T} + \mathrm{e} \times \mathrm{diag(d)}, \tag{20}$$

where

$$\mathrm{d} = [\mathrm{x_1}, \ldots, \mathrm{x_m}, \ldots, \mathrm{x_k}]^{\mathrm{T}},\ \mathrm{m} = 1, \ldots, \mathrm{k}. \tag{21}$$

The determinant of this sub-matrix is

$$\begin{aligned}\mathrm{Det}(\mathcal{H}_{\mathrm{k}\times\mathrm{k}}) &= \mathrm{Det}\left[\mathrm{e}\times\mathrm{diag}\left(\mathrm{d}\right)\right]\left\{1+\frac{\mathrm{c}}{\mathrm{e}}\left[\mathrm{d}^{\mathrm{T}}\left(\mathrm{diag}(\mathrm{d})\right)^{-1}\mathrm{d}\right]\right\} \\ &= e^{k}\times\mathrm{Det}\left[\mathrm{diag}\left(\mathrm{d}\right)\right]\left(\frac{\mathrm{e}+\mathrm{c}\times\sum_{\mathrm{m}=1}^{\mathrm{k}}\mathrm{x}_{\mathrm{m}}}{\mathrm{e}}\right).\end{aligned} \tag{22}$$

The above equation id derived by[41]

$$\mathrm{Det}(\mathbf{X}+\mathrm{cr}^{\mathrm{T}}) = \mathrm{Det}(\mathbf{X})(1+\mathrm{c}^{\mathrm{T}}\mathbf{X}^{-1}\mathrm{r}). \tag{23}$$

It has been proven that

$$\breve{x}\left\{\psi(\breve{x}+\breve{y})-\psi(\breve{x})+\breve{x}\left[\psi'(\breve{x}+\breve{y})-\psi'(\breve{x})\right]\right\}>0 \tag{24}$$

where $\breve{x}\geq 0$ and $\breve{y}\geq 1$.

If we substitute $\breve{x}=\sum_{k=1}^{K}x_k$ and $\breve{y}=\sum_{k=1}^{K}y_k$ into (24), then we have (recall that $\sum_{k=1}^{K}x_k>1$ and $\sum_{k=1}^{K}y_k>1$, since we force the beta density function to be unimodal)

$$e+c\times\sum_{m=1}^{k}x_m\geq\psi(\breve{x}+\breve{y})-\psi(\breve{x})+\breve{x}\left[\psi'(\breve{x}+\breve{y})-\psi'(\breve{x})\right]>0. \tag{25}$$

The above inequality was obtained by the facts that ψ' is a non-increasing function (then $c0$) and $\sum_{m=1}^{k}x_m\breve{x}$. As $\psi(.)$ is an increasing function, we have $c>0$. Then we can conclude that $\mathrm{Det}(\mathcal{H}_{\mathrm{k}\times\mathrm{k}})>0$. Since for any $k=1,.....,K$, the leading principal minors of the Hessian is positive, the Hessian is a positive definite matrix. Thus F_{pt} is *convex relative* to $\ln x$.

3.3.2 A Lower Bound Approximation of $\mathbf{F}_{pt}$

With this relative convexity and by restricting that $\sum_k A_{pk}H_{kt}$ and $\sum_k B_{pk}H_{kt}$ are both greater than 1, the expectation of the LIB function can be lower-bounded as

$$\begin{aligned}\mathbf{E}_{q(\mathbf{Z})}\left[\mathrm{F}_{\mathrm{pt}}\right] \geq & -\ln\mathcal{B}\left(\sum_k\bar{A}_{pk}\bar{H}_{kt},\sum_k\overline{B}_{pk}\bar{H}_{kt}\right) \\ & +\left[\psi\left(\sum_k(\bar{A}_{pk}\bar{H}_{kt}+\overline{B}_{pk}\bar{H}_{kt})\right)-\psi\left(\sum_k\bar{A}_{pk}\bar{H}_{kt}\right)\right] \\ & \times\sum_k\bar{A}_{pk}\bar{H}_{kt}\left\{\mathbf{E}_{q(A_{p,k})q(H_{k,t})}\left[\ln(A_{pk}H_{kt})\right]-\ln(\bar{A}_{pk}\bar{H}_{kt})\right\} \\ & +\left[\psi\left(\sum_k(\bar{A}_{pk}\bar{H}_{kt}+\overline{B}_{pk}\bar{H}_{kt})\right)-\psi\left(\sum_k\overline{B}_{pk}\bar{H}_{kt}\right)\right] \\ & \times\sum_k\overline{B}_{pk}\bar{H}_{kt}\{\mathbf{E}_{q(B_{p,k})q(H_{k,t})}[\ln(B_{pk}H_{kt})]-\ln(\overline{B}_{pk}\bar{H}_{kt})\} \\ & \triangleq\mathbf{E}_{q(\mathbf{Z})}[\widetilde{\mathrm{F}}_{pt}].\end{aligned} \tag{26}$$

Proof

By the relative convex property 3.1, the first-order expansion of the LIB function with respect to $\ln x$ around $\ln \bar{x}$ is a lower bound of the LIB function. Then we have the following inequality as 3

$$\begin{aligned}\mathbf{E}_{q(\mathbf{x},\mathbf{y})}\left[\mathrm{F}_{\mathrm{pt}}\right] \geq\ & \mathbf{E}_{q(\mathbf{x},\mathbf{y})}\left[-\ln \mathcal{B}\left(\sum_k \overline{x}_k, \sum_k y_k\right)\right] \\ & + \mathbf{E}_{q(\mathbf{x},\mathbf{y})}\left\{\left[\psi\left(\sum_k (\overline{x}_k + y_k)\right) - \psi\left(\sum_k \overline{x}_k\right)\right]\right. \\ & \left.\times \sum_k \overline{x}_k (\ln x_k - \ln \overline{x}_k)\right\}. \end{aligned} \tag{27}$$

As the LIB function in (27) is also *relative convex* to $\ln y$ for any x, the expectation of the LIB function can be further lower-bounded as

$$\mathbf{E}_{q(\mathbf{x},\mathbf{y})}\left[\mathrm{F}_{\mathrm{pt}}\right] \geq \mathbf{E}_{q(\mathbf{x},\mathbf{y})}\left[-\ln \mathcal{B}\left(\sum_k \overline{x}_k, \sum_k \overline{y}_k\right)\right]$$

$$+\ \mathbf{E}_{q(\mathbf{x},\mathbf{y})}\left\{\left[\psi\left(\sum_k (\overline{x}_k + \overline{y}_k)\right) - \psi\left(\sum_k \overline{y}_k\right)\right] \quad \times \sum_k \overline{y}_k (\ln y_k - \ln \overline{y}_k)\right\}$$

$$+\ \mathbf{E}_{q(\mathbf{x},\mathbf{y})}\left\{\left[\psi\left(\sum_k (\overline{x}_k + y_k)\right) - \psi\left(\sum_k \overline{x}_k\right)\right] \quad \times \sum_k \overline{x}_k (\ln x_k - \ln \overline{x}_k)\right\}.$$

In the above equation, the first term is a constant which does not contain variable x or y. The second term contains only the variable y, thus the expectation with respect to x can be ignored. The third term contains both x and y. As x and y are not mutually independent ($x_k = ApkHkt$ and $y_k = BpkHkt$ share the same H_{kt}), the expectation can not be carried out separately. However, as x_i and y_i, $i \neq j$, are mutually independent, this term can be written as in (28), where we used the fact that $q(x_k, y_k) = q(A_{pk})q(B_{pk})q(H_{kt})$.

$$
\begin{aligned}
&\mathbf{E}_{q(\mathbf{x},\mathbf{y})}\left\{\left[\psi\left(\sum_k(\overline{x}_k+y_k)\right)-\psi\left(\sum_k\overline{x}_k\right)\right]\right.\\
&\qquad\left.\times\sum_k\overline{x}_k(\ln x_k-\ln\overline{x}_k)\right\}\\
&=\sum_k\mathbf{E}_{q(x_k,y_k)}\left\{\mathbf{E}_{\setminus q(y_k)}\left[\psi\left(\sum_k(\overline{x}_k+y_k)\right)-\psi\left(\sum_k\overline{x}_k\right)\right]\right.\\
&\qquad\left.\times\overline{x}_k(\ln x_k-\ln\overline{x}_k)\right\}.\\
&=\sum_k\mathbf{E}_{q(H_{kt})}\left\{\underbrace{\mathbf{E}_{q(B_{pk}),\setminus q(y_k)}\left[\psi\left(\sum_k(\overline{x}_k+y_k)\right)-\psi\left(\sum_k\overline{x}_k\right)\right]}_{a(H_{kt})}\right.\\
&\qquad\left.\times\underbrace{\mathbf{E}_{q(A_{pk})}\left[\overline{x}_k(\ln x_k-\ln\overline{x}_k)\right]}_{b(H_{kt})}\right\}
\end{aligned}
\tag{28}
$$

For two increasing functions $a(x)$ and $b(x)$, we know that [42], [43]

$$
\mathbf{E}_{f(x)}\left[a(x)b(x)\right]\geq\mathbf{E}_{f(x)}\left[a(x)\right]\mathbf{E}_{f(x)}\left[b(x)\right],
\tag{29}
$$

where $f(x)$ is the PDF of Formula. Then the third term in equation before (28) can be lower-bounded as

$$
\begin{aligned}
&\mathbf{E}_{q(\mathbf{x},\mathbf{y})}\left\{\left[\psi\left(\sum_k(\overline{x}_k+y_k)\right)-\psi\left(\sum_k\overline{x}_k\right)\right]\right.\\
&\qquad\left.\times\sum_k\overline{x}_k(\ln x_k-\ln\overline{x}_k)\right\}\\
&\geq\sum_k\left\{\mathbf{E}_{q(\mathbf{y})}\left[\psi\left(\sum_k(\overline{x}_k+y_k)\right)-\psi\left(\sum_k\overline{x}_k\right)\right]\right.\\
&\qquad\left.\times\,\mathbf{E}_{q(x_k)}\left[\overline{x}_k\left(\ln x_k-\ln\overline{x}_k\right)\right]\right\}\\
&=\mathbf{E}_{q(\mathbf{y})}\left\{\left[\psi\left(\sum_k(\overline{x}_k+y_k)\right)-\psi\left(\sum_k\overline{x}_k\right)\right]\right\}\\
&\qquad\times\mathbf{E}_{q(\mathbf{x})}\left[\sum_k\overline{x}_k(\ln x_k-\ln\overline{x}_k)\right],
\end{aligned}
\tag{30}
$$

since both $\psi(x)$ and $\ln x$ are increasing functions. Moreover, both $\psi(x)$ and $\ln x$ are concave functions in x. So we have the following inequalities by the Jensen inequality as

$$
\begin{aligned}
\mathbf{E}_{q(x)}\left[\ln x\right]-\ln\overline{x}&\leq 0\\
\mathbf{E}_{q(x)}\left[\psi(x)\right]&\leq\psi(\overline{x}).
\end{aligned}
\tag{31}
$$

Substituting these relations into (30) and with some algebra, we have

$$\begin{aligned}
&\mathbf{E}_{q(\mathbf{x},\mathbf{y})}\left\{\left[\psi\left(\sum_k(\overline{x}_k+y_k)\right)-\psi\left(\sum_k\overline{x}_k\right)\right]\right. \\
&\qquad\left.\times\sum_k\overline{x}_k(\ln x_k-\ln\overline{x}_k)\right\} \\
&\geq\left\{\left[\psi\left(\sum_k(\overline{x}_k+\overline{y}_k)\right)-\psi\left(\sum_k\overline{x}_k\right)\right]\right\} \\
&\qquad\times\mathbf{E}_{q(\mathbf{x})}\left\{\sum_k\overline{x}_k(\ln x_k-\ln\overline{x}_k)\right\}.
\end{aligned} \tag{32}$$

Finally, the expectation of the LIB function in is lower-bounded as

$$\begin{aligned}
\mathbf{E}_{q(\mathbf{x},\mathbf{y})}\left[\mathrm{F_{pt}}\right] &\geq \mathbf{E}_{q(\mathbf{x},\mathbf{y})}\left[-\ln\mathcal{B}\left(\sum_k\overline{x}_k,\sum_k\overline{y}_k\right)\right] \\
&+\left\{\left[\psi\left(\sum_k(\overline{x}_k+\overline{y}_k)\right)-\psi\left(\sum_k\overline{y}_k\right)\right]\right. \\
&\qquad\left.\times\sum_k\overline{y}_k(\mathbf{E}_{q(\mathbf{y})}\left[\ln y_k\right]-\ln\overline{y}_k)\right\} \\
&+\left\{\left[\psi\left(\sum_k(\overline{x}_k+\overline{y}_k)\right)-\psi\left(\sum_k\overline{x}_k\right)\right]\right. \\
&\qquad\left.\times\sum_k\overline{x}_k(\mathbf{E}_{q(\mathbf{x})}\left[\ln x_k\right]-\ln\overline{x}_k)\right\}.
\end{aligned} \tag{33}$$

Thus, the lower bound approximation in (26) is proved by substituting x_k and y_k by $A_{pk}H_{kt}$ and $B_{pk}H_{kt}$, respectively.

3.3.3 Tightness of the Approximation to the LIB Function

In (26), we approximated the LIB function by a lower bound in (33). This lower bound approximation was obtained by utilizing the first-order Taylor expansion around $\bar{x}$, $\bar{y}$ and applying Jensen inequality. As several approximations were used, it is interesting to discuss the tightness of this lower bound and check if $\bar{x}$ and $\bar{y}$ are the reasonable choices for tightening the lower bound.

First, let us look at the first-order Taylor expansion around $\widetilde{\ln}x$. For the expectation of the LIB function $-\ln\mathcal{B}(x,y)$, we have the following inequality as

$$\mathbf{E}_{f(x)}\left[-\ln\mathcal{B}(x,y)\right] \;\geq \mathbf{E}_{f(x)}\Big\{-\ln\mathcal{B}(e^{\widetilde{\ln x}},y)\;+\Big[\psi(e^{\widetilde{\ln x}}+y)-\psi(e^{\widetilde{\ln x}})\Big]e^{\widetilde{\ln x}}(\ln x-\widetilde{\ln x})\Big\}.$$

Taking the derivative of the equation above (34) with respect to $\widetilde{\ln}x$ can maximize this first-order Taylor expansion. With some calculations, the optimal $\widetilde{\ln}x$ is

$$\widetilde{\ln x}^* = \mathbf{E}_{f(x)}\left[\ln x\right]. \tag{34}$$

If x is gamma distributed as

$$f(x) = \text{Gamma}(\text{x}; \mu, \alpha), \tag{35}$$

the optimal $\widetilde{\ln x}$ writes

$$\widetilde{\ln x}^* = \psi(\mu) - \ln \alpha. \tag{36}$$

Second, we study the usage of Jensen inequality in (32). As $\psi(x)$ is a concave function, we have

$$\mathbf{E}_{f(x)}\left[\psi(x)\right] \leq \mathbf{E}_{f(x)}\left[\psi(x_0) + \psi'(x_0)(x - x_0)\right]. \tag{37}$$

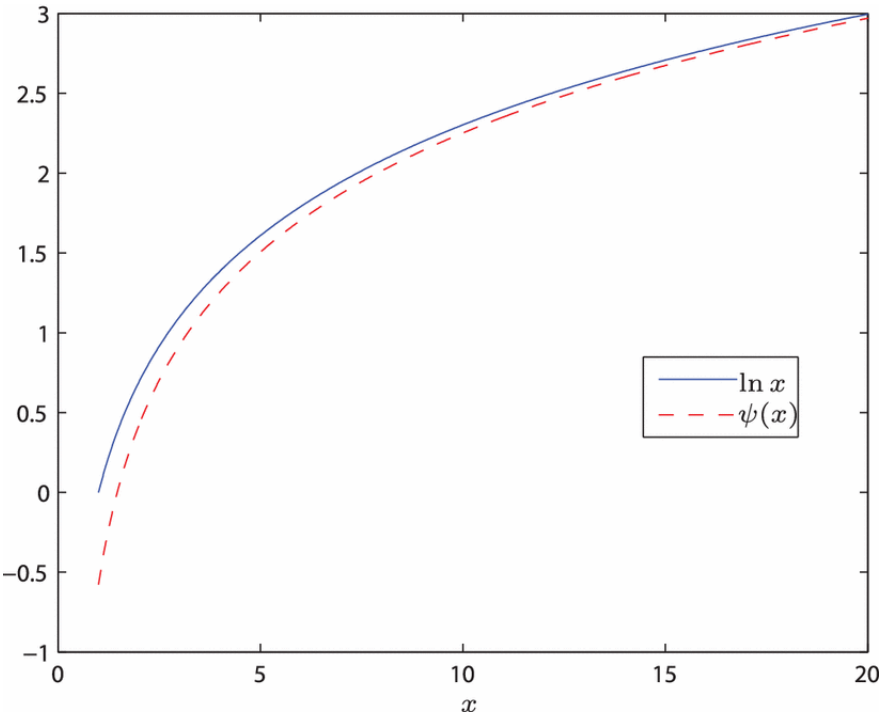
Similarly, the optimal x_0 that minimizes the first-order Taylor expansion is

$$x_0^* = \mathbf{E}_{f(x)}\left[x\right] = \overline{x} = \frac{\mu}{\alpha}. \tag{38}$$

When we take $x_0^* = \bar{x}$, (37) is exactly the same as the Jensen inequality. Thus, the first-order Taylor expansion reaches the optimal approximation when $\widetilde{\ln x}^* = \mathbf{E}_{f(x)}[\ln x]$ and the Jensen inequality for $\psi(x)$ is already optimal.

As shown in Fig. 2, $\ln x$ and $\psi(x)$ are very close to each other, especially when x becomes large, say $x \geq 5$. To simplify the expression and facilitate the calculation, we used $\ln x$ to approximate $\psi(x)$ in (36) throughout this paper. Then the optimal $\widetilde{\ln x}$ is approximated as

$$\widetilde{\ln x}^* \approx \ln \mu - \ln \alpha = \ln \overline{x}. \tag{39}$$



3.3.4 Optimal Estimation via the EFA

With (26), an auxiliary function that satisfies (15) can be obtained as

$$\mathbf{E}_{q(\mathbf{Z})}\left[\ln \widetilde{p}(\mathbf{X},\mathbf{Z})\right] = \mathbf{E}_{q(\mathbf{Z})}\left[\sum_{p,t}(\widetilde{\mathrm{F}}_{pt}+\mathrm{R}_{\mathrm{pt}})\right], \tag{40}$$

where $\mathbf{R_{pt}}$ denotes the unchanged parts in the log-likelihood function (the logarithm of(8) as

$$\begin{aligned}\mathrm{R}_{pt} &= \sum_{k}\left(A_{pk}H_{kt}-1\right)\ln X_{pt}\\ &+\sum_{k}\left(B_{pk}H_{kt}-1\right)\left(1-\ln X_{pt}\right).\end{aligned} \tag{41}$$

Combining (40) and (16) together, the objective function that we want to maximize is finally lower-bounded as

$$\begin{aligned}&\mathbf{E}_{q(\mathbf{Z})}\left[\ln\frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\right]\\ \geq\ &\mathbf{E}_{q(\mathbf{Z})}\left[\ln\widetilde{p}(\mathbf{X},\mathbf{Z})\right]-\mathbf{E}_{q(\mathbf{Z})}\left[q(\mathbf{Z})\right]\\ =\ &\mathbf{E}_{q(\mathbf{Z})}\left[\sum_{p,t}\left(\widetilde{\mathrm{F}}_{pt}+\mathrm{R}_{pt}\right)\right]-\mathbf{E}_{q(\mathbf{Z})}\left[q(\mathbf{Z})\right].\end{aligned} \tag{42}$$

In order to get the optimal solutions to Formula, Formula, and Formula, the principle of the VI framework can be applied and the optimal updates are

$$\ln q^{*}(A_{pk}) = \mathbf{E}_{\backslash q(A_{pk})}\left[\sum_{t}(\widetilde{\mathrm{F}}_{pt}+\mathrm{R}_{pt})\right]+\mathrm{const}, \tag{43}$$

$$\ln q^{*}(B_{pk}) = \mathbf{E}_{\backslash q(B_{pk})}\left[\sum_{t}(\widetilde{\mathrm{F}}_{pt}+\mathrm{R}_{pt})\right]+\mathrm{const}, \tag{44}$$

$$\ln q^{*}(H_{kt}) = \mathbf{E}_{\backslash q(H_{kt})}\left[\sum_{p}(\widetilde{\mathrm{F}}_{pt}+\mathrm{R}_{pt})\right]+\mathrm{const}, \tag{45}$$

respectively.

3.3.5 An Analytically Tractable Solution via the EFA

By skipping all the terms that do not contain A_{pk}, we can obtain an analytically tractable expression for $\ln q^{*}(A_{pk})$ as

$$\ln q^*(A_{pk}) \approx \left\{ \sum_t \left[\psi\left(\sum_k (\bar{A}_{pk} + \overline{B}_{pk}) \bar{H}_{kt} \right) - \psi\left(\sum_k \bar{A}_{pk} \bar{H}_{kt} \right) \right] \bar{A}_{pk} \bar{H}_{kt} + \mu_0 - 1 \right\} \ln A_{pk} - \left(\alpha_0 - \sum_t \bar{H}_{kt} \ln X_{pt} \right) A_{pk} + \text{const}, \tag{46}$$

which has the logarithmic forms of the gamma densities. Thus the conjugate match between the prior $p(A_{pk})$ and the posterior $q^*(A_{pk})$ is satisfied. The analytically tractable expressions for $\ln q^*(B_{pk})$ and $\ln q^*(H_{kt})$ can be obtained in similar manner.

Since $\psi(.)$ is a monotonous increasing function, the shape parameters in (46) and other similar solutions are always positive, which satisfies the definition of the gamma distribution. Furthermore, the inverse-scale parameters in (11), (12), and (13) are all positive since X_{pt} is in $(0, 1)$ With the update equations in (12)-(14) and solutions similar to (46) including it, we can update the posterior distributions of A, B, and H sequentially. Instead of maximizing the objective function directly, we maximize a lower bound of the objective function, which yields an analytically tractable approximation for $q(Z) = q(A)q(B)q(H)$ to approximate the true posterior distribution $p(Z \mid X)$.

3.3.6 Convergence

In the above sections, we factorize the latent variable Z into three disjoint groups A, B, and H. For the convenience of calculation, we introduced a single lower bound to the objective function. We then maximized this SLB, instead of the original objective function, to approximate the true posterior distribution. Maximizing this SLB is equivalent to maximizing the original one asymptotically. Furthermore, since this SLB is the only function that is maximized in every update step, convergence of the proposed algorithm is guaranteed. However, a local maximum may be reached. This effect is a general phenomenon whenever VI is employed.

3.4 The BG-NMF Algorithm

To facilitate the update, we express the update of the hyper-parameters in matrix form. The algorithm of the BG-NMF is summarized in Algorithm 1. Global optimum may not be reached because of the multi-modal property of the posterior distribution. It is observed that this objective function is non-decreasing during iterations of the proposed BG-NMF algorithm.

Input: Observation $\mathbf{X}$, number of basis K
Initialize α_0, β_0, ζ_0, μ_0, ν_0, ρ_0, $maxIter$;
Generate $\overline{\mathbf{A}}$, $\overline{\mathbf{B}}$, and $\overline{\mathbf{H}}$ from (3) as $\overline{\mathbf{A}} = \boldsymbol{\mu} \oslash \boldsymbol{\alpha}$, $\overline{\mathbf{B}} = \boldsymbol{\nu} \oslash \boldsymbol{\beta}$, $\overline{\mathbf{H}} = \boldsymbol{\rho} \oslash \boldsymbol{\zeta}$ [†].
repeat

$$\boldsymbol{\alpha} = \alpha_0 - (\ln \mathbf{X})\overline{\mathbf{H}}^{\mathrm{T}}$$
$$\boldsymbol{\mu} = \mu_0 + \left\{\psi\left[(\overline{\mathbf{A}} + \overline{\mathbf{B}})\overline{\mathbf{H}}\right] - \psi(\overline{\mathbf{A}}\,\overline{\mathbf{H}})\right\}\overline{\mathbf{H}}^{\mathrm{T}} \odot \overline{\mathbf{A}}$$ [†]
$$\boldsymbol{\beta} = \beta_0 - [\ln(1 - \mathbf{X})]\overline{\mathbf{H}}^{\mathrm{T}}$$
$$\boldsymbol{\nu} = \nu_0 + \left\{\psi\left[(\overline{\mathbf{A}} + \overline{\mathbf{B}})\overline{\mathbf{H}}\right] - \psi(\overline{\mathbf{B}}\,\overline{\mathbf{H}})\right\}\overline{\mathbf{H}}^{\mathrm{T}} \odot \overline{\mathbf{B}}$$
$$\boldsymbol{\zeta} = \zeta_0 - \overline{A}^{\mathrm{T}} \ln \mathbf{X} - \overline{\mathbf{B}}^{\mathrm{T}} \ln(1 - \mathbf{X})$$
$$\boldsymbol{\rho} = \rho_0 + \psi\left[(\overline{\mathbf{A}} + \overline{\mathbf{B}})\overline{\mathbf{H}}\right]\overline{\mathbf{H}}^{\mathrm{T}} \odot (\overline{\mathbf{A}} + \overline{\mathbf{B}}) - \psi(\overline{\mathbf{A}}\,\overline{\mathbf{H}})\overline{\mathbf{H}}^{\mathrm{T}} \odot \overline{\mathbf{A}} - \psi(\overline{\mathbf{B}}\,\overline{\mathbf{H}})\overline{\mathbf{H}}^{\mathrm{T}} \odot \overline{\mathbf{B}}$$

Optional: Calculate the objective function numerically.
until The number of iteration is equal to $maxIter$ or some criteria are reached.
Output: Hyper-parameters $\boldsymbol{\alpha}$, $\boldsymbol{\mu}$, $\boldsymbol{\beta}$, $\boldsymbol{\nu}$, $\boldsymbol{\zeta}$, and $\boldsymbol{\rho}$.
[†] $\oslash$ and $\odot$ denote element-wise division and multiplication, respectively.

3.5 Sparseness Constraints

The gamma distribution is a unimodal distribution with two parameters: the shape parameter k and the inverse-scale parameter θ. The expected value of gamma distribution is $\frac{k}{\theta}$ and the variance is $\frac{k}{\theta^2}$. When the mean value is fixed, a small shape parameter could force the variable to have a very high probability near zero, hence it favors a sparse representation of the variables. In our BG-NMF model, we can either include this sparseness constraint to the priors of the basis matrices A and B, which could make the basis matrices represent local features, or apply this constraint to the excitation matrix H such that only a few basis vectors are selected to recover the original signal.

3.6 Usage of the Proposed Method

For the beta distribution, the expected value of the variable is $\bar{x} = \frac{u}{u+v}$. Thus in this proposed generative BG-NMF model (see (3),(4)), the expected value of X_{pt} is $\bar{X}_{pt} = a_{pt}a_{pt} + b_{pt}$. If we take the point estimate to A_{pk}, B_{pk}, and H_{kt}, then the expected value of X_{pt} can be approximated as

$$\overline{X}_{pt} \approx \frac{\sum_k \bar{A}_{pk}\bar{H}_{kt}}{\sum_k \bar{A}_{pk}\bar{H}_{kt} + \sum_k \overline{B}_{pk}\bar{H}_{kt}}, \tag{47}$$

which can be expressed in matrix form as

$$\overline{\mathbf{X}} \approx (\overline{\mathbf{A}}\,\overline{\mathbf{H}}) \oslash (\overline{\mathbf{A}}\,\overline{\mathbf{H}} + \overline{\mathbf{B}}\,\overline{\mathbf{H}}), \tag{48}$$

where $\oslash$ means element-wise division.

For the purpose of visualization, we can combine A and B together to create a pseudo-basis matrix which could play a similar role as the basis matrix W (see (1)) in the conventional NMF. Generally, we have

$$\overline{\mathbf{X}} \approx (\overline{\mathbf{A}}\,\overline{\mathbf{H}}) \oslash (\overline{\mathbf{A}}\,\overline{\mathbf{H}} + \overline{\mathbf{B}}\,\overline{\mathbf{H}}) \neq \left[\overline{\mathbf{A}} \oslash (\overline{\mathbf{A}} + \overline{\mathbf{B}})\right] \overline{\mathbf{H}}. \tag{49}$$

Hence this reconstruction mentioned above is not linear in terms of $\bar{A}(\bar{A} + \bar{B})$. However, if the columns in H are highly sparse, the reconstruction in (49) could be approximated as a linear combination of $\bar{A} \oslash (\bar{A} + \bar{B})$ (if the column sparseness is 1, it is exactly linear). Thus in the experimental part, this pseudo-basis matrix s used to represent a kind of âĂIJbasisâĂİ matrix for the convenience of visualization.

$$\widehat{\mathbf{W}} = \overline{\mathbf{A}} \oslash (\overline{\mathbf{A}} + \overline{\mathbf{B}}) \tag{50}$$

4 Experimental Results and Discussion

4.1 Source seperation

We tested the ability of source separation of the BG-NMF model with sparseness constraint byreal life data evaluations. For the image data evaluation,we randomly selected three images as the source images from the Olivetti faces database and we generated a basis matrix $\mathbf{W}$ with size 4096×3, where each column represented one 4096 dimensional source with element value in [0,1] (shown below). Then a non-negative mixing matrix $\mathbf{V}$ with size 3×150 was generated. Each element in $\mathbf{V}$ is sampled from a gamma distribution with the shape parameter equal to 0.1 and the inverse-scale parameter equal to 1. Each column in $\mathbf{V}$ is normalized to be a unit vector. An observation matrix was then obtained as $\mathbf{X} = \mathbf{WV}$ and we applied the BG-NMF to separate the mixed images. The BG-NMF model was trained on the observation matrix and K was set to be 3 (assuming that we know the number of sources). For the purpose of visualization, the sparse constraint was applied to the basis matrix by setting the inverse-scale parameter and the shape parameter in the gamma prior equal to 0.0001 and 1, respectively. After convergence, the estimated basis matrix Formula was approximated by (50) and figure shows the separation performance of the BG-NMF model.

i) *left- is output after 1 iteration* ii) *right - is output after 10 iterations*

```
load('olivettifaces.mat');
numInputImages = 5;
initImage = 11;
numIter = 10;
w = faces(:,initImage:initImage-1+numInputImages)/255;
r = gamrnd(0.1,1,numInputImages,40);
y = sqrt(diag(r'*r));
z = repmat(y',size(r,1),1);
v = r./z;
```

```
y = w*v;
x = tanh(y);
shapeParam1=10^0; scaleFactor1 = 10^4;
shapeParam2=10^-1; scaleFactor2 = 10^0;
imageVecSize = size(w,1);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
alpha0 = gamrnd(shapeParam1,scaleFactor1,imageVecSize,numInputImages);
mue0 = gamrnd(shapeParam1,scaleFactor1,imageVecSize,numInputImages);
v0 = gamrnd(shapeParam1,scaleFactor1,imageVecSize,numInputImages);
beta0 = gamrnd(shapeParam1,scaleFactor1,imageVecSize,numInputImages);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
e0 = gamrnd(shapeParam2,scaleFactor2,numInputImages,40);
p0 = gamrnd(shapeParam2,scaleFactor2,numInputImages,40);

A = mue0./alpha0;
B = v0./beta0;
H = p0./e0;
for i = 1:numIter
    alpha = alpha0 - log(x)*H';
    mue  = mue0 + (log((A+B)*H) - log(A*H))*H'.*A;
    beta = beta0 - (log(1-x))*H';
    v  = v0 + (log((A+B)*H) - log(B*H))*H'.*B;
    e = e0 - A'*log(x) - B'*log(1-x);
    p = p0 + (A+B)'*(log((A+B)*H)).*H - A'*(log(A*H)).*H - ...
        B'*(log(B*H)).*H;
    A = mue./alpha;
    B = v./beta;
    H = p./e;
end
w1 = A./(A+B);
show(w1,1);
```

a) Code snippet for Source seperation problem.

4.2 Prediction of Missing Data

In this section, we apply the BG-NMF on the task of predicting missing data. We randomly selected five images from each person in the Olivetti faces database (which gives $5 \times 40 = 200$ in total) to train a model. We removed a patch from each of the remaining images and try to predict the missing part from the trained model. Due to the advantage of the Bayesian framework, the posterior distributions in the trained model are now used as the prior distributions when we predict the missing part. For an image with a patch removed, we used the remaining parts to update the generative BG-NMF model and obtain the mean value of the excitation matrix as Formula. Then the missing pixel values are reconstructed from the generative BG-NMF model as described in (47), which are the means of the corresponding beta distributions as


i) True image


ii)Images with some missing pixels

iii)BG-NMF estimation

```
load('olivettifaces.mat');
totalPixels = 4096;
imagesPerPerson  = 5;
K = 5;
iterations1 = 10;
iterations2 = 50;
z = zeros(totalPixels,1);
r = 0;
h = zeros(totalPixels,1);
g = 0;
for i = 1:40
    for j = 1:imagesPerPerson
        z = cat(2,z,faces(:,g+j));
    end
    g = g+10;
end
g = 0;
for i = 1:40
    for j = 6:10
        h = cat(2,h,faces(:,g+j));
    end
    g = g+10;
end
z = z(:,2:201);
h = h(:,2:201);
for i = 1:200
    for j = 2000:2256
        h(j,i) = 0.01;
    end
end
shapeParam1=10^0; scaleFactor1 = 10^4;
shapeParam2=10^-1; scaleFactor2 = 10^0;
x = z/255;
alpha0 = gamrnd(shapeParam1,scaleFactor1,totalPixels,K);
mue0 = gamrnd(shapeParam1,scaleFactor1,totalPixels,K);
v0 = gamrnd(shapeParam1,scaleFactor1,totalPixels,K);
beta0 = gamrnd(shapeParam1,scaleFactor1,totalPixels,K);
e0 = gamrnd(shapeParam2,scaleFactor2,K,200);
p0 = gamrnd(shapeParam2,scaleFactor2,K,200);
A = mue0./alpha0;
B = v0./beta0;
H = p0./e0;
for i = 1:iterations1
    alpha = alpha0 - log(x)*H';
    mue  = mue0 + (log((A+B)*H) - log(A*H))*H'.*A;
    beta = beta0 - (log(1-x))*H';
    v  = v0 + (log((A+B)*H) - log(B*H))*H'.*B;
```

```
48      e = e0 - A'*log(x) - B'*log(1-x);
49      p = p0 + (A+B)'*(log((A+B)*H)).*H - A'*(log(A*H)).*H - ...
            B'*(log(B*H)).*H;
50      A = mue./alpha;
51      B = v./beta;
52      H = p./e;
53  end
54  x = (h+0.01)/255;
55  alpha0 = alpha;
56  mue0 = mue;
57  v0 = v;
58  beta0 = beta;
59  e0 = e;
60  p0 = p;
61  for i = 1:iterations2
62      alpha = alpha0 - log(x)*H';
63      mue  = mue0 + (log((A+B)*H) - log(A*H))*H'.*A;
64      beta = beta0 - (log(1-x))*H';
65      v  = v0 + (log((A+B)*H) - log(B*H))*H'.*B;
66      e = e0 - A'*log(x) - B'*log(1-x);
67      p = p0 + (A+B)'*(log((A+B)*H)).*H - A'*(log(A*H)).*H - ...
            B'*(log(B*H)).*H;
68      A = mue./alpha;
69      B = v./beta;
70      H = p./e;
71  end
72  w1 = A./(A+B);
73  show(w1,1);
```

b) Code snippet for missing data problem.

where $\bar{A}$ and $\bar{B}$ are the means of the recently updated posterior distribution, S denotes the location indices of the missing pixel values. The $PSNR$ is utilized as the measure of prediction performance. I_{max} is the maximum possible value of the image and MSE is the mean squared error between the true and estimated images.Due to the boundedness constraint and assuming it to the beta distribution which is continuous has certainly provided a better results when compared to other techniques like P-NMF with gradient method and BPMF with MCMC.

References

[1] Zhanyu Ma, Andrew E.Teschendorff, Arne Leijion, Yuanyuan Qiao, Honggang Zhnag, **Variational Bayesian Matrix Factorization for Bounded Support Data** - IEEE *Transactionas on Pattern Analysis and Machine Intelligence,* VOL 37, NO. 4, APRIL 2015

[2] C.M.Bishop, **Pattern Recognition Machine Learning.** New York, NY, USA: Springer, 2006.