Variational Bayesian Matrix Factorization for Bounded Support Data

Zhanyu Ma, *Member, IEEE*, Andrew E. Teschendorff, Arne Leijon, Yuanyuan Qiao, Honggang Zhang, *Senior Member, IEEE*, and Jun Guo

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.

Index Terms—Nonnegative matrix factorization, Bayesian estimation, bounded support data, variational inference, extended factorized approximation, relative convexity, collaborative filtering, bioinformatics

1 Introduction

The nonnegative matrix factorization (NMF) was introduced by [1], [2] 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 nonnegative matrix into a product of two nonnegative matrices (a basis matrix and an excitation matrix). It is a fundamental technique for low rank nonnegative matrix approximation and has been widely used in information retrieval [3], image analysis [2], [4], source separation [5], [6], [7], [8], speech denoising [9], [10], collaborative filtering [11], [12], [13], [14], and other applications.

In the previous research, a lot of algorithms were proposed to realize the matrix factorization efficiently. By minimizing the l_2 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 [15]. To emphasize the effect of presenting the local features in the face images, the optimization with sparseness constraints was introduced in [16]. Also, assigning different weights to the vectors in the basis matrix could also improve the local representation [17].

Bayesian estimation, in general, can provide robust solution to parameter estimation. In [18], the 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 in [19], [20] where the reconstruction error $E_{pt} = X_{pt} - [\mathbf{WV}]_{pt}$ 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) [21] 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 in [5]. Furthermore, the NMF with beta divergence, which is a general form of distance measure, was introduced in [7]. These two works were applied successfully on audio/musical data source separation. In [22], 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 [5].

- *Z. Ma, H. Zhang, and J. Guo are with Pattern Recognition and Intelligent System Lab., Beijing University of Posts and Telecommunications, Beijing 100876, China. E-mail: {mazhanyu, zhhg, guojun}@bupt.edu.cn.*
- *A. E. Teschendorff is with the Computational Systems Genomics, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute for Biological Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai 200031, China, and the Statistical Genomics Group, Paul O'Gorman Building UCL Cancer Institute, University College London, 72 Huntley Street, London WC1E 6BT, United Kingdom. E-mail: a.teschendorff@ucl.ac.uk.*
- *A. Leijon is with School of Electrical and Engineering, KTH - Royal Institute of Technology, SE-100 44 Stockholm, Sweden. E-mail: leijon@kth.se.*
- *Y. Qiao is with Network Monitor and Control Laboratory, Beijing University of Posts and Telecommunications, Beijing 100876, China. E-mail: qyybupt@126.com.*

Manuscript received 6 May 2013; revised 1 Apr. 2014; accepted 6 Aug. 2014. Date of publication 3 Sept. 2014; date of current version 3 Mar. 2015. Recommended for acceptance by B. Taskar.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2014.2353639

Secondly, this paper [22] 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 $X_{pt} = \hat{X}_{pt} + E_{pt}$, where $\hat{X}_{pt}$ are Poisson distributed and E_{pt} is uniformly distributed on $[0, 1]$. A recent paper for the Bayesian treatment of the NMF derived a nonparametric Bayesian NMF with gamma process for music record processing [6]. 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 distributions.

In some real world applications, the data we are processing has bounded support. For example, the digitalized image pixels are distributed in the interval $[0, 2^R - 1]$, where R denotes the number of bits to store the pixel value. The correlations of gene-expression levels and the DNA methylation data are measured and recorded in a fixed range. The line spectral frequencies used in speech coding are strictly bounded in the range $[0, \pi]$. For topic discovery, the vector that denotes the probabilities a document belongs to all the topics has its l_1 norm equals 1, the latent Dirichlet allocation model was proposed to capture such property [23]. In order to describe the data more efficiently, some matrix factorization related work took this bounded support into account by involving link functions. Schmidt et al. [19] proposed a Bayesian treatment of NMF via Gaussian process prior (GPP-NMF), where a link function is employed to connect the nonnegativity with Gaussian prior. In principle, it is possible to derive a suitable link function (e.g., the logit function) that addresss bounded support. For the purpose of capturing the ordinal property of data, a hierarchical model for ordinal matrix factorization (OMF) was introduced in [14]. The authors applied an ordinal regression likelihood function to get the possible discrete rank so that the ordering is obtained. This OMF method can also be extended to derive a probabilistic model for bounded support data. However, all the above mentioned methods involved link function to capture the bounded support. So far as we know, there is no matrix factorization strategy proposed *directly* for bounded support data, without using any link function. This motivates us to study the matrix factorization method for bounded support data.

Several researches have shown that, compared to some conventional used statistical model, e.g., Gaussian distribution, the beta distribution can model the bounded support data more efficiently and lead to better performance in many applications [24], [25], [26], [27], [28], [29]. In this paper, a Bayesian matrix factorization method for bounded support data is presented. We assume a generative model such that each bounded support element in the observation matrix is generated from a beta distribution. Different from the conventional NMF methods which applied the factorization directly on the observed matrix, we apply the NMF strategy to the parameter matrices of the beta distribution. There are two parameter matrices for the beta distribution. As all the elements in the parameter matrix are nonnegative, each parameter matrix is nonnegatively factorized into the product of a basis matrix and an excitation matrix. To handle the correlation between these two parameter matrices, the excitation matrix is chosen to be the same for both factorizations. Each entry in the basis matrices and the excitation matrices is assigned with a gamma prior. Therefore, we name the proposed Bayesian matrix factorization method as beta-gamma NMF (BG-NMF). By the relative convexity [30], [31] of the log-inverse-beta (LIB) function, we approximate the objective function with a single lower bound (SLB). Thus, a single function is maximized during each update step and the convergence of the proposed is guaranteed. The tightness of the lower bound approximation is also discussed. With the variational inference framework [32], [33] and the methodology of the extended factorized approximation (EFA) [6], [26], [32], [34], [35], [36], [37], [38], an optimal solution to approximately calculate the posterior distribution can be obtained in an analytically tractable form. This solution retains the conjugate match between the prior and the posterior distribution, which is favorable in practice. In addition, a sparse BG-NMF can also be obtained by including a sparseness constraint to the gamma prior. The performance of the proposed method was evaluated with both synthesized data and several real-life applications in source separation, collaborative filtering, and bioinformatics areas.

The rest of this paper is organized as follows: the Bayesian NMF is introduced in Section 2. For the bounded support data, we introduce a generative Bayesian NMF model in Section 3. Also, the parameter estimation method and the corresponding algorithm are proposed. In Section 4, the experimental evaluations and comparisons are shown. Finally, some conclusions are drawn in Section 5.

2 Nonnegative 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 nonnegative values X_{pt}, W_{pk}, and V_{kt} respectively and $p = 1, \ldots, P$, $t = 1, \ldots, T$, $k = 1, \ldots, K$. Usually, we choose $K < T$ such that the NMF is a low rank matrix approximation. $\mathbf{W}$ and $\mathbf{V}$ are usually named as the basis matrix and the excitation matrix, respectively. Denoting the tth column in $\mathbf{X}$ as $\mathbf{x}_t$, we have that $\mathbf{x}_t$ 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

As shown in literature [24], [25], [26], [27], [28], [29], [39], the bounded support data (usually defined in the interval $[x_l, x_h]$, 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 The 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 [6, Eq. (1)], 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 nonnegative, we assign a gamma prior to each entry. With the above description, we assume that the matrix $\mathbf{X}$ (with element $X_{pt} \in [0, 1]$) is drawn according to the following generative model[1]

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

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

$$\mathrm{Gamma}(x; k, \theta) = \frac{\theta^k}{\Gamma(k)} x^{k-1} e^{-\theta x},\ k, \theta > 0, \tag{4}$$

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

$$\mathrm{Beta}(x; u, v) = \frac{1}{\mathcal{B}(u, v)} x^{u-1}(1-x)^{v-1},\ u, v > 0, \tag{5}$$

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

1. When the beta distribution is unimodally distributed, both the parameters are greater than 1. This is a typical case in practical problems and we only study this case in this paper. To this end, we assume the probability that $\sum_k A_{pk}H_{kt} < 1$ is very small (almost close to zero), which is similar as that in [26]. The same assumption applies to $\sum_k B_{pk}H_{kt}$.

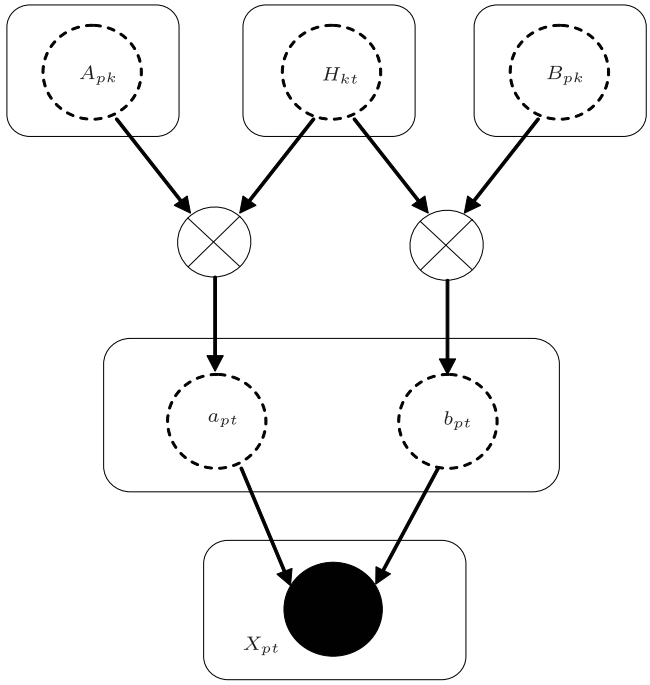


Fig. 1. Graphical model of the BG-NMF. All the dashed circles in the graphical figure represent variables. A_{pk}, B_{pk}, and H_{kt} are assumed to be gamma distributed. X_{pt} is assumed to be beta distributed with parameter a_{pt} and b_{pt}. Arrows show the relationship between variables. The variables in the box are independent from each other.

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 (VI) [32], [33], [34], [35], we have already divided the latent variables $\mathbf{Z} = \{\mathbf{A}, \mathbf{B}, \mathbf{H}\}$ into disjoint groups $\mathbf{A}$, $\mathbf{B}$, $\mathbf{H}$ and assigned a gamma prior to each entry in those matrices (see Section 3.1). Thus the prior distributions of the latent variables are

$$\begin{aligned}p(\mathbf{A}) &= \prod_{p,k} p(A_{pk}),\\ p(\mathbf{B}) &= \prod_{p,k} p(B_{pk}),\\ p(\mathbf{H}) &= \prod_{k,t} p(H_{kt}),\\ p(\mathbf{Z}) &= p(\mathbf{A})p(\mathbf{B})p(\mathbf{H}).\end{aligned} \tag{6}$$

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

$$\begin{aligned}p(\mathbf{X}\,|\,\mathbf{Z}) = \prod_{p,t} &\frac{1}{\mathcal{B}\big(\sum_k A_{pk}H_{kt}, \sum_k B_{pk}H_{kt}\big)}\\ &\times (X_{pt})^{\sum_k A_{pk}H_{kt}-1}\big(1-X_{pt}\big)^{\sum_k B_{pk}H_{kt}-1}.\end{aligned} \tag{7}$$

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 $\mathbf{X}$ as

$$\begin{aligned}\ln\ p(\mathbf{X}) &= \mathbf{E}_{q(\mathbf{Z})}\left[\ln\frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\right] - \mathbf{E}_{q(\mathbf{Z})}\left[\ln\frac{p(\mathbf{Z}\,|\,\mathbf{X})}{q(\mathbf{Z})}\right] \\ &= \mathbf{E}_{q(\mathbf{Z})}\left[\ln\frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\right] + \mathrm{KL}(q(\mathbf{Z})\|p(\mathbf{Z}\,|\,\mathbf{X})).\end{aligned} \tag{8}$$

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

$$q(\mathbf{Z}) \approx q(\mathbf{A})q(\mathbf{B})q(\mathbf{H}) = \prod_{p,k} q(A_{pk})q(B_{pk}) \prod_{t} q(H_{kt}). \tag{9}$$

To minimize the KL divergence from $q(\mathbf{Z})$ to $p(\mathbf{Z}\,|\,\mathbf{X})$ is equivalent to maximizing $\mathbf{E}_{q(\mathbf{Z})}[\ln\frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}]$, which is the objective function in the variational inference [33]. 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

$$\begin{aligned}\ln q^*(A_{pk}) &= \mathbf{E}_{\backslash q(A_{pk})}[\ln p(\mathbf{X},\mathbf{Z})] + \text{const} \\ &= \sum_t \mathbf{E}_{\backslash q(A_{pk})} \underbrace{\left[-\ln\mathcal{B}\left(\sum_k A_{pk}H_{kt}, \sum_k B_{pk}H_{kt}\right)\right]}_{\mathrm{F}(A_{p,:},B_{p,:},H_{:,t})^2} \\ &\quad + \left(\sum_t \overline{H}_{kt}\ln X_{pt}\right)A_{pk} \\ &\quad + (\mu_0 - 1)\ln A_{pk} - \alpha_0 A_{pk} + \text{const},\end{aligned} \tag{10}$$

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

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 $\ln q^*(A_{pk})$ to have the logarithmic form of the gamma distribution.

If we assume that $\mathrm{F}(A_{p,:}, B_{p,:}, H_{:,t})$ in (10) 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

$$\alpha^*_{pk} = \alpha_0 - \sum_t \overline{H}_{kt}\ln X_{pt}. \tag{11}$$

Also, we have

$$\beta^*_{pk} = \beta_0 - \sum_t \overline{H}_{kt}\ln(1 - X_{pt}). \tag{12}$$

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

$$\zeta^*_{kt} = \zeta_0 - \sum_p [\bar{A}_{pk}\ln X_{pt} + \bar{B}_{pk}\ln(1 - X_{pt})]. \tag{13}$$

2. $A_{p,:}$ means a row vector which is the pth row of $\mathbf{A}$. $B_{p,:}$ means a row vector which is the pth row of $\mathbf{B}$. Similarly, $H_{:,t}$ is the tth column of $\mathbf{H}$.

One way to have an analytical tractable solution to the shape parameters is that the sum-expectation parts in (10), (55), and (56) only contain $\ln A_{pk}$, $\ln B_{pk}$, and $\ln H_{pk}$, respectively. However, due to the integral expression of the gamma function $\Gamma(\cdot)$, the expectation of $\ln\Gamma(\cdot)$ 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 $\mathbf{E}_{q(\mathbf{Z})}[\ln\widetilde{p}(\mathbf{X},\mathbf{Z})]$, which satisfies

$$\mathbf{E}_{q(\mathbf{Z})}[\ln p(\mathbf{X},\mathbf{Z})] \geq \mathbf{E}_{q(\mathbf{Z})}[\ln\widetilde{p}(\mathbf{X},\mathbf{Z})]. \tag{14}$$

Then a lower bound to the objective function $\mathbf{E}_{q(\mathbf{Z})}[\ln\frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}]$ in (8) can be obtained as

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

Maximizing this lower bound is asymptotically equivalent to maximizing the objective function in (8) [26], [32], [34], [40]. 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 $\mathrm{F}_{pt} \triangleq \mathrm{F}(A_{p,:}, B_{p,:}, H_{:,t})$.

Property 3.1. *The log-inverse-beta function*

$$\mathrm{F}_{pt} = -\ln\mathcal{B}\left(\sum_k x_k, \sum_k y_k\right) \tag{16}$$

is convex relative to [30], [36] $\ln\mathbf{x}$ *for arbitrary* $\mathbf{y}$, *if and only if* $\sum_k y_k > 1$. *In the above property, we used* x_k *and* y_k *to denote* $A_{pk}H_{kt}$ *and* $B_{pk}H_{kt}$ *respectively and* $\mathbf{x} = [x_1, \ldots, x_K]^{\mathrm{T}}$, $\mathbf{y} = [y_1, \ldots, y_K]^{\mathrm{T}}$.

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

$$\mathcal{H}_{mn} = \frac{\partial^2\mathrm{F}_{pt}}{\partial\ln x_m \ln x_n} = \begin{cases} cx_m^2 + ex_m & m = n \\ cx_m x_n & m \neq n, \end{cases} \tag{17}$$

where

$$\begin{aligned} c &= \psi'\left(\sum_{k=1}^K (x_k + y_k)\right) - \psi'\left(\sum_{k=1}^K x_k\right), \\ e &= \psi\left(\sum_{k=1}^K (x_k + y_k)\right) - \psi\left(\sum_{k=1}^K x_k\right). \end{aligned} \tag{18}$$

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

$$\mathcal{H}_{k\times k} = c \times \mathrm{d}\mathrm{d}^{\mathrm{T}} + e \times \mathrm{diag}(\mathrm{d}), \tag{19}$$

where

$$\mathrm{d} = [x_1, \ldots, x_m, \ldots, x_k]^{\mathrm{T}},\ m = 1, \ldots, k. \tag{20}$$

The determinant of this sub-matrix is

$$\begin{aligned}\mathrm{Det}(\mathcal{H}_{k\times k}) &= \mathrm{Det}[e \times \mathrm{diag}(\mathrm{d})]\Big\{1 + \frac{c}{e}\Big[\mathrm{d}^{\mathrm{T}}(\mathrm{diag}(\mathrm{d}))^{-1}\mathrm{d}\Big]\Big\} \\ &= e^k \times \mathrm{Det}[\mathrm{diag}(\mathrm{d})]\left(\frac{e + c \times \sum_{m=1}^{k} x_m}{e}\right).\end{aligned} \tag{21}$$

The above equation is derived by [41]

$$\mathrm{Det}(\mathbf{X} + cr^{\mathrm{T}}) = \mathrm{Det}(\mathbf{X})(1 + c^T\mathbf{X}^{-1}r). \tag{22}$$

It has been proven (see [26, Appendix A]) that

$$\breve{x}\{\psi(\breve{x} + \breve{y}) - \psi(\breve{x}) + \breve{x}[\psi'(\breve{x} + \breve{y}) - \psi'(\breve{x})]\} > 0, \tag{23}$$

when $\breve{x} > 0$ and $\breve{y} > 1$.

If we substitute $\breve{x} = \sum_{k=1}^{K} x_k$ and $\breve{y} = \sum_{k=1}^{K} y_k$ into (23), 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)

$$e + c \times \sum_{m=1}^{k} x_m \geq \psi(\breve{x} + \breve{y}) - \psi(\breve{x}) + \breve{x}[\psi'(\breve{x} + \breve{y}) - \psi'(\breve{x})] > 0. \tag{24}$$

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

3.3.2 A Lower Bound Approximation to 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

$$\begin{aligned}\mathbf{E}_{q(\mathbf{Z})}\big[\mathrm{F}_{pt}\big] \geq\ & -\ln \mathcal{B}\left(\sum_k \overline{A}_{pk}\overline{H}_{kt}, \sum_k \overline{B}_{pk}\overline{H}_{kt}\right) \\ &+ \left[\psi\left(\sum_k (\overline{A}_{pk}\overline{H}_{kt} + \overline{B}_{pk}\overline{H}_{kt})\right) - \psi\left(\sum_k \overline{A}_{pk}\overline{H}_{kt}\right)\right] \\ &\times \sum_k \overline{A}_{pk}\overline{H}_{kt}\Big\{\mathbf{E}_{q(A_{p,k})q(H_{k,t})}\big[\ln(A_{pk}H_{kt})\big] - \ln(\overline{A}_{pk}\overline{H}_{kt})\Big\} \\ &+ \left[\psi\left(\sum_k (\overline{A}_{pk}\overline{H}_{kt} + \overline{B}_{pk}\overline{H}_{kt})\right) - \psi\left(\sum_k \overline{B}_{pk}\overline{H}_{kt}\right)\right] \\ &\times \sum_k \overline{B}_{pk}\overline{H}_{kt}\{\mathbf{E}_{q(B_{p,k})q(H_{k,t})}[\ln(B_{pk}H_{kt})] - \ln(\overline{B}_{pk}\overline{H}_{kt})\} \\ \triangleq\ & \mathbf{E}_{q(\mathbf{Z})}[\widetilde{\mathrm{F}}_{pt}].\end{aligned} \tag{25}$$

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

$$\begin{aligned}\mathbf{E}_{q(\mathbf{x},\mathbf{y})}\big[\mathrm{F}_{pt}\big] \geq\ & \mathbf{E}_{q(\mathbf{x},\mathbf{y})}\left[-\ln \mathcal{B}\left(\sum_k \overline{x}_k, \sum_k y_k\right)\right] \\ &+ \mathbf{E}_{q(\mathbf{x},\mathbf{y})}\left\{\left[\psi\left(\sum_k (\overline{x}_k + y_k)\right) - \psi\left(\sum_k \overline{x}_k\right)\right]\right. \\ &\left.\times \sum_k \overline{x}_k(\ln x_k - \ln \overline{x}_k)\right\}.\end{aligned} \tag{26}$$

As the LIB function in (26) is also *relative convex to* $\ln \mathbf{y}$ for *any* $\mathbf{x}$, the expectation of the LIB function can be further lower-bounded as

$$\mathbf{E}_{q(\mathbf{x},\mathbf{y})}\big[\mathrm{F}_{pt}\big] \geq \mathbf{E}_{q(\mathbf{x},\mathbf{y})}\left[-\ln \mathcal{B}\left(\sum_k \overline{x}_k, \sum_k \overline{y}_k\right)\right] \tag{27a}$$

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

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

In the above equation, the first term (27a) is a constant which does not contain variable $\mathbf{x}$ or $\mathbf{y}$. The second term (27b) contains only the variable $\mathbf{y}$, thus the expectation with respect to $\mathbf{x}$ can be ignored. The third term (27c) contains both $\mathbf{x}$ and $\mathbf{y}$. As $\mathbf{x}$ and $\mathbf{y}$ are *not* mutually independent ($x_k = A_{pk}H_{kt}$ and $y_k = B_{pk}H_{kt}$ share the same H_{kt}), the expectation can not be carried out separately. However, as x_i and y_j, $i \neq j$, are mutually independent, this term can be written as in (28), where we used the fact that $q(x_k, y_k) = q(A_{pk})q(B_{pk})q(H_{kt})$.

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

3. Recall that we denote $x_k = A_{pk}H_{kt}$ and $y_k = B_{pk}H_{kt}$. The vector $\mathbf{x}$ is $\mathbf{x} = [x_1, x_2, \ldots, x_K]^{\mathrm{T}}$ and $\mathbf{y}$ is $\mathbf{y} = [y_1, y_2, \ldots, y_K]^{\mathrm{T}}$.

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

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

where $f(x)$ is the PDF of x. Then the third term (27c) can be lower-bounded as

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

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's inequality as

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

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

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

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

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

Thus, the lower bound approximation in (25) 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 (25), 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 $\overline{\mathbf{x}}$, $\overline{\mathbf{y}}$ and applying Jensen's inequality. As several approximations were used, it is interesting to discuss the tightness of this lower bound and check if $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ are the reasonable choices for tightening the lower bound.

First, let's 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

$$\begin{aligned}
&\mathbf{E}_{f(x)}[-\ln\mathcal{B}(x,y)]\\
&\quad\geq\mathbf{E}_{f(x)}\Big\{-\ln\mathcal{B}(e^{\widetilde{\ln x}},y).\\
&\qquad+\Big[\psi(e^{\widetilde{\ln x}}+y)-\psi(e^{\widetilde{\ln x}})\Big]e^{\widetilde{\ln x}}(\ln x-\widetilde{\ln x})\Big\}.
\end{aligned}\tag{34}$$

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

$$\widetilde{\ln x^*}=\mathbf{E}_{f(x)}[\ln x].\tag{35}$$

If x is gamma distributed as

$$f(x)=\mathrm{Gamma}(x;\mu,\alpha),\tag{36}$$

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

$$\widetilde{\ln x^*}=\psi(\mu)-\ln\alpha.\tag{37}$$

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

$$\mathbf{E}_{f(x)}[\psi(x)]\leq\mathbf{E}_{f(x)}[\psi(x_0)+\psi'(x_0)(x-x_0)].\tag{38}$$

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

$$x_0^*=\mathbf{E}_{f(x)}[x]=\overline{x}=\frac{\mu}{\alpha}.\tag{39}$$

When we take $x_0^*=\overline{x}$, (38) is exactly the same as the Jensen's inequality. Thus, the first-order Taylor expansion reaches the optimal approximation when $\widetilde{\ln x^*}=\mathbf{E}_{f(x)}[\ln x]$ and the Jensen's 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 (37) throughout this paper. Then the optimal $\widetilde{\ln x}$ is approximated as

$$\widetilde{\ln x^*}\approx\ln\mu-\ln\alpha=\ln\overline{x}.\tag{40}$$

In summary, taking the first-order Taylor expansions around $\ln\overline{x}$ and applying the Jensen's inequality are the nearly optimal choices for LIB approximation.

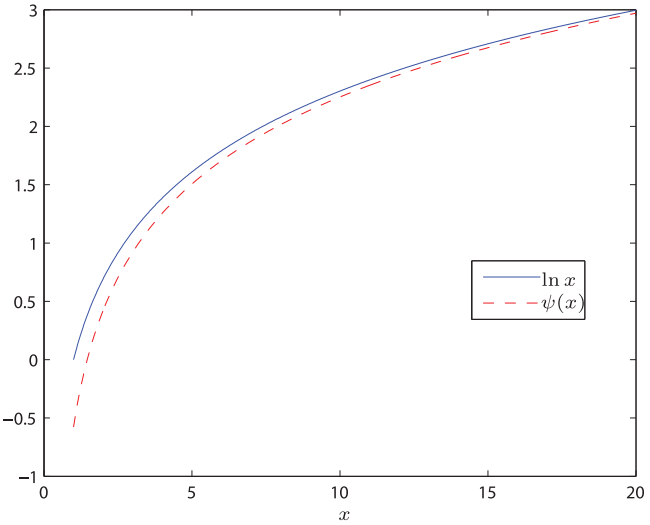


Fig. 2. Comparison of $\ln x$ and $\psi(x)$.

To illustrate the accuracy of the approximation, we list the comparisons between the true expectation of the LIB function and the approximating one (see (25)) in Table 1. In this comparison, we generate x and y, 100,000 samples for each variable, from a known Gamma distribution. The true expectation of the LIB function $\mathbf{E}_{q(x,y)}[\mathrm{F}_{pt}]$ is calculated numerically. Meanwhile, the approximating one $\mathbf{E}_{q(x,y)}[\widetilde{F}_{pt}]$ is also calculated with the mean values of x and y and the expected values of $\ln x$ and $\ln y$, respectively. It can be observed that the proposed approximation to the expectation of the LIB function works well under different conditions.

3.3.4 Optimal Estimation via the EFA

With (25), an auxiliary function that satisfies (14) can be obtained as

$$\mathbf{E}_{q(\mathbf{Z})}[\ln \widetilde{p}(\mathbf{X},\mathbf{Z})] = \mathbf{E}_{q(\mathbf{Z})}\left[\sum_{p,t}(\widetilde{\mathrm{F}}_{pt}+\mathrm{R}_{pt})\right], \tag{41}$$

where R_{pt} denotes the unchanged parts in the log-likelihood function (the logarithm of (7)) as

$$\begin{aligned}\mathrm{R}_{pt} = &\sum_{k}\big(A_{pk}H_{kt}-1\big)\ln X_{pt}\\ &+\sum_{k}\big(B_{pk}H_{kt}-1\big)(1-\ln X_{pt}).\end{aligned} \tag{42}$$

Combining (41) and (15) together, the objective function that we want to maximize is finally lower-bounded as

$$\begin{aligned}&\mathbf{E}_{q(\mathbf{Z})}\left[\ln\frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}\right]\\ &\geq\ \mathbf{E}_{q(\mathbf{Z})}[\ln\widetilde{p}(\mathbf{X},\mathbf{Z})]-\mathbf{E}_{q(\mathbf{Z})}[q(\mathbf{Z})]\\ &=\ \mathbf{E}_{q(\mathbf{Z})}\left[\sum_{p,t}\big(\widetilde{\mathrm{F}}_{pt}+\mathrm{R}_{pt}\big)\right]-\mathbf{E}_{q(\mathbf{Z})}[q(\mathbf{Z})].\end{aligned} \tag{43}$$

In order to get the optimal solutions to $q^*(A_{pk})$, $q^*(B_{pk})$, and $q^*(H_{kt})$, the principle of the VI framework [33] can be applied and the optimal updates are

TABLE 1
Comparisons of the Approximation Accuracy of the LIB Function

	$\mathbf{E}_{q(x,y)}[\mathrm{F}_{pt}]$	$\mathbf{E}_{q(x,y)}[\widetilde{F}_{pt}]$	SNR (in dB)
$x\sim\mathrm{Gamma}(x;4,3)$ $y\sim\mathrm{Gamma}(y;8,5)$	0.6319	0.6008	26.15
$x\sim\mathrm{Gamma}(x;40,30)$ $y\sim\mathrm{Gamma}(y;80,50)$	0.8358	0.8335	51.10
$x\sim\mathrm{Gamma}(x;400,300)$ $y\sim\mathrm{Gamma}(y;800,500)$	0.8562	0.8559	71.41

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

$$\ln q^*(B_{pk}) = \mathbf{E}_{\setminus q(B_{pk})}\left[\sum_{t}(\widetilde{\mathrm{F}}_{pt}+\mathrm{R}_{pt})\right]+\mathrm{const}, \tag{45}$$

and

$$\ln q^*(H_{kt}) = \mathbf{E}_{\setminus q(H_{kt})}\left[\sum_{p}(\widetilde{\mathrm{F}}_{pt}+\mathrm{R}_{pt})\right]+\mathrm{const}, \tag{46}$$

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

$$\begin{aligned}\ln q^*(A_{pk}) \ \approx\ &\left\{\sum_{t}\left[\psi\left(\sum_{k}(\overline{A}_{pk}+\overline{B}_{pk})\overline{H}_{kt}\right)\right.\right.\\ &\left.\left.-\psi\left(\sum_{k}\overline{A}_{pk}\overline{H}_{kt}\right)\right]\overline{A}_{pk}\overline{H}_{kt}+\mu_0-1\right\}\ln A_{pk}\\ &-\left(\alpha_0-\sum_{t}\overline{H}_{kt}\ln X_{pt}\right)A_{pk}+\mathrm{const},\end{aligned} \tag{47}$$

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, which are provided in Appendix B.

Since $\psi(\cdot)$ is a monotonous increasing function, the shape parameters in (47) and (B) 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 $(0,1)$.[4] With the update equations in (11)-(13) and (47)-(B), we can update the posterior distributions of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{H}$ sequentially. Instead of maximizing the objective function directly, we maximize a lower bound of the objective function, which yields an analytically tractable approximation for $q(\mathbf{Z})=q(\mathbf{A})q(\mathbf{B})q(\mathbf{H})$ to approximate the true posterior distribution $p(\mathbf{Z}\,|\,\mathbf{X})$.

3.3.6 Convergence

In the above sections, we factorize the latent variable $\mathbf{Z}$ into three disjoint groups $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{H}$. For the convenience of

4. To avoid the infinity quantity in the practical implementation, we assign ε_1 to X_{pt} when $X_{pt}=0$ and $1-\varepsilon_2$ to X_{pt} when $X_{pt}=1$. Both ε_1 and ε_2 are slightly positive real numbers.

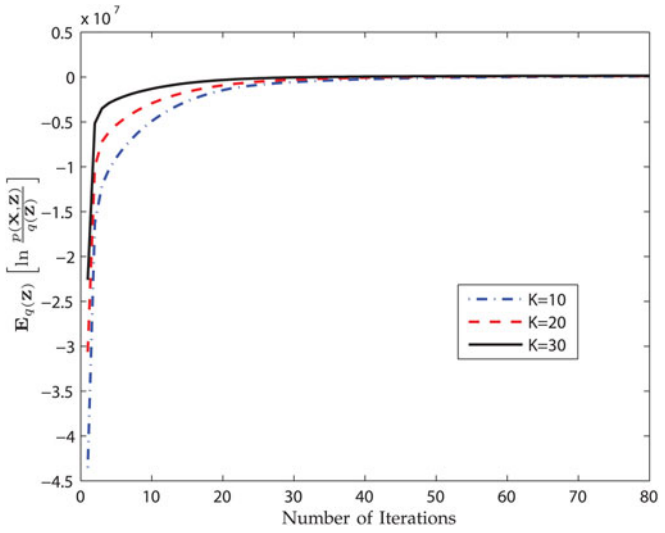


Fig. 3. Illustration of the convergence of the BG-NMF algorithm. We randomly selected 80 images from the Olivetti faces database [46] and downsampled them to size 32×32. We set $K = 10$, 20, and 30. The algorithm could always converge after about $60 \sim 80$ rounds of iterations. The objective function is numerically calculated by generating samples from the posterior distributions.

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. As shown in [26], [40], 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 [26], [33], [40]. This effect is a general phenomenon whenever VI is employed [33].

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. Fig. 3 illustrates the convergence of the proposed BG-NMF algorithm.

3.5 Computational Complexity

For Bayesian estimation framework, either the Markov chain Monte Carlo (MCMC) method (e.g., Gibbs sampling) can be employed to numerically simulate the posterior distributions of the parameters or analytically tractable solution can be derived by introducing lower-bound approximation with the principles of variational inference. Generally speaking, the VI procedure has several advantages. Firstly, it circumvents sampling from high-dimensionally multinomial variables, which is the main computational bottleneck with the Gibbs sampler. Secondly, it is straightforward to calculate the hyper-parameters which avoid maximizing the marginal likelihood via Monte Carlo EM procedure [44], [45]. Hence, we only compare the computational complexity of proposed BG-NMF method with other methods which use link function and simple Bayesian matrix factorization approach.

Algorithm 1. BG-NMF

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.

According to the algorithm presented in Algorithm 1, the operations of the proposed BG-NMF algorithm are mainly based on matrix multiplication. Therefore, the computational complexity is linear $w.r.t.$ the dimension of samples P, the amount of samples T or the dimension of the reduced feature space K. For the OMF method [14] which involves link function, both Gibbs sampling-based method and VI-based method were proposed. The VI-based method is aiming for estimate the posterior distribution of the mean and the covariance matrix in the multi-variate Gaussian density function. The computational cost for the hyper-parameters estimation are mainly spent on matrix multiplication, which are also linear to the size of data matrix. In the BG-NMF method, parameters from three matrices (two basis matrices and one excitation matrix) are required to estimate while the parameters to be estimated in [14] are from two matrix (one basis matrix and one excitation matrix). Hence, given the same number of iterations, the scale of computational demand for BG-NMF is the same as that required in the OMF method, but the practical computational cost of BG-NMF might be slightly higher. When applying suitable link function (e.g., logit transform), the GPP-NMF introduced in [19] can be applied for bounded support data as well. By ignoring the computational cost of the link function, the main computational cost comes from the MAP estimation of $\boldsymbol{\delta}$ and $\boldsymbol{\eta}$ [19, Section 2.6]. Thus, the overall computational cost of this method mainly depends on the optimization method used in MAP estimation, which usually contains gradient search. Hence, we speculate that the computational cost of the BG-NMF method is less than the GPP-NMF.

3.6 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 k/θ and the variance is k/θ^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.

3.7 Usage of the Proposed Method

For the beta distribution, the expected value of the variable is $\overline{x} = \frac{u}{u+v}$. Thus in this proposed generative BG-NMF model (see (3)), the expected value of X_{pt} is $\overline{X}_{pt} = \frac{a_{pt}}{a_{pt}+b_{pt}}$. 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

$$\overline{X}_{pt} \approx \frac{\sum_k \overline{A}_{pk}\overline{H}_{kt}}{\sum_k \overline{A}_{pk}\overline{H}_{kt} + \sum_k \overline{B}_{pk}\overline{H}_{kt}}, \tag{48}$$

which can be expressed in matrix form as

$$\overline{\mathbf{X}} \approx (\overline{\mathbf{A}}\ \overline{\mathbf{H}}) \oslash (\overline{\mathbf{A}}\ \overline{\mathbf{H}} + \overline{\mathbf{B}}\ \overline{\mathbf{H}}), \tag{49}$$

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

$$\overline{\mathbf{X}} \approx (\overline{\mathbf{A}}\ \overline{\mathbf{H}}) \oslash (\overline{\mathbf{A}}\ \overline{\mathbf{H}} + \overline{\mathbf{B}}\ \overline{\mathbf{H}}) \neq \left[\overline{\mathbf{A}} \oslash (\overline{\mathbf{A}} + \overline{\mathbf{B}})\right]\overline{\mathbf{H}}. \tag{50}$$

Hence this reconstruction mentioned above is not linear in terms of $\overline{\mathbf{A}}\oslash(\overline{\mathbf{A}} + \overline{\mathbf{B}})$. However, if the columns in **H** are highly sparse, the reconstruction in (50) could be approximated as a linear combination of $\overline{\mathbf{A}} \oslash (\overline{\mathbf{A}} + \overline{\mathbf{B}})$ (if the column sparseness is 1, it is exactly linear). Thus in the experimental part, this pseudo-basis matrix

$$\widehat{\mathbf{W}} = \overline{\mathbf{A}} \oslash (\overline{\mathbf{A}} + \overline{\mathbf{B}}) \tag{51}$$

is used to represent a kind of "basis" matrix for the convenience of visualization.

4 Experimental Results and Discussion

The BG-NMF model is proposed for bounded support data. We have conducted several experiments to demonstrate the performance of this proposed method.

First, in 4.1 and 4.2, we apply our BG-NMF model to the Olivetti faces database [46], which contains 400 human face images in 8 bits gray scale. The 400 face images are from 40 persons and each person has ten face images. We compressed the pixel value linearly to $[0, 1]$ by dividing each pixel value by 255. Also, we downsampled each image from the size 64×64 to the size 32×32. Each image was then rearranged into a column vector with dimension 1,024. The point estimate from the BG-NMF model (see (49)) was considered as the reconstructed signal $\widehat{\mathbf{X}}$. We compared our BG-NMF with different NMF methods and took the peak signal-to-noise ratio

$$\mathrm{PSNR} = 10 \log_{10} \frac{\mathrm{I}_{\max}}{\mathrm{MSE}} \tag{52}$$

as the objective criterion. Here, MSE is the mean square error between the true and estimated images defined as

(a) Source (b) Estimated Source

Fig. 4. Synthetic source separation example. Each column is rearranged into a 6×6 matrix for visual clarity. The order of the matrices has also been rearranged for the purpose of easy comparison. See Section 4.1 for more details.

$$\mathrm{MSE} = \frac{1}{PT}\sum_{p,t}(X_{pt} - \widehat{X}_{pt})^2 \tag{53}$$

and $\mathrm{I}_{\max}$ denotes the maximum possible value in an image.

Secondly, in Section 4.3, the proposed BG-NMF model is applied to the Netflix problem. We consider the movie ratings from 1 to 5 as sampled from a continuous variable and model it by the beta distribution. The BG-NMF is used to model the relations between the movie indices and the ratings given by different reviewers. With the obtained model, we predict the missing values by the generative framework introduced in Section 3.1.

Thirdly, we apply the BG-NMF model in DNA methylation analysis in Section 4.4. The proposed BG-NMF model is used in retrieving the components of variation associated with normal/cancer status. Also, it serves as an efficient tool in dimension reduction.

4.1 Source Separation

We tested the ability of source separation of the BG-NMF model with sparseness constraint by synthesized data and real life data evaluations. For the synthesized data evaluation, we generated a basis matrix **W** with size 36×6, where each column represented one 36 dimensional source with element value in [0,1] (shown in Fig. 4a). Then a nonnegative mixing matrix **V** with size 6×150 was generated. Each element in **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 **V** is normalized to be a unit vector. An observation matrix was then obtained as $\mathbf{X} = \mathbf{WV}$. The BG-NMF model was trained on the observation matrix and K was set to be six (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 $\widehat{\mathbf{W}}$ was approximated by (51) and shown in Fig. 4b.

Furthermore, for the real life data evaluation, we randomly selected three images as the source images from the Olivetti faces database. The basis matrix **W** is of size $1{,}024 \times 3$. A mixing matrix **V** of size 3×40 was generated from a gamma distribution with the shape parameter equal to 0.1 and the inverse-scale parameter equal to 1. Then each column in the mixing matrix was normalized to be a unit vector. The source images were mixed by the mixing matrix to obtain a observation matrix $\mathbf{X} = \mathbf{WV}$. We applied the BG-NMF to separate the mixed images

(a) The source images (upper panel) and the estimated source images (lower panel). (b) Examples of mixed images.

Fig. 5. Source separation with images. See Section 4.1.

with K set to be three. Fig. 5 shows the separation performance of the BG-NMF model.

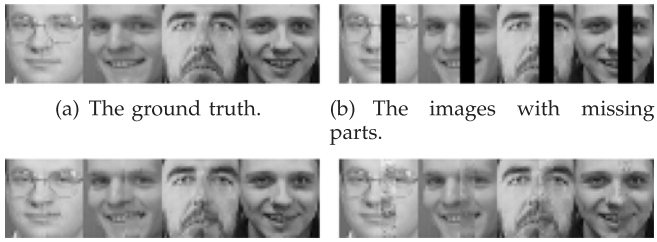
4.2 Prediction of Missing Data

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 $\check{\mathbf{H}}$. Then the missing pixel values are reconstructed from the generative BG-NMF model as described in (48), which are the means of the corresponding beta distributions as

$$\widehat{\mathbf{X}}_{ij} = \frac{\sum_k \check{A}_{ik}\check{H}_{kj}}{\sum_k \check{A}_{ik}\check{H}_{kj} + \sum_k \check{B}_{ik}\check{H}_{kj}}, \ (i,j) \in S, \tag{54}$$

where $\check{A}$ and $\check{B}$ 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. We compare the BG-NMF with the P-NMF [22],[5] which is a recently proposed Bayesian NMF method focusing on image processing. In total, 20 rounds of simulations were done for each model and the mean result were reported. The PSNR for the reconstruction obtained by BG-NMF is 21.94 dB while the PSNR obtained by P-NMF is 19.44 dB. Fig. 6 shows the prediction performance of different methods with some examples. It can be observed that the BG-NMF can predict the missing part better than the P-NMF. The P-NMF assumed that the data was generated from a discrete source, which makes itself weak at describing the continuous data. Also, applying a semi-bounded distribution to describe the distribution of the bounded support data suffers from the mismatch between the model and the source. Thus, we believe that the proposed BG-NMF, which is based on the Bayesian framework, is more suitable for the continuous bounded support data. In this part, the sparse constraint was the same as that in Section 4.1.

5. We ran the P-NMF method on the original image pixel value, which is in $[0, 255]$.



(a) The ground truth. (b) The images with missing parts. (c) BG-NMF. (d) P-NMF (with ICM).

Fig. 6. Examples from the missing pixels prediction. See Section 4.2.

4.3 Collaborative Filtering

Collaborative filtering is used to predict a person's preferences by using other people's preferences in some recommendation systems. Generally speaking, the commonly used CF methods can be categorized into two types: the memory-based method and the model-based method [47]. The memory-based method measures the similarities between persons [48]. The model-based method, on the other hand, can investigate and recognize the patterns hidden in the database and make an prediction from a probabilistic perspective, given the pre-learned model [48], [49]. Generally speaking, the model-based CF method performs better than the memory-based method because it addresses the sparsity better, improves the prediction performance, and gives an intuitive rationale for recommendations [47].

The Netflix problem [50] is such a problem.[6] The data set is divided into two parts: a training matrix and a probing matrix. These two matrices have the same sizes. In each matrix, the row denotes different movies and the column denote different reviewers. The element in the matrix is the rating. The ratings that appear in the training matrix will not appear in the probing matrix. All the ratings are integers from 1 to 5. The task of the Netflix problem is to predict the missing ratings given some reviewers' existing rating behaviors. It is worth to note that, the data matrix in the Netflix problem is highly sparse, as not all the movies have ratings and not all the reviewers have scored all the movies.

The proposed BG-NMF model fits the Netflix problem nicely. Firstly, it seeks for a low-rank matrix approximation, which can reduce the model complexity. Secondly, the BG-NMF model captures the data's bounded property (ratings are from 1 to 5), and, therefore, the prediction performance is improved. Finally, the BG-NMF model can address the fact the the Netflix data matrix is highly sparse by involving a proper prior information. To apply the BG-NMF model to the Netflix problem, we scale each rating from the interval $[1, 5]$ to the interval $(0, 1)$ to fit the beta distribution. The scaling is carried out by $y = \frac{x-1+\delta}{4+2\delta}$, where x is the rating value. We set $\delta = 0.5$ empirically. Since training data set $\mathbf{X}$ is highly sparse, we set both the basis and excitation matrices with sparseness constraints. To handle the missing values, we introduce a mask matrix $\mathbf{M}$, which has the same size as the training matrix $\mathbf{X}$. Each entry in $\mathbf{M}$ is either 0 or 1, indicating

6. The Netflix contest was finished. Unfortunately, we cannot access to the original database. However, a subset ($6{,}040 \times 3{,}952$) of the original database is available at http://www.mit.edu/~rsalakhu/BPMF.html. The following evaluations and comparisons were carried out based on this subset.

TABLE 2
Comparisons of BG-NMF, PMF, BPMF, and Sparse PCA

Method	Probing RMSE
BG-NMF with point estimate	0.8624
PMF with gradient method	0.8787
BPMF with MCMC	0.8406
sparse PCA with point estimate	0.8665

We ran 20 rounds of simulations with $K = 20$ for each method and the mean values are reported.

the value is present or not in the corresponding position in $\mathbf{X}$. Then we replace $\ln \mathbf{X}$ with $\mathbf{M} \odot \ln \mathbf{X}$ and replace $\ln(\mathbf{1} - \mathbf{X})$ with $\mathbf{M} \odot \ln(\mathbf{1} - \mathbf{X})$ respectively in Algorithm 1. With the estimated parameters, we predicted the missing values in $\mathbf{X}$. The root mean squared error (RMSE) between the predicted ratings and the true ratings is used to evaluate the prediction performance.

Table 2 shows the RMSE obtained by the BG-NMF, the probabilistic matrix factorization (PMF) [12], Bayesian PMF (BPMF) [11] and sparse PCA [13]. The BPMF and sparse PCA methods were proposed recently and shown to be efficient for this problem. The smaller the RMSE is, the better the prediction is. It can be observed that the BG-NMF performs better than both the PMF and the sparse PCA. This is because 1) the PMF utilized the gradient method, which is not a Bayesian framework; 2) the sparse PCA (with Gaussian assumption) did not consider or utilize the bounded support property of the data. Unlike the BG-NMF or the sparse PCA, which use the posterior point estimate, the BPMF applied MCMC algorithm to generate samples, which could (in principle) simulate the true posterior distribution more accurately than the other approximation methods. This is probably why the BPMF performs the best among all the methods, even though it does not consider the bounded property either. Generally speaking, the proposed BG-NMF is an efficient method for the Netflix problem.

4.4 Cancer Epigenomics Analysis

In order to demonstrate another practical utility of BG-NMF, we considered its application to the analysis of DNA methylation data, which is naturally beta distributed (see e.g., [28]). Currently, there is a lot of interest in DNA methylation as a key regulatory mechanism of gene expression, since DNA methylation patterns are widely altered in many complex genetic diseases, including cancer [51], [52]. DNA methylation is naturally quantified by bounded support data, and although standard dimensional reduction methods like NMF, singular value decomposition (SVD), and PCA have been applied to DNA methylation data [53], most of these standard methods do not take the bounded support nature of the data into account. Thus, an attractive feature of BG-NMF is the ability to perform dimensional reduction on fairly large beta distributed data matrices, retaining the beta distributed distribution in the pseudo-basis matrix as in (51). This then allows the option for further clustering algorithms designed for beta distributed data to be applied, such as the Recursive Partitioning Beta-Mixture Model (RPBMM) presented in [54].

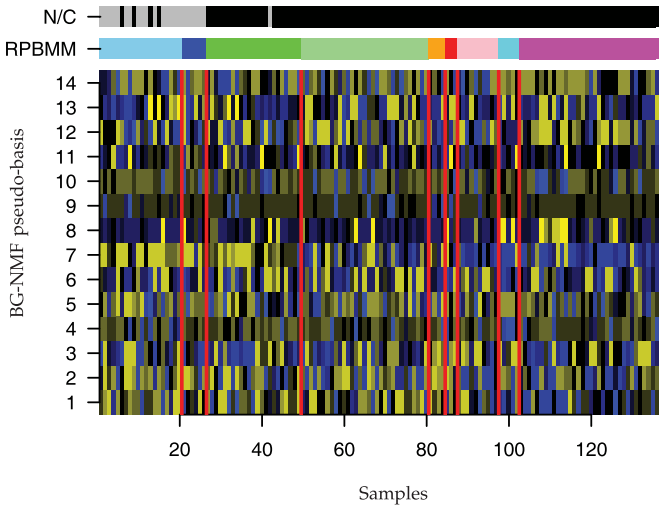


Fig. 7. RPBMM clustering results based on the transposed BG-NMF pseudo-basis matrix. The $1st$ bar from the top is the normal (grey) and breast cancer (black) examples. The clustering results are shown in the $2nd$ bar from the top. The heatmap shows the pseudo-basis matrix of the BG-NMF. The heatmap itself has been standardized so that blue indicates high level ($\sim$1), and yellow denotes low level ($\sim$0).

We considered a DNA methylation data matrix over 5,000 features (specifically, CpG dinucleotides) and 136 breast tissue samples, which was generated using a modern beadarray platform interrogating 27,578 CpGs [55]. The 5,000 CpGs were selected as those with the highest variance across the 136 samples. Thus, each entry in the data matrix represents the methylation level (bounded between 0 and 1) of a CpG in a given sample. Of the 136 samples, 23 were normal (healthy) specimens, with the remaining 113 representing breast cancers. Normal and cancer samples are known to be well discriminated at the level of DNA methylation [52], [56], hence application of BG-NMF to this data matrix provides a benchmark to assess if BG-NMF can efficiently retrieve the components of variation associated with normal/cancer status. Cancers specially are also known to be highly heterogeneous [57], hence the expectation is that BG-NMF can retrieve some of this heterogeneity.

In applying BG-NMF, we need to specify the dimensionality, i.e., the number of components to search for. To obtain an estimate of the dimensionality, we used Random Matrix Theory (RMT) [58]. Although data is distinctly non-normal (even after mean-centering each CpG), RMT provides a reasonable approximation of the dimensionality as shown by us previously [59]. For our data matrix of $136 \times 5{,}000$, we estimated a total of 14 dimensions out of the total 5,000 dimensions. Thus, setting the number of basis vectors equal to 14 and applying BG-NMF to this data matrix resulted in a 136×14 pseudo-basis matrix and a $14 \times 5{,}000$ excitation matrix. The hypothesis is that the dimensionally reduced pseudo-basis matrix, whose element remains bounded supported and is assumed to be beta distributed, captures the salient patterns of variation. To assess this, we used RPBMM to cluster the 136 samples over the 14 BG-NMF pseudo-bases. RPBMM inferred a total of nine clusters (see Fig. 7), which correlated significantly with normal/cancer status as assessed using the Adjusted Rand Index (ARI), an index designed to evaluate the concordance of two partitions. Although the ARI value was small ($ARI = 0.12$) this only

TABLE 3
RPBMM Clustering Details for all the 136 Samples over 14 BG-NMF Pseudo-Basis Vectors

Cluster	Normal	Cancer
rLLLL	16	4
rLLLR	6	0
rLLR	1	32
rLR	0	31
rRLLL	0	4
rRLLR	0	3
rRLRL	0	10
rRLRR	0	5
rRR	0	34

reflects the relatively large number of clusters inferred (i.e., 9), and indeed the ARI value was highly significant as assessed using 1,000 randomizations of phenotype labels ($P < 0.001$, i.e., in none of the 1,000 randomization we observed an ARI value as large as 0.12). The exact distribution of normal/cancer samples among the 9 clusters is shown in Table 3. Thus, we can see from both Fig. 7 and Table 3 that normal and cancer samples are well discriminated, but also that breast cancers are highly heterogeneous, as expected.

Finally, we compared these results with what would have been obtained had we used RPBMM directly on the $136 \times 5{,}000$ data matrix. Without prior dimensional reduction using BG-NMF, RPBMM also predicted nine clusters with a similar ARI value ($ARI = 0.11$, $P < 0.001$). However, while RPBMM on the excitation matrix only took 23 seconds to run on a modern Dell Precision Workstation, on the full $136 \times 5{,}000$ data matrix it took about 55 times longer (1,275 seconds). Our BG-NMF implementation on the full data matrix took about 116 seconds, hence the overall efficiency gain of using BG-NMF prior to RPBMM was by a factor of 8. Thus, we can conclude that BG-NMF can not only retrieve biologically relevant patterns of variation in DNA methylation data, but most importantly that it provides a more efficient means of dimensional reduction. This is an important consideration given that future applications will require effective dimensional reduction and clustering on even larger DNA methylation data matrices.

5 Conclusion

To explicitly utilize the bounded support property of the data, a new Bayesian matrix factorization model, the beta-gamma nonnegative matrix factorization (BG-NMF) model, was proposed for the continuous data with bounded support. The data distribution is described by the beta density function and each of the parameters in the beta density function is assigned with a gamma prior. By approximating the objective function with a single lower bound, an analytically tractable solution was obtained. With this solution, we can approximately calculate the posterior distribution of the parameters in the BG-NMF model. The approximated posterior distribution for each parameter is also gamma distributed. Therefore, the conjugacy of the model is retained. With the estimated posterior distributions, the original matrix could be reconstructed efficiently. This BG-NMF approach can be used in several important applications, such as source separation, collaborative filtering, and cancer epigenomics analysis. By comparing the proposed method with some recently introduced and widely used Bayesian matrix factorization methods, we demonstrated the good performance of the proposed BG-NMF model.

Appendix A
Optimal Solutions to $q^*(B_{pk})$ and $q^*(H_{kt})$

With the principles of variational inference, the optimal solution for $q^*(B_{pk})$ is

$$\begin{aligned}\ln q^*(B_{pk}) &= \mathbf{E}_{\setminus q(B_{pk})}[\ln p(\mathbf{X},\mathbf{Z})] + \text{const} \\ &= \sum_t \mathbf{E}_{\setminus q(B_{pk})} \underbrace{\left[-\ln \mathcal{B}\left(\sum_k A_{pk}H_{kt}, \sum_k B_{pk}H_{kt}\right)\right]}_{\mathrm{F}(A_{p,:},B_{p,:},H_{:,t})} \\ &\quad + \left[\sum_t \overline{H}_{kt}\ln(1-X_{pt})\right] B_{pk} \\ &\quad + (\nu_0 - 1)\ln B_{pk} - \beta_0 B_{pk} + \text{const.}\end{aligned} \tag{55}$$

When considering H_{kt} as the variable, the optimal solution for $q^*(H_{pk})$ is

$$\begin{aligned}\ln q^*(H_{kt}) &= \mathbf{E}_{\setminus q(H_{kt})}[\ln p(\mathbf{X},\mathbf{Z})] + \text{const} \\ &= \sum_p \mathbf{E}_{\setminus q(H_{kt})} \underbrace{\left[-\ln \mathcal{B}\left(\sum_k A_{pk}H_{kt}, \sum_k B_{pk}H_{kt}\right)\right]}_{\mathrm{F}(A_{p,:},B_{p,:},H_{:,t})} \\ &\quad + \left[\sum_p \overline{A}_{pk}\ln X_{pt} + \sum_p \overline{B}_{pk}\ln(1-X_{pt})\right] H_{kt} \\ &\quad + (\rho_0 - 1)\ln H_{kt} - \zeta_0 H_{kt} + \text{const.}\end{aligned} \tag{56}$$

Appendix B
Analytically Tractable Solutions for $q^*(B_{pk})$ and $q^*(H_{kt})$

With the EFA method, we can obtain the posterior distribution in gamma PDF form. Hence, an analytically tractable expression for $\ln q^*(B_{pk})$ can be obtained as

$$\begin{aligned}\ln q^*(B_{pk}) \approx &\left\{\sum_t \left[\psi\left(\sum_k (\overline{A}_{pk} + \overline{B}_{pk})\overline{H}_{kt}\right)\right.\right. \\ &\left.\left. - \psi\left(\sum_k \overline{B}_{pk}\overline{H}_{kt}\right)\right]\overline{B}_{pk}\overline{H}_{kt} + \nu_0 - 1\right\}\ln B_{pk} \\ &- \left[\beta_0 - \sum_t \overline{H}_{kt}\ln(1-X_{pt})\right] B_{pk} + \text{const.}\end{aligned} \tag{57}$$

With a similar approach, we can approximate the optimal solution of $q^*(H_{kt})$ by a gamma distribution as

$$
\begin{aligned}
&\ln q^*(H_{kt}) \\
&\approx \left\{ \sum_p \left[\psi\left(\sum_k (\overline{A}_{pk} + \overline{B}_{pk}) \overline{H}_{kt} \right) (\overline{A}_{pk} + \overline{B}_{pk}) \overline{H}_{kt} \right.\right. \\
&\left.\left. - \psi\left(\sum_k \overline{A}_{pk} \overline{H}_{kt} \right) \overline{A}_{pk} \overline{H}_{kt} - \psi\left(\sum_k \overline{B}_{pk} \overline{H}_{kt} \right) \overline{B}_{pk} \overline{H}_{kt} \right] \right\} \ln H_{kt} \\
&+ \left[\sum_p \overline{A}_{pk} \ln X_{pt} + \sum_p \overline{B}_{pk} \ln(1 - X_{pt}) \right] H_{kt} + (\rho_0 - 1) \\
&\ln H_{kt} - \zeta_0 H_{kt} + \text{const.}
\end{aligned}
$$

ACKNOWLEDGMENTS

The authors would like to thank Professor Cédric Févotte for his fruitful discussion and suggestions. This work was partly supported by the National Nature Science Foundation of China Grant No. 61402047, No. 61175011, No. 61273217, Chinese 111 program of Advanced Intelligence and Network Service Grant No. B08004, and EU FP7 IRSES MobileCloud Project Grant No. 612212.

REFERENCES

[1] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, pp. 111–126, 1994.
[2] D. D. Lee and H. S. Seung, "Learning the parts of objects with non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.
[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.
[4] D. Guillamet and J. Vitrià, "Non-negative matrix factorization for face recognition," in *Topics in Artificial Intelligence*, ser. Lecture Notes in Computer Science, vol. 2504. New York, NY, USA: Springer, 2002, pp. 336–344.
[5] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, 2009.
[6] M. Hoffman, D. Blei, and P. Cook, "Bayesian nonparametric matrix factorization for recorded music," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 439–446.
[7] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *Neural Comput.*, vol. 23, no. 9, pp. 2421–2456, Sep. 2011.
[8] S. Kirbiz and P. Smaragdis, "An adaptive time-frequency resolution approach for non-negative matrix factorization based single channel sound source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2011, pp. 253–256.
[9] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 4029–4032.
[10] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising," in *Proc. Interspeech Int. Conf. Spoken Lang. Process.*, 2008, pp. 411–414.
[11] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization using Markov chain Monte Carlo," in *Proc. Int. Conf. Mach. Learn.*, 2008, pp. 880–887.
[12] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1257–1264.
[13] T. Raiko, A. Ilin, and J. Karhunen, "Principal component analysis for sparse high-dimensional data," in *Neural Information Processing*, M. Ishikawa, K. Doya, H. Miyamoto, and T. Yamakawa, Eds, Berlin, Germany: Springer-Verlag, 2008, pp. 566–575.
[14] U. Paquet, B. Thomson, and O. Winther, " A hierarchical model for ordinal matrix factorization," *Statist. Comput.*, vol. 22, no. 4, pp. 945–957, 2012.
[15] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst. 13*, 2001, pp. 556–562.
[16] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, Dec. 2004.
[17] D. Guillamet, M. Bressan, and J. Vitrià, "A weighted non-negative matrix factorization for local representations," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, 2001, pp. 942–947.
[18] P. A. Højen-Sørensen, O. Winther, and L. K. Hansen, "Mean-field approaches to independent component analysis," *Neural Comput.*, vol. 14, no. 4, pp. 889–918, 2002.
[19] M. N. Schmidt and H. Laurberg, "Nonnegative matrix factorization with Gaussian process priors," *Comput. Intell. Neurosci.*, vol. 2008, p. 8, 2008.
[20] M. N. Schmidt, O. Winther, and L. K. Hansen, "Bayesian non-negative matrix factorization," in *Independent Component Analysis and Signal Separation*, ser. Lecture Notes in Computer Science. New York, NY, USA: Springer, 2009.
[21] J. Besag, "On the statistical analysis of dirty pictures," *J. Royal Statist. Soc.*, vol. 48, no. 3, pp. 259–302, 1986.
[22] A. T. Cemgil, "Bayesian inference in non-negative matrix factorisation models," *Comput. Intell. Neurosci.*, vol. 2009, p. 4, July 2009.
[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.
[24] Y. Ji, C. Wu, P. Liu, J. Wang, and K. R. Coombes, "Application of beta-mixture models in bioinformatics," *Bioinformat. Appl. Note*, vol. 21, pp. 2118–2122, 2005.
[25] Z. Ma and A. Leijon, "PDF-optimized LSF vector quantization based on beta mixture models," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 2374–2377.
[26] Z. Ma and A. Leijon, "Bayesian estimation of beta mixture models with variational inference," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2160–2173, Nov. 2011.
[27] Z. Ma, "Non-Gaussian statistical models and their applications," Ph.D. dissertation, KTH-Royal Institute of Technology, Stockholm, Sweden, 2011.
[28] A. E. Teschendorff, F. Marabita, M. Lechner, T. Bartlett, J. Tegner, D. Gomez-Cabrero, and S. Beck, "A beta-mixture quantile normalisation method for correcting probe design bias in illumina infinium 450k DNA methylation data," *Bioinformatics*, vol. 29, no. 2, pp. 189–196, Nov. 2012.
[29] Z. Ma and A. E. Teschendorff, "A variational Bayes beta mixture model for feature selection in DNA methylation studies," *J. Bioinformat. Comput. Biol.*, vol. 11, no. 4, p. 1350005, 2013.
[30] J. A. Palmer, "Relative convexity," ECE Dept., Univ. California, San Diego, San Diego, CA, USA, Tech. Rep., 2003.
[31] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[32] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, no. 2, pp. 183–233, 1999.
[33] C. M. Bishop, *Pattern Recognition Machine Learning*. New York, NY, USA: Springer, 2006.
[34] T. S. Jaakkola and M. I. Jordan, "Bayesian parameter estimation via variational methods," *Statist. Comput.*, vol. 10, pp. 25–37, 2000.
[35] T. S. Jaakkola, "Tutorial on variational approximation methods," in *Proc. Adv. Mean Field Methods*, 2001, pp. 129–159.
[36] D. M. Blei and J. D. Lafferty, "Correlated topic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006.
[37] D. M. Blei and J. D. Lafferty, "A correlated topic model of Science," *Ann. Appl. Statist.*, vol. 1, pp. 17–35, 2007.
[38] M. Braun and J. McAuliffe, "Variational inference for large-scale models of discrete choice," *J. Amer. Statist. Assoc.*, vol. 105, pp. 324–335, 2010.
[39] N. Bouguila, D. Ziou, and E. Monga, "Practical Bayesian estimation of a finite beta mixture through Gibbs sampling and its applications," *Statist. Comput.*, vol. 16, pp. 215–225, 2006.
[40] Z. Ma, P. K. Rana, J. Taghia, M. Flierl, and A. Leijon, "Bayesian estimation of Dirichlet mixture model with variational inference," *Pattern Recognit.*, vol. 47, no. 9, pp. 3143–3157, Sep. 2014.
[41] M. Brookes, *The Matrix Reference Manual*, 2005, http://www.psi.toronto.edu/matrix/intro.html
[42] K. D. Schmidt, "On the covariance of monotone functions of a random variable," Lehrstuhl für Versicherungsmathematik, Technische Universität Dresden, Tech. Rep., 2003.
[43] M. Egozcue, L. F. Garcia, and W. K. Wong, "On some covariance inequalities for monotonic and non-monotonic functions," *J. Inequalities Pure Appl. Math.*, vol. 10, pp. 1–16, 2009.

[44] F. A. Quintana, J. S. Liu, and G. E. Pino, "Monte Carlo em with importance reweighting and its applications in random effects models," *Comput. Statist. Data Anal.*, vol. 29, no. 4, pp. 429–444, Feb. 1999.
[45] G. Moffa and J. Kuipers, "Sequential Monte Carlo em for multivariate probit models," *Comput. Statist. Data Anal.*, vol. 72, no. 0, pp. 252–272, 2014.
[46] [Online]. Available: http://cs.nyu.edu/~roweis/data.html
[47] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Adv. Artif. Intell.*, vol. 2009, p. 4, Jan. 2009.
[48] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," in *Proc. 14th Conf. Uncertainty Artif. Intell.*, 1998, pp. 43–52.
[49] J. L. Herlocker, J. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 5–53, Jan. 2004.
[50] [Online]. Available: http://www.netflixprize.com/
[51] A. P. Feinberg, R. Ohlsson, and S. Henikoff, "The epigenetic progenitor origin of human cancer," *Nat. Rev. Genetics*, vol. 7, pp. 21–33, Jan. 2006.
[52] P. A. Jones and S. B. Baylin, "The epigenomics of cancer," *Cell*, vol. 128, no. 4, pp. 683–692, Feb. 2007.
[53] J. Zhuang, M. Widschwendter, and A. E. Teschendorff, "A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform," *BMC Bioinformatics*, vol. 13, p. 59, 2012.
[54] E. A. Houseman, B. C. Christensen, R. F. Yeh, C. J. Marsit, M. R. Karagas, M. Wrensch, H. H. Nelson, J. Wiemels, S. Zheng, J. K. Wiencke, and K. T. Kelsey, "Model-based clustering of DNA methylation array data: A recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions," *Bioinformatics*, vol. 9, p. 365, 2008.
[55] M. Bibikova, Z. Lin, L. Zhou, E. Chudin, E. W. Garcia, B. Wu, D. Doucet, N. J. Thomas, Y. Wang, E. Vollmer, T. Goldmann, C. Seifart, W. Jiang, D. L. Barker, M. S. Chee, J. Floros, and J. B. Fan, "High-throughput DNA methylation profiling using universal bead arrays," *Genome Res.*, vol. 16, no. 3, pp. 383–393, Mar. 2006.
[56] M. Widschwendter, H. Fiegl, D. Egle, E. Mueller-Holzner, G. Spizzo, C. Marth, D. J. Weisenberger, M. Campan, J. Young, I. Jacobs, and P. W. Laird, "Epigenetic stem cell signature in cancer," *Nat. Genetics*, vol. 39, no. 2, pp. 157–158, Feb 2007.
[57] C. Curtis, S. P. Shah, S. F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Grf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, A. Langerod, A. Green, E. Provenzano, G. W. G, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A. L. Borresen-Dale, J. D. Brenton, S. Tavare, C. Caldas, and S. Aparicio, "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature*, vol. 486, no. 7403, pp. 346–352, Apr. 2012.
[58] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. Amaral, T. Guhr, and H. E. Stanley, "Random matrix approach to cross correlations in financial data," *Phys. Rev. E: Statist., Nonlinear, and Soft Matter Phys.*, vol. 65, p. 066126, Jun. 2002.
[59] A. E. Teschendorff, J. Zhuang, and M. Widschwendter, "Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies," *Bioinformatics*, vol. 27, no. 11, pp. 1496–1505, Jun. 2011.

Zhanyu Ma received the MEng degree in signal and information processing from Beijing University of Posts and Telecommunications (BUPT), China, and the PhD degree in electrical engineering from Royal Institute of Technology (KTH), Sweden, in 2007 and 2011, respectively. He has been an assistant professor at the Beijing University of Posts and Telecommunications, Beijing, China, since 2013. From 2012 to 2013, he has been a Postdoctoral research fellow in the School of Electrical Engineering, KTH, Sweden. His research interests include pattern recognition and machine learning fundamentals with a focus on applications in multimedia signal processing, data mining, biomedical signal processing, and bioinformatics.

Andrew E. Teschendorff received his PhD degree in theoretical particle physics at the University of Cambridge, United Kingdom. He now leads the Statistical Cancer Genomics group at the UCL Cancer Institute, University College London, London, United Kingdom. His research interests includes statistical genomics and epigenomics with a focus on applications to cancer, as well as network physics.

Arne Leijon received the MS degree in engineering physics in 1971, and the PhD degree in information theory in 1989, both from Chalmers University of Technology, Gothenburg, Sweden. He is a professor in hearing technology at the Royal Inst of Technology (KTH) Sound and Image Processing Lab, Stockholm, Sweden, since 1994. His main research interest concerns applied signal processing in aids for people with hearing impairment, and methods for individual fitting of these aids, based on psychoacoustic modelling of sensory information transmission and subjective sound quality.

Yuanyuan Qiao received the BE degree from Xidian University in 2009. She is an assistant professor in the School of Information and Communication Engineering, BUPT. Her research focuses on broadband IP network, traffic measurement and classification, mobile Internet traffic analysis and cloud computing optimization and management.

Honggang Zhang received the BS degree from the Department of Electrical Engineering, Shandong University in 1996, the master's and PhD degree from the School of Information Engineering, Beijing University of Posts and Telecommunications (BUPT) in 1999 and 2003, respectively. He worked as a visiting scholar in School of Computer Science, Carnegie Mellon University (CMU) from 2007-2008. He is currently an associate professor and director of web search center at BUPT. His research interests include image retrieval, computer vision and pattern recognition. He published more than 40 papers on TPAMI, SCIENCE, Machine Vision and Applications, Neurocomputing. He is a senior member of IEEE.

Jun Guo received the BE and ME degrees from Beijing University of Posts and Telecommunications (BUPT), China in 1982 and 1985, respectively, and the PhD degree from the Tohuku-Gakuin University, Japan in 1993. He is currently a professor and a vice president of BUPT. His research interests include pattern recognition theory and application, information retrieval, content based information security, and bioinformatics. He has published more than 200 papers on the journals and conferences including SCIENCE, Nature Scientific Reports, *IEEE Transactions on PAMI, Pattern Recognition, AAAI, CVPR, ICCV, SIGIR*, etc.