mathematics

MDPI

Article

Probabilistic Non-Negative Matrix Factorization with Binary Components

Xindi Ma [1], Jie Gao [1], Xiaoyu Liu [2,*], Taiping Zhang [1] and Yuanyan Tang [3]

[1] College of Computer Science, Chongqing University, Chongqing 400044, China; xindima@cqu.edu.cn (X.M.); jie.gao@cqu.edu.cn (J.G.); tpzhang@cqu.edu.cn (T.Z.)
[2] National Center for Applied Mathematics in Chongqing, Chongqing Normal University, Chongqing 400044, China
[3] Zhuhai UM Science & Technology Research Institute, Zhuhai 519000, China; yytang@umac.mo
* Correspondence: 20210001@cqnu.edu.cn

Abstract: Non-negative matrix factorization is used to find a basic matrix and a weight matrix to approximate the non-negative matrix. It has proven to be a powerful low-rank decomposition technique for non-negative multivariate data. However, its performance largely depends on the assumption of a fixed number of features. This work proposes a new probabilistic non-negative matrix factorization which factorizes a non-negative matrix into a low-rank factor matrix with $\{0,1\}$ constraints and a non-negative weight matrix. In order to automatically learn the potential binary features and feature number, a deterministic Indian buffet process variational inference is introduced to obtain the binary factor matrix. Further, the weight matrix is set to satisfy the exponential prior. To obtain the real posterior distribution of the two factor matrices, a variational Bayesian exponential Gaussian inference model is established. The comparative experiments on the synthetic and real-world datasets show the efficacy of the proposed method.

Keywords: Indian buffet process; binary components; non-negative matrix factorization; exponential Gaussian model

Citation: Ma, X.; Gao, J.; Liu, X.; Zhang, T.; Tang, Y. Probabilistic Non-Negative Matrix Factorization with Binary Components. *Mathematics* **2021**, *9*, 1189. https://doi.org/10.3390/math9111189

Academic Editor: Luca Gemignani

Received: 19 April 2021
Accepted: 20 May 2021
Published: 24 May 2021

1. Introduction

Non-negative matrix factorization (NMF) is an important nonlinear technique to find purely additive, parts-based representations of non-negative multivariate data [1]. It aims to reveal the latent structure or pattern in the data. Compared with the classical low-rank matrix factorization methods, such as the singular value decomposition and principal component analysis, NMF has turned out to be a powerful tool in data-mining analysis because of its stronger explanatory properties (parts constitute the whole) and rationality (decomposing factors must be non-negative in some real scenarios). For example, image information can be compressed and summarized into a combination of some features [2], and some feature words can be learned from the document to represent the document for clustering tasks [3]. This original information is generally stored in the computer as a data matrix, which is decomposed by NMF into the product of a basis factor matrix and a weight.

Most research on the NMF algorithm is based on the specific expected characteristics to impose other restrictions except negativity on the factorization factor matrix [4–6]. Although these works have made some achievements, some significant problems remain to be solved, such as the way to construct efficient algorithm to avoid complex iterative calculation, and the initialization problem which greatly weaken the flexibility of the model due to its unknowability. Wild et al. [7] proposed that the use of a structured initial value can improve the efficiency of the NMF algorithm at the cost of converging to a poor local solution.

The probability non-negative matrix factorization from statistical view may be enlightening. Bayesian NMF is the NMF model with KL penalty that was extended to the Bayesian setting [8]. Using the Bayesian nonparametric model of infinite dimensional parameter space, the number of features is determined as a random quantity in the process of posterior inference. Indian buffet process (IBP) is a classical latent feature model. This method generates a prior of a binary matrix, which has finite dimensional rows (number of objects) and infinite dimensional columns (number of features) [9]. The real latent finite element information can be obtained completely by the posterior inference with IBP. Many studies have proposed the application of IBP in NMF. For example, Yang et al. projected a non-negative binary matrix tri-factorization model that better reveals the underlying structure between samples and features, including two binary matrices and a weight matrix [10].

In this work, we come up with a new special probabilistic non-negative matrix factorization which requires the basis matrix Z atisfies $\{0,1\}$ constraints and the weighted matrix A is non-negative. This is clearer and more intuitive than the definition of tri-factorization proposed in [10]. For the binary matrix Z, the IBP is introduced to solve, and we assume that the weight matrix A obeys the exponential distribution under the non-negative constraint. In addition, based on synthetic dataset and real-world datasets, we compare our approach with benchmark methods to verify its flexibility and effectiveness. Main contributions are following three aspects:

(1) We present a pNMF with single binary component. Compared with other methods, the binary matrix can be regarded as the mapping from real object to binary codes, which is more intuitive.
(2) IBP is applied to the pNMF and we explain its use as a prior in the variational Bayesian exponential Gaussian model. The real latent information is completely obtained by inference and the sensitivity of the model to initialization parameter setting is greatly reduced.
(3) The experiments on the synthesized dataset and real-world datasets show the validity of the proposed method.

The rest of this paper is organized as follows: the related work is outlined in the next section. Section 3 explains the proposed method and the derivation process. Section 4 gives the empirical research results on four different types of dataset. The last section concludes the work.

2. Related Work

Recent research has inspired us to consider pNMF with single components on the basis of IBP. This section lists the literature review of NMF and IBP.

2.1. Non-Negative Matrix Factorization

NMF was firstly posed by Paatero and tapper [1], but enjoyed wide popularity due to the research results of Lee and Seung [2]. According to its definition, the factor elements after factorization must be non-negative, and simultaneously achieves nonlinear dimension reduction. Mathematically, given the non-negative matrix V, the problem is to find two factor matrices $W \geq 0$ and $H \geq 0$ to approximate V, so that $V \approx WH$. It is also usually interpreted as $V = WH + \varepsilon$ by researchers with blind source signal separation background [11], where ε is a noise matrix. Lee and Seung [2] also introduced two simpler algorithms whose objective functions are based on Euclid distance and Kullback–Leibler divergence (KL), respectively, and they proved the convergence of iterative rules. As an unsupervised learning method, NMF is widely used in many fields.

Many improved NMF algorithms have been advocated in recent years. Hoyer et al. [12] combined sparse penalty term with NMF to construct the sparse NMF model, in which parameters were solved by the gradient projection method and the EM algorithm. Wang et al. [13] combined Fisher's discriminant criterion with the objective function of

non-negative matrix factorization to extract local features. Lin et al. [14] brought forth an optimization algorithm of NMF based on projection gradient method, which refined the convergence rate of factorization. Liu et al. [6] found a restrictive NMF algorithm under semi-supervised learning that effectively combined the category information. Guillamet et al. [15] pointed out a weighted NMF algorithm to solve the image classification problem by better representation ability of local features.

For many NMF methods, the rank needs to be determined in advance, and its value was adjusted and searched through trial and error. Hinrich et al. [16] mentioned a probabilistic sparse NMF model that extends the variational Bayesian NMF and explicitly considers sparsity. Mohammadiha et al. [17] developed a hidden Markov model whose output density function is a gamma distribution. This model was reformulated as a probabilistic NMF, and it can capture the time dependence internally. To derive the maximum a posterior (MAP) of the non-negative factors, Shterenberg et al. [18] extended the NMF deterministic framework to the probabilistic such that the original data are no longer deterministic, which is considered to be a sample drawn from a multinomial distribution.

2.2. Indian Buffet Process

Griffiths et al. [9] proposed the IBP in 2005. They defined the probability distribution over equivalence classes of a binary matrix with a finite number of rows and an unbounded number of columns, and illustrated the use of it as a prior in an infinite binary linear–Gaussian model. At the same time, they have proved that this distribution as a prior is suitable for probabilistic models that use potentially infinite feature arrays to represent objects [19]. Many studies have been extended on the basis of IBP to generate more general distribution categories.

Ghahramani et al. [20] introduced a two-parameter generalization of the IBP that focused on separating the coupling relationship between the total number of features and the distribution on the feature number of each object. The et al. [21] analyzed the stick-breaking construction of IBP and developed a slice sampler with higher efficiency and easier application. Thibaux and Jordan [22] demonstrated the exchangeable distribution produced by IBP corresponds to the use of a latent measure based on the beta process. Gael et al. [23] rendered a strategy for modifying the IBP to capture the dependency which can arise as the observations being generated in a specific sequence. Miller et al. [24] used the phylogenetic IBP to seek a dependency that is taken as the result of known degrees of relatedness among observations.

The significance of IBP is to produce a distribution of infinite sparse binary matrix, which can be used to define many new non-parametric Bayesian models with different likelihood functions. Miller et al. [25] defined a class of non-parametric latent feature models that can be used for link prediction. Meeds et al. [26] presented a model of dyadic data based on the two-parameter IBP and developed some novel Metropolis-Hasting proposals for inference. Wood et al. [27] adopted a non-parametric Bayesian approach based on the IBP to model the structure graphs with many hidden causes. Navarro et al. [28] proposed a non-parametric Bayesian model for inferring features from similarity judgements.

3. The Proposed Methods

3.1. Model Framework

Given a non-negative matrix $X \in \mathbb{R}^{N \times D}$, the product of two non-negative factor matrices Z and A is expected to be as close to X as possible. Our data model is shown in Figure 1.

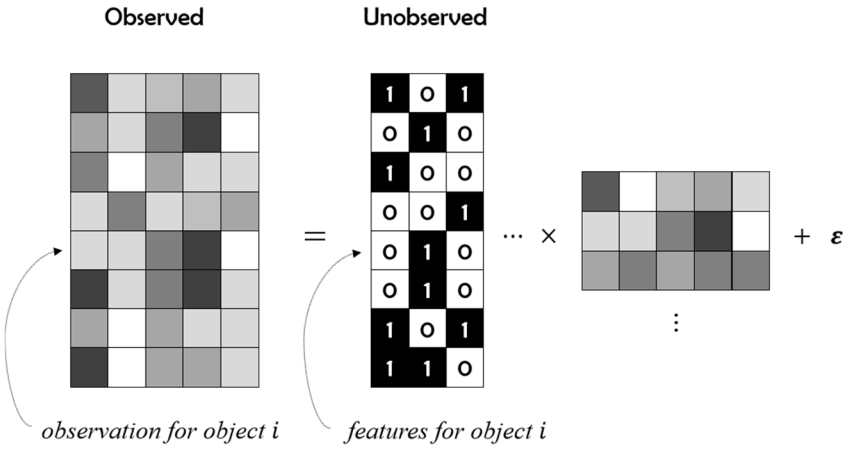


Figure 1. Non-negative matrix factorization with a single binary component.

Where Z is a binary matrix of $N \times K$ dimension, and A is a non-negative weight matrix of $K \times D$ dimension. In this model, X follows the Gaussian conditional distribution:

$$X \approx ZA \text{ s.t. } Z \in \{0,1\}^{N \times K} and\ A \in \mathbb{R}_{+}^{K \times D}. \tag{1}$$

$$X_n \sim Normal(Z_n A, \sigma_n^2), for\ n \in \{1, \ldots, N\}. \tag{2}$$

According to Bayesian rule:

$$p(Z, A|X) \propto \underbrace{p(X|Z, A)p(A)}_{model\ specific} \times \underbrace{p(Z)}_{prior\ on\ binary\ matrix}. \tag{3}$$

It can be seen that each column in Z corresponds to the existence of a latent feature, and all elements of Z are non-negative due to the $\{0,1\}$ constraints. $p(Z)$ is just the prior probability function defined by IBP on the binary matrix Z in (1). For A, we need a prior to express rules about its value, so we assumed that A obeys the exponential distribution to meet the non-negative limitation. Thus, each element of A obeys the exponential distribution [8].

Therefore, the defined model can be considered as an exponential Gaussian model, and the prior representation of these two matrices is:

$$\begin{gathered} A_{jk} \sim \varepsilon\left(A_{jk}|\lambda_k\right),\ \lambda_k \sim \mathcal{G}(\lambda_k|\alpha_0, \beta_0),\ z_{nk} \sim Bernoulli(\pi_k), \\ s.t.\ k \in \{1, \ldots, \infty\},\ j \in \{1, \ldots, D\},\ n \in \{1, \ldots, N\}. \end{gathered} \tag{4}$$

In accordance with the automatic association determination algorithm proposed by Thomas et al. [29], the parameter λ is set using gamma prior, which helps the automatic selection of the model. The whole factor k is either "activated" or "off", and only a few effective factors are set to 0 based on the magnitude of λ. Figure 2 indicates the graph structure of the exponential Gaussian model.

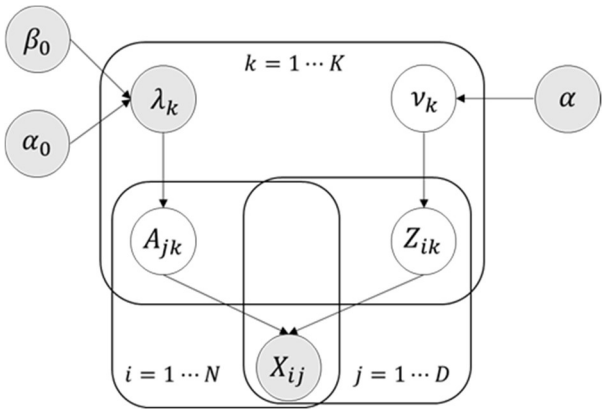


Figure 2. The exponential Gaussian model.

According to (3), the goal of our model is to obtain the true posterior distribution $ln(\,p(Z, A|X, \sigma_n^2, \alpha, \lambda))$ of factor matrices. It is very difficult to solve this problem with conventional methods, but the variational approximation algorithm is likely to be an effective method. Many studies have revealed that this algorithm simplifies inference process and guarantees accuracy to a certain extent.

Like Gibbs sampling, the mean field variational inference is a common approximate Bayesian inference method, which is to simplify the calculation when the introduced variational distribution $q(\theta)$ is as close to the real distribution $p(\theta|X)$ as possible. The theory assumes that the variational distribution q can be completely decomposed as $q(\theta) = \prod_i q(\theta_i)$, and all variables are calculated independently in the process of posteriori inference. It has been proved in [30] that the optimal distribution $q^*(\theta_i)$ of the i^{th} parameter can be expressed as follows:

$$log\, q^*\,(\theta_i) = \mathbb{E}_{q(-\theta_i)}[log\, p\,(\theta, X)] + C. \tag{5}$$

where C is a constant. The objective function is usually to maximize the lower bound of evidence (ELBO) or minimize the KL divergence as much as possible, and it is essential to choose a suitable approximate distribution q for solving the variational problems.

On the basis of the model framework, our work is mainly to find the best approximate distribution of two factor matrices that are consistent with the constraints.

3.2. The Solution of Weighted Matrix A

For the weight matrix A, the exponential distribution is used as a prior in the model to limit its non-negativity. According to Bayes' rule, the posterior of A is approximately a truncated normal distribution if Z is fixed. The formula is as follows:

$$\begin{aligned}
q^*(A_{jk}) &\propto exp\left\{\mathbb{E}_{q\left(\theta_{-A_{jk}}\right)}[log\, p\,(X|\theta)] + log\, p\,(\theta)\right\} \\
&\propto exp\left\{\mathbb{E}_{q(\theta_{-A_{jk}})}\left[\sum_{i\in\Omega_j} log\left[\sqrt{\frac{\tau}{2\pi}} exp\left\{-\frac{\tau}{2}(X_{ij} - Z_i A_j)^2\right\}\right] + log\left[\lambda_k\, exp\{-\lambda_k A_{jk}\}\right]\right]\right\} \times u(x) \\
&\propto exp\left\{\mathbb{E}_{q(\theta_{-A_{jk}})}\left[\sum_{i\in\Omega_j} -\frac{\tau}{2}(X_{ij} - Z_i A_j)^2 - \lambda_k A_{jk}\right]\right\} \times u(x) \\
&\propto exp\left\{\mathbb{E}_{q(\theta_{-A_{jk}})}\left[-\frac{\tau}{2}\sum_{i\in\Omega_j}\left[Z_{ik}^2 A_{jk}^2 - 2Z_{ik}A_{jk}(X_{ij} - \sum_{k'\neq k} Z_{ik'}A_{jk'})\right]\right] - A_{jk}\widetilde{\lambda_k}\right\} \times u(x)
\end{aligned} \tag{6}$$

$$\propto exp\left\{A_{jk}\left[-\widetilde{\lambda_k}+\tilde{\tau}\sum_{i\in\Omega_j}\left(X_{ij}-\sum_{k'\neq k}\widetilde{Z_{ik'}}\widetilde{A_{jk'}}\right)\widetilde{Z_{ik}}\right]-\frac{A_{jk}^2}{2}\left[\tilde{\tau}\sum_{i\in\Omega_j}\widetilde{Z_{ik}^2}\right]\times u(x)\right\}$$

$$\propto exp\left\{-\frac{\tau_{jk}}{2}\left(A_{jk}-\mu_{jk}\right)^2\right\}\times u(x)$$

$$\propto \mathcal{TN}(A_{jk}|\mu_{jk},\tau_{jk}).$$

Here,

$$\mathcal{TN}(x|\mu,\tau)=\begin{cases}\dfrac{\sqrt{\dfrac{\tau}{2\pi}}exp\{-\dfrac{\tau}{2}(x-\mu)^2\}}{1-\Phi\left(-\mu\sqrt{\tau}\right)} & if\ x\geq 0\\ 0 & if\ x<0\end{cases}, \tag{7}$$

where A_{-jk} represents all elements in the matrix A_{jk} except the $k-th$ column, and it can be seen that the posterior probabilities of non-positive elements are all set to 0. This process is not a strict optimization of A, which may cause some errors in the final reconstruction results.

According to the derivation process of (6), the mean and variance of the true posterior $q^*(A_{jk})$ can be obtained as follows:

$$\begin{aligned}\tau_{jk}&=\tilde{\tau}\sum_{i\in\Omega_j}\widetilde{Z_{ik}^2},\\ \mu_{jk}&=\frac{1}{\tau_{jk}}\left(-\widetilde{\lambda_k}+\tilde{\tau}\sum_{i\in\Omega_j}\left(X_{ij}-\sum_{k'\neq k}\widetilde{Z_{ik'}}\widetilde{A_{jk'}}\right)\widetilde{Z_{ik}}\right).\end{aligned} \tag{8}$$

It can be found that the smaller the mean (negative), the larger the variance. The truncated distribution is closest to the exponential distribution when the expected scale parameter is $|\mu*\tau|$. In order to prevent the calculation of mean and variance effectively from being invalidated by numerical errors, it is necessary to restrict $|\mu|$. The setting of parameter λ is also an important factor affecting the lower bound of X. We assume that it satisfies the following distribution and update the relevant parameters:

$$\begin{aligned}q(A_{jk})=\mathcal{TN}(A_{jk};\mu_{jk},\tau_{jk}),\ \ q(\lambda_k)=\mathcal{G}(\lambda_k;\alpha_k^*,\beta_k^*),\\ \alpha_k^*=\alpha_0+D,\ \ \beta_k^*=\beta_0+\textstyle\sum_{j=1}^{D}\widetilde{A_{jk}}.\end{aligned} \tag{9}$$

3.3. The Solution of Binary Matrix Z

The generation process of IBP is similar to many Indian cafeterias in London. Endless plates line up, and customers enter one by one, choosing food from left to right. If selected by a customer, this dish is marked as "1", otherwise "0". The number of dishes selected by each customer satisfies the Poisson distribution [9]. The process is shown in Figure 3.

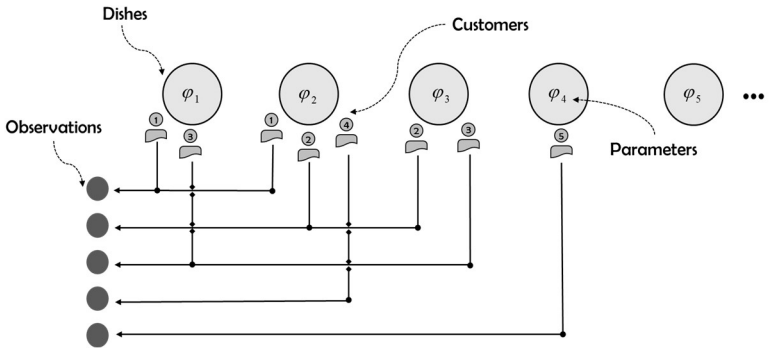


Figure 3. Illustration of Indian Buffet process.

As shown in Figure 3, the binary sequence representing the choice of the third customer can be expressed as $[1, 0, 1, 0, 0, \cdots]$. By connecting N pieces of such sequence vertically, IBP generates a priori of a binary matrix with finite number of rows (number of data) and infinite number of columns (number of latent features) to define many new nonparametric Bayesian models with different likelihood functions. Combined with the analysis of the model framework, IBP is used to approximate the binary basis matrix Z. In order to facilitate the process of variational derivation, the stick breaking construction based on the research of The et al. [16] is introduced as follows:

$$v_j \sim Beta(\alpha, 1),\ \pi_k = \prod_{j=1}^{k} v_j,\ z_{nk} \sim Bernoulli(\pi_k). \tag{10}$$

where v_j is the variable, α is a parameter of the beta distribution and π_k is the truncated weight of column k. The basis matrix Z in our model adopts the approximate form similar to the infinite variational approach of linear Gaussian model proposed by Doshi et al. [12]:

$$q_{v_{nk}}(z_{nk}) = Bernoulli(z_{nk}; v_{nk}),\ q_{\tau_k}(v_k) = Beta(v_k; \tau_{k1}, \tau_{k2}). \tag{11}$$

To facilitate calculation, the π is replaced by v. The derivation of v and its parameters are given in [12]. We will present this information in detail in the following derivation.

3.4. Variational Inference Process

Based on the above analysis, the variation distribution of our model is shown in Figure 4.

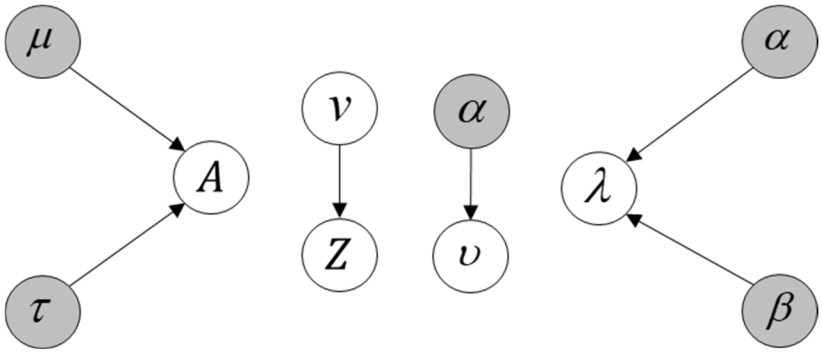


Figure 4. Approximate variational distribution diagram.

Since it is difficult to take the real expectation of the log joint likelihood $\log p(X|\theta)$, this paper gives its lower bound through reasoning. The variational objective function of our model is as follows:

$$\begin{aligned} log\ p\ (X|\theta) \geq & \sum_{k=1}^{K} \mathbb{E}_v[log\ p\ (v_k|\alpha)] + \sum_{k=1}^{K}\sum_{n=1}^{N} \mathbb{E}_{v,Z}[log\ p\ (Z_{nk}|v)] \\ & + \sum_{k=1}^{K} \mathbb{E}_\lambda[log\ p\ (\lambda_k|\alpha_0, \beta_0)] + \sum_{k=1}^{K}\sum_{j=1}^{D} \mathbb{E}_{\lambda,A}[log\ p\ (A_{jk}|\lambda_{jk})] \\ & + \sum_{n=1}^{N} \mathbb{E}_{Z,A}[log\ p\ (X_n|Z, A, \sigma_n^2)] + H[q], \end{aligned} \tag{12}$$

where $\theta = \{\alpha, \alpha_0, \beta_0, \sigma_n^2\}$ is the parameter. From (12), It can be found from (12) that except the second term, all items are calculated by exponential distribution family, and their expected value derivation can refer to the study of Doshi et al. [12]. The $\mathbb{E}_\lambda[*]$ is:

$$\prod_{k=1}^{K} \mathbb{E}_{\lambda}[log\, p\,(\lambda_k|\alpha_0,\beta_0)] = \alpha_0\, log\, \beta_0 - ln\, \Gamma\,(\alpha_0) + (\alpha_0 - 1)\sum_{k=1}^{K} \psi(\alpha_k) + \beta_0 \sum_{k=1}^{K} \frac{\alpha_k}{\beta_k}. \quad (13)$$

According to the research results of Chopin and Mazet [31,32], we can draw the truncated distribution formula and simulate the Gaussian distribution defined on the finite interval $[a, b]$. The semi finite interval can still be considered by setting $b = +\infty$. The principle of this method is to divide the interval into the same area and use a properly distributed accept–reject algorithm. Since the A_{jk} is a truncated normal distribution, the $\mathbb{E}_{\lambda,A}[*]$ is:

$$\begin{aligned} \mathbb{E}_{\lambda,A}[log\, p\,(A_k|\lambda_k)] = \mathbb{E}_{\lambda,A}[log(\,\lambda_k \cdot exp(-\lambda_k A_k))] &= \mathbb{E}_{\lambda,A}[log\,\lambda_k - \lambda_k A_k] \\ &= log\,\lambda_k - \lambda_k \mathbb{E}[A_k] \\ = log\frac{\alpha_k}{\beta_k} - \frac{\alpha_k}{\beta_k}\left(\mu_k + \tau_k \cdot \frac{1}{\sqrt{2\pi \cdot \tau_k}} \cdot exp\left(-\frac{u_k^2}{2\tau_k}\right)\Big/\frac{1}{2} erfc\left(-\sqrt{\frac{\tau_k}{2}} \cdot \mu_k\right)\right). \end{aligned} \quad (14)$$

For the lower truncated normal distribution, the expectation is as follows:

$$\mathbb{E}(x|x < c) = \mu + \sigma^2 \frac{f(c)}{S(c)}. \quad (15)$$

Here $f(x)$ is the probability density function and $S(x)$ is the complementary cumulative distribution function, and their formulas are as follows:

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \\ S(x) = 1 - F(x) &= 1 - cdf(x) = 0.5 * \mathrm{erfc}\left(\frac{x}{\sqrt{2}}\right). \end{aligned} \quad (16)$$

The entropy $H[q]$ in (11) is:

$$\begin{aligned} H[q] = -\mathbb{E}_q\, log\left[\prod_{k=1}^{K} q_{\tau_k}(\pi_k) \prod_{k=1}^{K} q(\lambda_k) \prod_{k=1}^{K} q_{\mu_k}(A_k) \prod_{k=1}^{K}\prod_{n=1}^{N} q_{v_{nk}}(z_{nk})\right] \\ = \sum_{k=1}^{K} \mathbb{E}_{\pi}\left(-log\, q_{\tau_k}\,(\pi_k)\right) + \sum_{k=1}^{K} \mathbb{E}_{\lambda}\big(-log\, q\,(\lambda_k)\big) \\ + \sum_{k=1}^{K} \mathbb{E}_{A}(-log\, q_{\mu_k}\,(\boldsymbol{A}_k)) + \sum_{k=1}^{K}\sum_{n=1}^{N} \mathbb{E}_{Z}(-log\, q_{v_{nk}}\,(z_{nk})) \end{aligned} \quad (17)$$

where,

$$\begin{aligned} &\mathbb{E}_{\pi}\left(-log\, q_{\tau_k}\,(\pi_k)\right) = log\left(\frac{\Gamma(\tau_{k1})\Gamma(\tau_{k2})}{\Gamma(\tau_{k1}+\tau_{k2})}\right) - (\tau_{k1}-1)\psi(\tau_{k1}) - (\tau_{k2}-1)\psi(\tau_{k2}) \\ &+(\tau_{k1}+\tau_{k2}-2)\psi(\tau_{k1}+\tau_{k2}), \\ &\mathbb{E}_{\lambda}\big(-log\, q\,(\lambda_k)\big) = ln\,\Gamma\,(\alpha_k^*) - (\alpha_k^*-1)\psi(\alpha_k^*) - ln\,\beta_k^* + \alpha_k^*, \\ &\mathbb{E}_{A}\left(-log\, q_{\mu_k}\,(A_k)\right) = 0.5^* log(\,(2\pi e)^D|\tau_k|), \\ &\mathbb{E}_{Z}(-log\, q_{v_{nk}}\,(z_{nk})) = -(1-v_{nk})\, log(\,1-v_{nk}) - v_{nk}\, log\, v_{nk}. \end{aligned} \quad (18)$$

The $\psi(\cdot)$ is a digamma function. According to the analysis, the derivation formula of the final objective function of (12) is as follows:

$$log\, p\,(X|\theta) \geq \sum_{k=1}^{K}\big[log\,\alpha + (\alpha-1)\big(\psi(\tau_{k1}) - \psi(\tau_{k1}+\tau_{k2})\big)\big] \quad (19)$$

$$+\sum_{k=1}^{K}\sum_{n=1}^{N}\left[v_{nk}\left(\sum_{m=1}^{k}\psi(\tau_{k2})-\psi(\tau_{k1}+\tau_{k2})\right)+(1-v_{nk})\mathbb{E}_{v}\left[log\left(1-\prod_{m=1}^{k}v_{m}\right)\right]\right]$$

$$+\alpha_0\, log\,\beta_0-ln\,\Gamma(\alpha_0)+(\alpha_0-1)\sum_{k=1}^{K}\psi(\alpha_k)+\beta_0\sum_{k=1}^{K}\frac{\alpha_k}{\beta_k}$$

$$+log\frac{\alpha_k}{\beta_k}-\frac{\alpha_k}{\beta_k}\left(\mu_k+\tau_k\cdot\frac{1}{\sqrt{2\pi\cdot\tau_k}}\cdot\ exp\left(-\frac{u_k^2}{2\tau_k}\right)\Big/\frac{1}{2}erfc(-\sqrt{\frac{\tau_k}{2}}\cdot\mu_k)\right)$$

$$+\sum_{n=1}^{N}\left[-\frac{D}{2}log(2\pi\sigma_n^2)\right]$$

$$+\sum_{n=1}^{N}\left[-\frac{1}{2\sigma_n^2}\left(X_nX_n^T-2\sum_{k=1}^{K}v_{nk}\,\mu_kX_n^T+2\sum_{k<k'}v_{nk}v_{nk'}\,\mu_k\mu_k^T+\sum_{k=1}^{K}v_{nk}(tr(\tau_k)+\mu_k\mu_k^T)\right)\right]$$

$$+\sum_{k=1}^{K}\left[log\left(\frac{\Gamma(\tau_{k1})\Gamma(\tau_{k2})}{\Gamma(\tau_{k1}+\tau_{k2})}\right)-(\tau_{k1}-1)\psi(\tau_{k1})-(\tau_{k2}-1)\psi(\tau_{k2})+(\tau_{k1}+\tau_{k2}-2)\psi(\tau_{k1}+\tau_{k2})\right]$$

$$+\sum_{k=1}^{K}\frac{1}{2}log(\,(2\pi e)^{D}|\tau_k|)+\sum_{k=1}^{K}ln\,\Gamma(\alpha_k^*)-(\alpha_k^*-1)\psi(\alpha_k^*)-ln\,\beta_k^*+\alpha_k^*$$

$$+\sum_{k=1}^{K}\sum_{n=1}^{N}[-v_{nk}\,log\,v_{nk}-(1-v_{nk})\,log(\,1-v_{nk})].$$

4. Experimental Result

The ability of the proposed method to perform matrix factorization and learn the number of latent features through the synthetic dataset is first demonstrated, and then the performance of the proposed method is verified by comparing with the benchmark methods on real datasets, including Swimmer dataset, Cora document dataset, and CBCL face dataset. While showing the effect of the model, we list the initialization setting of the parameters in each task, which have a great impact on the quality of convergence.

The main parameters involved in the experiment include IBP parameter α, lambda parameter α_0, β_0, Gaussian parameter σ_n and truncated feature number K. The optimal effect parameter will be given for reference at the end of each experiment. The algorithm experiment is carried out in the environment of Matlab2015b.

4.1. Sythetic Dataset

We apply the model to a simple composite dataset generated from known features. The dataset is simple enough, and the accuracy of graph reconstruction can reach almost 100%, and there are enough differences to resolve the qualitative features of latent feature inference. We have generated a specific 20×16-dimensional non-negative matrix, some of which is shown in Figure 5.

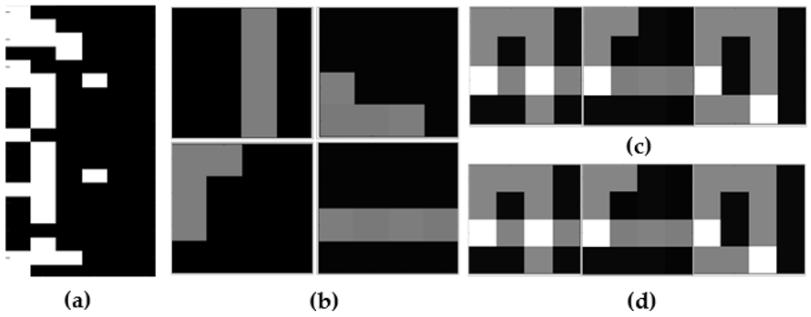


Figure 5. The analysis and reconstruction results on synthetic dataset: (**a**) the binary matrix Z, (**b**) the extracted latent features, (**c**) the original sample, (**d**) the reconstructed result. Here the white part indicates that the object has this feature while the black part indicates the opposite.

Figure 5a shows the basis matrix Z with 20×4 dimension. Figure 5b shows the extracted latent features, there are four in total. The original image is sufficient to justify it. For example, the first image is made up of three feature elements in Figure 5b except for the item in the upper right corner, while the third image contains that single item.

Figure 5d shows the reconstructed image using the extracted latent features, and the comparison shows that the accuracy can reach almost 100%.

The above analysis fully proves the effectiveness of our algorithm for automatic learning of latent features. Then, the parameters are initialized as $\alpha = 1, \alpha_0 = 0.1, \beta_0 = 10, K = 6$.

4.2. Swimmer Dataset

The swimmer dataset contains all possible combinations of swimmer's body postures (including limbs and trunk), with a total of 256 grayscale images. Figure 6 shows some samples of the dataset, depicting four simplified human figures with limbs and showing the different joints of the limbs. In this paper, the dataset is represented as a 256×1024-dimensional matrix X, and each column represents a 32×32-dimensional image.

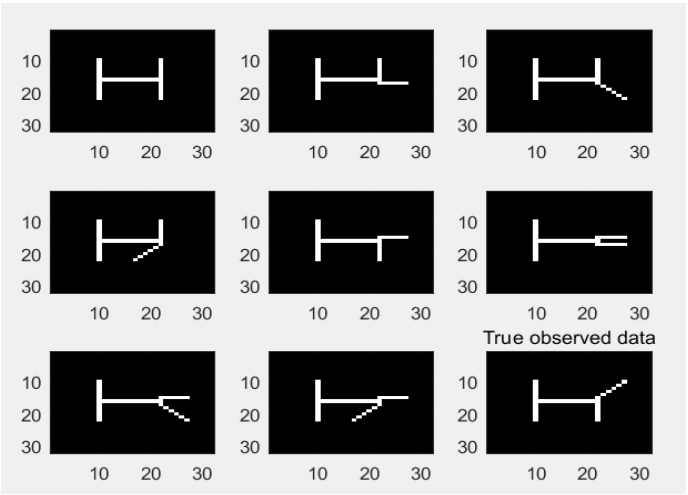


Figure 6. Samples of the swimmer dataset.

The experiment is designed to observe 16 variable limb features and 1 fixed trunk feature through model learning. The trunk and limbs should be separate parts. According to our model, 17 features are accurately extracted, as shown in Figure 7.

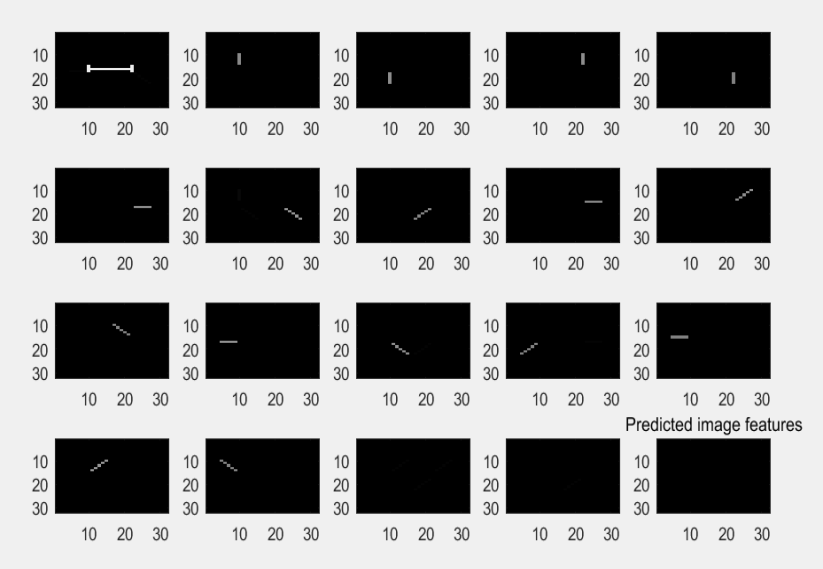
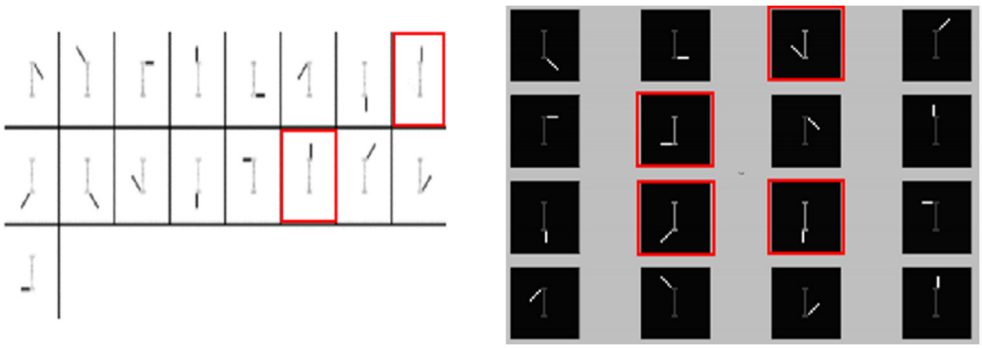


Figure 7. The extracted latent feature information.

On the contrary, most previous studies cannot successfully separate the backbone part. Since our algorithm is based on the NMF framework, the result of the basic NMF is first shown in Figure 8a. This result is derived from the research results of Gao Liang et al. in [33]. It can be seen that the eighth base image in the first line is almost identical to the sixth in the second line. The body part exists in each extracted feature, so the basic NMF method cannot independently extract the inherent or variable features of the dataset, and there is obvious overlap between the base feature values.



(**a**) (**b**)

Figure 8. Analysis and comparison results on swimmer dataset: (**a**) basic NMF decomposition results, (**b**) the comparative decomposition results including various combinations of limbs and the trunk.

Figure 8b shows the result of feature extraction in [34]. It can be seen that the four basic images that have been specially labeled are all a combination of limbs and the trunk. This is contrary to common sense.

Referring to two comparative experimental methods, a series of actual factors have been "polluted" to a certain extent. The results show that the proposed method has good data representation and feature recognition ability.

In addition, Figure 9 shows the basis matrix Z. The parameters are initialized as $\alpha = 1$, $\sigma_n = 0.1$, $\alpha_0 = 1$, $\beta_0 = 1$, $K = 20$.

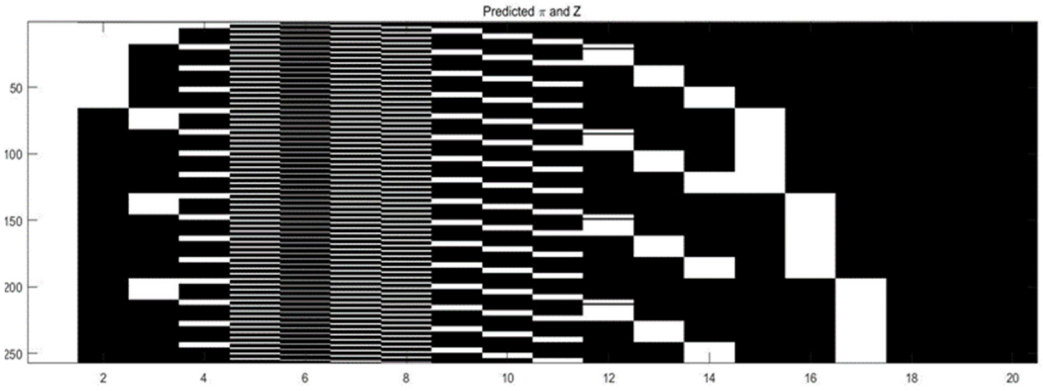


Figure 9. Binary matrix Z of the swimmer dataset.

4.3. Document Clustering

In this subsection, we apply the proposed model to a practical task: document clustering. The Cora dataset is used for experimental data that contains 2708 papers, including 1433 unique words after removing the stop words and the words that appear less than 10 times in the literature. The labels of these documents are their research field, about 7 categories. In order to better measure the effect of clustering task, we use three evaluation indicators: Jaccard coefficient (JC), Fowlkes and Mallows (FM), and F-measurement (F1). The larger their value, the better the clustering effect is, and their formulas are as follows:

$$\begin{aligned} JC &= a/(a+b+c), \\ FM &= \sqrt{a/(a+b)} \cdot \sqrt{a/(a+c)}, \\ F1 &= 2a^2/(2a^2+ac+ab). \end{aligned} \tag{20}$$

where a is the number of point pairs belonging to the same cluster in the benchmark results and test results, b is the number of point pairs in the same benchmark but different cluster, and c is the number of point pairs in the same cluster but different benchmark cluster.

The results of our method are compared with some parametric models, such as spectral clustering (SP), sparse NMF (SNMF), NMF, and also some Bayesian non-parametric models, which will be described in Table 1.

Table 1. Clustering result of the Cora dataset.

Index		JC	FM	F1
Parametric Model or Algorithms	SNMF	0.1678	0.2882	0.2870
	NMF	0.1524	0.2729	0.2688
	SP	0.1691	0.3123	0.2737
	IBP	0.1650	0.2770	0.2890
Bayesian Non-parametric NMF Models	Bivariate-Beta doubly IBP	0.1720	0.2790	0.2930
	Gaussian process doubly IBP	0.1650	0.2630	0.2610
	Ours	0.1749	0.3173	0.2977

The benchmark comparison data are obtained from [35]. The Bivariate-Beta distribution (BB-dIBP-NMF) and Gaussian process (GP-dIBP-NMF) are proposed based on doubly IBP. Finally, in our model, 14 latent features are obtained from the dataset, others are 17 (from BB-dIBP-NMF), 22 (from GP-dIBP-NMF). It can be seen that, compared with various methods mentioned in other papers, our algorithm obtains the best clustering results with the least features. Figure 10 shows this result more visually.

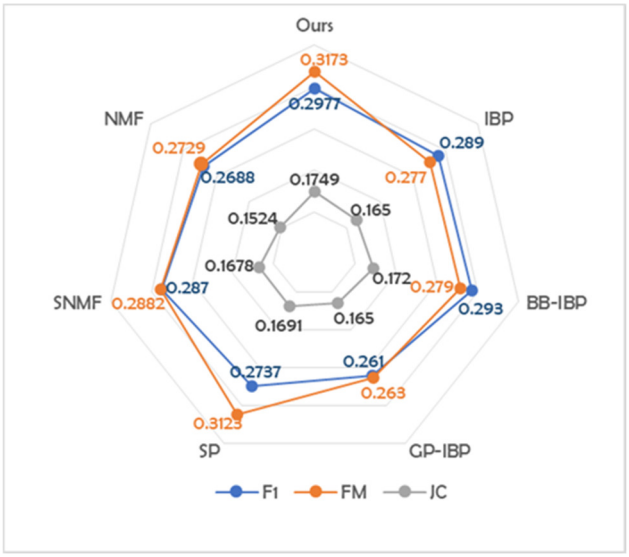


Figure 10. Comparative display of clustering results related to Cora dataset.

The parameters are initialized as $\alpha = 20$, $\sigma_n = 0.1$, $\lambda = 100$. And the square difference σ of spectral clustering is set to 0.2.

4.4. Face Feature Extraction

Not only can NMF decompose quickly and accurately, but the decomposition result has definite physical meaning. Lee and Seung [2] took face recognition as an example to decompose the face into separate parts such as nose, eyes, mouth and eyebrows. This is consistent with the intuitive visual concept of "the parts constitute the whole" in human thinking. And they compared the decomposition results of the NMF algorithm with those of common matrix factorization methods such as vector quantization (VQ) and PCA. The difference is that NMF is based on partial facial representations, while VQ and PCA are global representations. Therefore, we aim to apply the proposed model to the same face dataset (CBCL Face Database) [36] and observe whether our model can accurately extract similar local features and completely reconstruct the entire face, which is conducive to verifying the effectiveness of our model.

The CBCL Face Database contains many face and non-face images. It is widely used in the Biology and Computing Learning Center of the MIT. We use 2429 face images in the training set. Figure 11 shows some face images in this dataset.

Figure 11. The samples of CBCL Face Database.

According to the algorithm model, we extracte 51 "local" facial features from 2429 images, while 49 features are extracted by Lee and Seung [2]. Figure 12 shows facial feature extracted by our model, where white pixels represent features and the black is set as background.

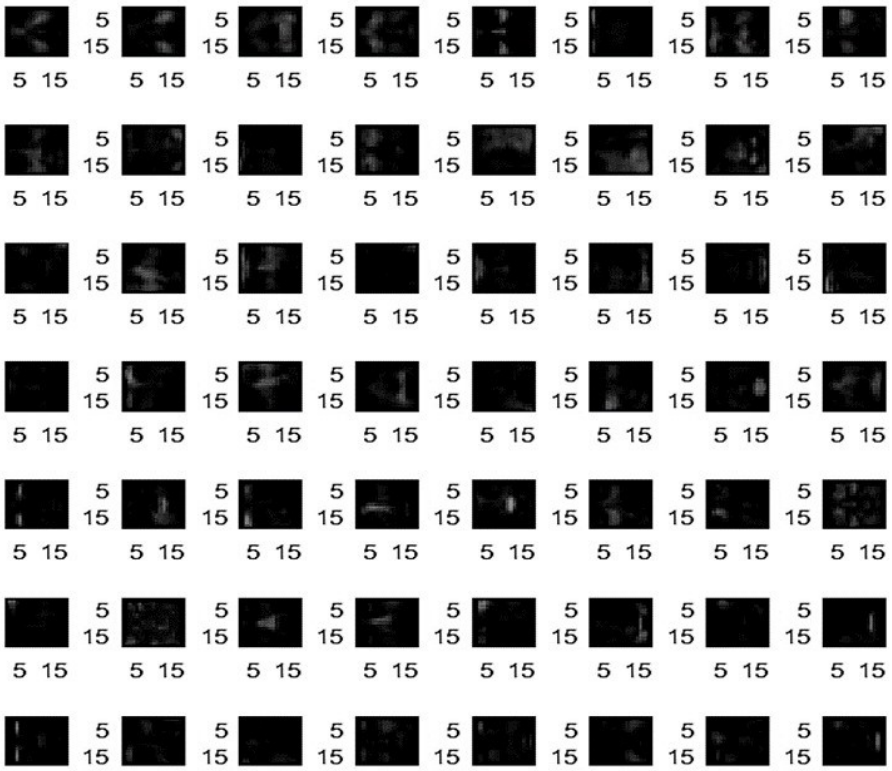


Figure 12. Extracted local features of human face.

Figure 13 shows the comparison between the reconstructed specific face image and the original image. By calculation, the mean square error rate of the reconstruction of the whole dataset is only 8%.

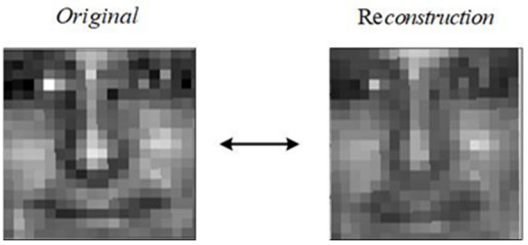


Figure 13. Comparison of original face and reconstructed result.

The parameters are initialized as $\alpha = 12, \sigma_n = 1, \alpha_0 = 3, \beta_0 = 1$.

5. Conclusions

As we know, traditional NMF models need to assume the number of basis to be fixed which makes them less flexible for many applications. IBP provides a priori of an infinite binary matrix, which is usually used to infer latent features and their quantities from a set of observation data. In this paper, we propose an improved IBP based probabilistic NMF framework, which uses exponential distribution and $\{0,1\}$ onstraints to limit the non-negativity of the weight matrix and the basis matrix. This method allows us to flexibly choose appropriate parameters as the basis to avoid the problems caused by the initial decomposition rank setting, and to automatically learn the latent features. At the same time, an exponential Gaussian model is constructed based on this framework, and the true posterior of two factor matrices is derived with the variational Bayes method. Compared with the benchmark methods, the proposed method has achieved better results on both syn-

thetic and real datasets, and it shows higher decomposition efficiency. However, the proposed model has great computational complexity. The efficiency of execution and the reduction of model complexity may be one of the future research contents.

Author Contributions: Conceptualization, X.M. and T.Z.; methodology, X.M., J.G. and T.Z.; writing—original draft preparation, X.M. and J.G.; writing—review and editing, X.M., X.L. and J.G.; visualization, X.M. and X.L.; supervision, T.Z. and Y.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China (No.62076043) and the Science and Technology Research Program of Chongqing Municipal Education Commission (No. KJQN201900110).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No data available.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Paatero, P.; Tapper, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* **1994**, *5*, 111–126.
2. Lee, D.; Seung, H. Learning the parts of objects by nonnegative matrix factorization. *Nature* **1999**, *401*, 788–791.
3. Ramsay, J.O. Functional Data Analysis. In *Encyclopedia of Statistical Sciences*; Everitt, B.S., Howell, D.C., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2004; p. 4.
4. Bach, F.; Mairal, J.; Ponce, J. Convex sparse matrix factorizations. *arXiv* **2008**, arXiv:0812.1869.
5. Zafeiriou, S.; Tefas, A.; Buciu, I.; Pitas, I. Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification. *IEEE Trans. Neural. Netw. Learn Syst.* **2006**, *17*, 683–695.
6. Liu, H. Constrained nonnegative matrix factorization for image representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 1299–1311.
7. Cai, D.; He, X.; Han, J.; Huang, T.S. Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 1548–1560.
8. Schmidt, M.N.; Winther, O.; Hansen, L.K. Bayesian Non-Negative Matrix Factorization. In *Independent Component Analysis and Signal Separation, Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation, Paraty, Brazil, 15 March 2009*; Adali, T., Christian, J., Romano, J.M.T., Barros, A.K., Eds.; Springer: Berlin, Germany, 2009; pp. 540–547.
9. Wild, S.; Curry, J.; Dougherty, A. Improving non-negative matrix factorizations through structured initialization. *Pattern Recognit.* **2004**, *37*, 2217–2232.
10. Yang, X.; Huang, K.; Zhang, R.; Hussain, A. Learning latent features with infinite nonnegative binary matrix trifactorization. *IEEE Trans. Emerg. Top. Comput.* **2018**, *2*, 450–463.
11. Zhang, Y.; Fang, Y. A NMF algorithm for blind separation of uncorrelated signals. In Proceedings of the International Conference on Wavelet Analysis and Pattern Recognition, Beijing, China, 2 November 2007; Curran Associates Inc.: New York, NY, USA, 2007; pp. 999–1003.
12. Hoyer, P.O. Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.* **2004**, *5*, 1457–1469.
13. Jia, Y.W.Y.; Turk, C.H.M. Fisher non-negative matrix factorization for learning local features. In Proceedings of the Asian Conference on Computer Vision, Seoul, Korea, 27 January 2004; Hong, K.S., Zhang, Z.Y., Eds.; Asian Federation of Computer Vision Societies: Seoul, Korea, 2004; pp. 27–30.
14. Lin, C.J. Projected gradient methods for nonnegative matrix factorization. *Neural. Comput.* **2007**, *19*, 2756–2779.
15. Guillamet, D.; Vitria, J.; Schiele, B. Introducing a weighted non-negative matrix factorization for image classification. *Pattern Recogn. Lett.* **2003**, *24*, 2447–2454.
16. Hinrich, J.L.; Mørup, M. Probabilistic Sparse Non-Negative Matrix Factorization. In *Latent Variable Analysis and Signal Separation, Proceedings of the International Conference on Latent Variable Analysis and Signal Separation, Guildford, UK, 2 July 2018*; Deville, Y., Gannot, S., Mason, R., Plumbley, M., Ward, D., Eds.; Springer: Berlin, Germany, 2018; pp. 488–498.
17. Mohammadiha, N.; Kleijn, W.B.; Leijon, A. Gamma hidden Markov model as a probabilistic nonnegative matrix factorization. In Proceedings of the 21st European Signal Processing Conference (EUSIPCO 2013), Marrakech, Morocco, 9 September 2013; IEEE: New York, NY, USA, 2013; pp. 1–5.
18. Bayar, B.; Bouaynaya, N.; Shterenberg, R. Probabilistic non-negative matrix factorization: Theory and application to microarray data analysis. *J. Bioinform. Comput. Biol.* **2014**, *12*, 1450001.
19. Griffiths, T.L.; Ghahramani, Z. The Indian buffet process: An introduction and review. *J. Mach. Learn. Res.* **2011**, *12*, 1185–1224.

20. Knowles, D.; Ghahramani, Z. Infinite Sparse Factor Analysis and Infinite Independent Components Analysis. In *International Conference on Independent Component Analysis and Signal Separation, ICA'07, Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation, London, UK, 12 August 2007*; Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D., Eds.; Springer: Berlin, Germany, 2007; pp. 381–388.
21. Teh, Y.W.; Grür, D.; Ghahramani, Z. Stick-Breaking Construction for the Indian Buffet Process. In *Artificial Intelligence and Statistics, Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistic, San Juan, PR, USA, 21–24 March 2007*; Meila, M., Shen, X., Eds.; PMLR: San Juan, PR, USA, 2007; pp. 556–563.
22. Thibaux, R.; Jordan, M.I. Hierarchical Beta Processes and the Indian Buffet Process. In *Artificial Intelligence and Statistics, Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistic, San Juan, PR, USA, 21–24 March 2007*; Meila, M., Shen, X., Eds.; PMLR: San Juan, PR, USA, 2007; pp. 564–571.
23. Gael, J.V.; The, Y.; Ghahramani, Z. The infinite factorial hidden Markov model. In Proceedings of the 21th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–10 December 2008; Koller, D., Schuurmans, D., Bengio, Y., Bottou, L., Eds.; Curran Associates Inc: New York, NY, USA, 2008; pp. 1697–1704.
24. Miller, K.T.; Griffiths, T.L.; Jordan, M.I. The phylogenetic Indian buffet process: A non-exchangeable nonparametric prior for latent features. In Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence, Helsinki, Finland, 9–12 July 2008; Mcallester, D., Myllymaki, P., Eds.; AUAI Press: Arlington, VA, USA, 2008; pp. 403–407.
25. Miller, K.T.; Griffiths, T.L.; Jordan, M.I. Nonparametric latent feature models for link prediction. In Proceedings of the 22nd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 7–10 December 2009; Bengio, Y., Schuurmans, D., Lafferty, J.D., Williams, C.K.I., Culotta, A., Eds.; Curran Associates Inc.: New York, NY, USA, 2009; pp. 1276–1284.
26. Meeds, E.; Ghahramani, Z.; Neal, R.; Roweis, S.T. Modeling dyadic data with binary latent factors. In Proceedings of the 19th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 4–7 December 2006; Schölkopf, B., Platt, J.C., Hoffman, T., Eds.; MIT Press: Cambridge, MA, USA, 2006; pp. 977–984.
27. Wood, F.; Griffiths, T.L.; Ghahramani, Z.A. Non-parametric Bayesian method for inferring hidden causes. In Proceedings of the 22nd Conference in Uncertainty in Artificial Intelligence, Cambridge, MA, USA, 13–16 July 2006; AUAI Press: MA, USA, 2006; 536–543.
28. Navarro, D.; Griffiths, T. A nonparametric Bayesian method for inferring features from similarity judgments. In Proceedings of the 19th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 4–7 December 2006; Schölkopf, B., Platt, J.C., Hoffman, T., Eds.; MIT Press: Cambridge, MA, USA; 2006; pp. 1033–1040.
29. Brouwer, T.; Frellsen, J.; Lió, P. Comparative study of inference methods for Bayesian nonnegative matrix factorization. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Skopje, Macedonia, 18 September 2017; Ceci, M., Hollmén, J., Todorovski, L., Vens, C., Džeroski S., Eds.; Springer: Berlin, Germany, 2017; pp. 513–529.
30. Bernardo, J.M.; Bayarri, M.J.; Berger, J.O.; Dawid, A.P.; Heckerman, D.; Smith, A.F.; West, M. The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures. *Bayesian Stat.* **2003**, *7*, 210.
31. Chopin, N. Fast simulation of truncated Gaussian distributions. *Stat. Comput.* **2012**, *21*, 275–288.
32. Mazet, V. *Simulation d'une Distribution Gaussienne Tronquée sur un Intervalle Fini*; Université de Strasbourg: Strasbourg, France, 2012.
33. Gao, L.; Yu, J.; Pan, J. Feature Re-factorization-based data sparse representation. *J. Beijing Univ. Technol.* **2017**, *43*, 1666–1672.
34. Pan, W.; Doshi-Velez, F. A characterization of the non-uniqueness of nonnegative matrix factorizations. *arXiv* **2016**, arXiv:1604.00653.
35. Xuan, J.; Lu, J.; Zhang, G.; Da Xu, R.Y.; Luo, X. Doubly nonparametric sparse nonnegative matrix factorization based on dependent Indian buffet processes. *IEEE Trans. Neural. Netw. Learn. Syst.* **2018**, *29*, 1835–1849.
36. MIT Center for Biological and Computation Learning. Available online: http://www.ai.mit.edu/projects/cbcl.old (accessed on 8 May 2020).