# Learning Latent Features with Infinite Non-negative Binary Matrix Tri-factorization

Xi Yang[1], Kaizhu Huang[1], Rui Zhang[1], and Amir Hussain[2]

[1]Xi'an Jiaotong-Liverpool University, SIP, Suzhou, China
{Xi.Yang,Kaizhu.Huang,Rui.Zhang02}@xjtlu.edu.cn
[2] Division of Computing Science & Maths,
School of Natural Sciences, University of Stirling, Stirling FK9 4LA, UK
ahu@cs.stir.ac.uk

**Abstract.** Non-negative Matrix Factorization (NMF) has been widely exploited to learn latent features from data. However, previous NMF models often assume a fixed number of features, say $p$ features, where $p$ is simply searched by experiments. Moreover, it is even difficult to learn binary features, since binary matrix involves more challenging optimization problems. In this paper, we propose a new Bayesian model called infinite non-negative binary matrix tri-factorizations model (iNBMT), capable of learning automatically the latent binary features as well as feature number based on Indian Buffet Process (IBP). Moreover, iNBMT engages a tri-factorization process that decomposes a nonnegative matrix into the product of three components including two binary matrices and a non-negative real matrix. Compared with traditional bi-factorization, the tri-factorization can better reveal the latent structures among items (samples) and attributes (features). Specifically, we impose an IBP prior on the two infinite binary matrices while a truncated Gaussian distribution is assumed on the weight matrix. To optimize the model, we develop an efficient modified maximization-expectation algorithm (ME-algorithm), with the iteration complexity one order lower than another recently-proposed Maximization-Expectation-IBP model [9]. We present the model definition, detail the optimization, and finally conduct a series of experiments. Experimental results demonstrate that our proposed iNBMT model significantly outperforms the other comparison algorithms in both synthetic and real data.

**Keywords:** Infinite non-negative binary matrix tri-factorization, Infinite latent feature model, Indian Buffet Process prior

## 1 Introduction

Non-negative matrix factorization (NMF), a popular matrix decomposition technique, has been widely applied in data analysis and machine learning [8]. Typically, NMF can be exploited to reveal from observations the latent features and consequently be used in semantic recognition or clustering. However, previous NMF models usually assume the number of features as a constant parameter,

which is generally tuned or searched by trial and error. Such algorithms include the methods proposed in [1][2][10]. Moreover, when the factor matrix is assumed as binary, NMF is even challenging, since binary matrices usually lead to more difficult optimization.

To tackle the above problems, we extend standard NMF to learn binary features with a novel Bayesian model called infinite non-negative binary matrix tri-factorization (iNBMT) in this paper. Different from traditional NMF, the novel iNBMT model can select automatically from infinite latent features an optimal set by applying Indian Buffet Process (IBP) prior to the factor matrices. In addition, we manage to decompose the input sample matrix $\mathbf{Y}$ into triple matrix factors i.e., $\mathbf{Y} = \mathbf{ZWX}^T$, where $\mathbf{Z}$ and $\mathbf{X}$ are two binary matrices, and non-negative matrix $\mathbf{W}$ can be considered as a weight matrix. Compared from bi-factorization typically involved in NMF, tri-factorization can better capture latent features and reveal hidden structures underlying the samples [2]. Importantly, although two binary matrices are involved, we further propose an efficient modified maximization-expectation algorithm (ME-algorithm), which can be even fast used in very large matrix decomposition. In particular, the time complexity of our proposed ME-algorithm proves one order lower than another competitive model called Maximization-Expectation-IBP (ME-IBP) [9].

In the literature, there have been several proposals of NMF for binary matrix decomposition. However, all of them have certain drawbacks. Binary Matrix Factorization (BMF) proposed in [10] limits the input data to be binary; this is however too strong in real cases. On the other hand, the correlated IBP-IBP model enforces a product of two binary matrices to be still binary; such assumption is in general invalid unfortunately. Despite of its good properties, the recently-proposed Maximization-Expectation-IBP (ME-IBP) model [9] is slow in optimization. In particular, the iteration complexity for the ME-IBP model is $O(\gamma ND)$, which is significantly higher than $O(\alpha N + \beta D)$, the iteration complexity of our iNBMT model. Here, $N$ and $D$, usually two big numbers, denote respectively the number of observations and the dimensionality. $\alpha, \beta$, and $\gamma$ are three coefficients.

## 2 Notation and Background

### 2.1 Indian Buffet Process

IBP can be considered as a prior defined on models with infinite binary matrices. It is typically used to infer how many latent features each observation processes. Suppose $\mathbf{Y} \in \mathbb{R}^{N \times D}$ be generated by linear combination with $K$-dimensional vector of latent factors $\mathbf{W} \in \mathbb{R}^{K \times D}$ and the assignment matrix $\mathbf{Z} \in \mathbb{R}^{N \times K}$. The observed data $\mathbf{Y}$ is then modeled as $\mathbf{Y} = \mathbf{ZW} + \epsilon$. $\epsilon$ is noise term of distributed independently over $\mathcal{N}(0, \sigma\mathbf{I})$.

Let $\mathbf{Z}$ be a binary matrix where $z_{nk} = 1$ presents the latent feature $k$ belongs to the observation $n$. The following IBP prior on binary feature matrix $\mathbf{Z}$ is derived by placing independent beta priors on Bernoulli. $\pi_k$'s are generated independently for each column following a Beta prior. And then each object possessing feature $k$ are generated independently from a Bernoulli with mean

$\pi_k$.

$$\pi_{\mathbf{k}} \mid (\alpha) \sim Beta(\alpha/\mathbf{K}, 1)\ , \qquad \mathbf{Z} \mid \pi_k \sim Bernoulli(\pi_k)\ ,$$
$$p([\mathbf{Z}]) = \frac{\alpha^K}{\prod_{h>0} K_h!} e^{\{-\alpha H_N\}} \prod_{k=1}^{K} \frac{(N-m_k)!(m_k-1)!}{N!}\ , \tag{1}$$

where $K_h$ is the number of rows corresponding to the non-zero number $h$, $m_k = \sum_{i=1}^{N} z_{ik}$ is the number of objects possessing feature $k$, and $H_N = \sum_{j=1}^{N} \frac{1}{j}$ is the $N^{th}$ harmonic number.

The IBP inspired several infinite-limit versions of classic matrix factorization models, e.g. infinite ICA models [6]. In infinite limit, Grinffiths et al. take the IBP prior into the infinite limit by defining equivalence classes on binary matrices [5]. The equivalence classes are matrices permutating the order of columns through eliminating all the null columns. Therefore, let $K$ be unbounded and assume that we allow the number of active features $K_+$ to be learned from the data while remaining finite with probability one. By defining a scheme to re-order the non-zero columns of $Z$ we can take $K \to \infty$ and find

$$p([\mathbf{Z}]) = \frac{\alpha^{K_+}}{\prod_{h>0} K_h!} e^{\{-\alpha H_N\}} \prod_{k=1}^{K_+} \frac{(N-m_k)!(m_k-1)!}{N!}\ . \tag{2}$$

### 2.2 Maximization-Expectation Algorithm

The ME algorithm just reverses the roles of two steps in the classical EM algorithm by maximization over hidden variables and marginalization over random parameters [7]. Given a dataset $\mathbf{Y}$, $p(\mathbf{Y}, \mathbf{Z}, \mathbf{W})$ is a probabilistic model where $\mathbf{Z}$ and $\mathbf{W}$ are all hidden random variables. To perform approximate MAP inference, it is necessary to compute posterior or marginal probabilities such as $p(\mathbf{Z}|\mathbf{Y}), p(\mathbf{W}|\mathbf{Y})$ or $p(Y)$. It can be viewed as a special case of a Mean-Field Variational Bayes (MFVB) approximation to a posterior that cannot be computed analytically. $p(\mathbf{Z}, \mathbf{W}|\mathbf{Y})$ is approximated by $q(\mathbf{Z})q(\mathbf{W})$ [4] if we assume independent variational distributions.

In MFVB, the variational Bayesian approximation alternatively estimates these distributions by minimizing the KL-divergence between the approximation and the exact distribution: $KL[q(\mathbf{Z})q(\mathbf{W})\|p(\mathbf{Z}, \mathbf{W}|\mathbf{Y})]$. The results are closeformed with the updates,

$$q(\mathbf{Z}) \propto exp(\mathbb{E}[\ln p(\mathbf{Y} \mid \mathbf{Z}, \mathbf{W})_{q(\mathbf{Z})}])\ , \quad q(\mathbf{W}) \propto exp(\mathbb{E}[\ln p(\mathbf{Y} \mid \mathbf{Z}, \mathbf{W})_{q(\mathbf{W})}])\ . \tag{3}$$

## 3 Infinite Non-negative Binary Matrix Tri-factorization

### 3.1 Model Description

The iNBMT model is applied on a real-valued observation data $\mathbf{Y} \in \mathbf{R}^{\mathbf{N}\times\mathbf{D}}$ where the rows and columns could be exchangeable. For a latent feature model, we use the matrix $\mathbf{F}$ to indicate the latent feature values. Then our focus will be on a distribution over observations conditioned on features $p(\mathbf{Y}|\mathbf{F})$, where $p(\mathbf{F})$

is the prior over features. $F$ can be expressed as the element-wise product of these three components, $\mathbf{F} = \mathbf{Z} \bigotimes \mathbf{W} \bigotimes \mathbf{X}$, where a latent feature binary vector $\mathbf{x_j}$ is associated with each attribute, each item has a potential binary vector $\mathbf{z_i}$, and a matrix $\mathbf{W}$ represents the interaction weights parameter. Furthermore, the prior of the features is also defined by $p(\mathbf{F}) = p(\mathbf{Z})p(\mathbf{W})p(\mathbf{X})$.

In effect, we factorized $\mathbf{Y}$ into the linear inner product of the features and weight, $\mathbf{ZWX^T}$, generated by a fixed observation process $f(\cdot)$, as illustrated in Fig. 1. This process is equivalent to factorization or approximation of the data:

$$\mathbf{Y} \mid \mathbf{Z}, \mathbf{W}, \mathbf{X} \sim f(\mathbf{ZWX}^T, \theta) \ ,$$

where $\theta$ are hyperparameters specific to the model variant.

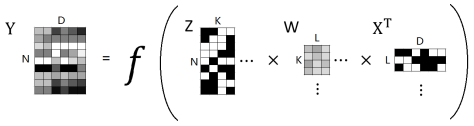



**Fig. 1.** Representation of the iNBMT model. The process $f(\cdot)$ applied to the linear inner product of the three components. Here $\mathbf{Z}, \mathbf{X}$ are infinite binary matrices, $\mathbf{W}$ present non-negative matrix.

We now develop our iNBMT model using Bayesian non-parametric priors. Specifically, IBP priors are imposed over binary matrices $Z$ and $X$, while any non-negative prior $\mathcal{F}$ (e.g. exponential and truncated Gaussian) is assumed on the weight matrix $W$:

$$\mathbf{Z} \sim \mathcal{IBP}(\alpha) \ , \qquad \mathbf{X} \sim \mathcal{IBP}(\lambda) \ , \qquad \mathbf{W} \sim \mathcal{F}(\mathbf{W}; \mu, \sigma_W^2) \ .$$

We assumed the hyperparameters were estimated from the data. By placing conjugate gamma hyperpriors on these parameters, we can have a straightforward extension to infer their values. Formally,

$$\mathbf{Y} \mid \mathbf{Z}, \mathbf{W}, \mathbf{X}, \theta \sim p(\mathbf{Y} \mid \boldsymbol{\theta}) \ , \qquad \boldsymbol{\theta} = \{\alpha, \lambda, \sigma_Y \sigma_W\} \sim Gamma(a, b) \ .$$

### 3.2 Linear-Gaussian iNBMT Model

To illustrate the iNBMT model for capturing the latent features, we set the linear-Gaussian model as the observation distribution with mean $\mathbf{ZWX}^T$ and covariance $(1/\theta)\mathbf{I}$ throughout this paper. This can be thought of a two-sided version of the linear-Gaussian model.

The marginal probabilities of the linear-Gaussian iNBMT model, is shown as below:

$$p(\mathbf{Y}|\mathbf{Z}, \mathbf{W}, \mathbf{X}, \sigma_X^2) = \frac{1}{(2\pi\sigma_Y^2)^{ND/2}} \exp -\frac{1}{2\sigma_Y^2} tr((\mathbf{Y} - \mathbf{ZWX}^T)^T(\mathbf{Y} - \mathbf{ZWX}^T)) \ .$$

The weight matrix $\mathbf{W}$ uses the truncated Gaussian priors with a zero-mean i.i.d.

$$p(\mathbf{W}|0, \sigma_W^2) = \prod_{k=1}^{K} \prod_{l=1}^{L} TN(a_{kl}; 0, \sigma_W^2) \ .$$

The marginal probabilities $p([\mathbf{Z}])$ and $p([\mathbf{X}])$ are specified with infinite IBP prior (given in Eq. (2)):

$$p(\mathbf{Z}|\alpha) = \frac{\alpha^{K_+}}{K_+!}\prod_{k=1}^{K}\frac{(N-m_k)!(m_k-1)!}{N!}\ ,$$
$$p(\mathbf{X}|\lambda) = \frac{\lambda^{L_+}}{L_+!}\prod_{l=1}^{L}[\frac{(D-m_l)!(m_l-1)!}{D!}]\ .$$

From the Bayesian theorem, the posterior can be write as follows:

$$p(\mathbf{Y},\mathbf{Z},\mathbf{W},\mathbf{X}|\boldsymbol{\theta}) = p(\mathbf{Y}|\mathbf{Z},\mathbf{W},\mathbf{X},\sigma_Y^2)p(\mathbf{W}|0,\sigma_W^2)p(\mathbf{Z}|\alpha)p(\mathbf{X}|\lambda)\ ,$$

where the hyperparameters $\boldsymbol{\theta}$ conjugate gamma priors on inference parameters.

### 3.3 Evidence of iNBMT

In this part, we will present the approximate MAP inference, derived from the ME algorithm, for the linear-Gaussian iNBMT model.

Given the MFVB constraint, we determine the variational distributions by minimizing the KL-divergence, $\mathcal{D}(q\|p)$, between the variational distribution and the true posterior; this is equivalent to maximizing a lower bound on the evidence:

$$\ln p(\mathbf{Y}|\theta) = \mathbb{E}_q[\ln p(\mathbf{Y},\mathbf{Z},\mathbf{W},\mathbf{X}|\theta)] + \mathcal{H}[q] + \mathcal{D}(q\|p) \tag{4}$$
$$\geq \mathbb{E}_q[\ln p(\mathbf{Y},\mathbf{Z},\mathbf{W},\mathbf{X}|\theta)] + \mathcal{H}[q] \tag{5}$$
$$\equiv \mathcal{T}\ , \tag{6}$$

where $\mathcal{H}[q]$ is the entropy of $q$. The lower bound of evidence, $\mathcal{T}$, for the linear-Gaussian iNBMT model is:

$$\begin{aligned}\mathcal{T} \equiv & \frac{1}{\sigma_Y^2}[-\frac{1}{2}(\mathbf{Z}\mathbb{E}[W]\mathbf{X}^T)(\mathbf{Z}\mathbb{E}[W]\mathbf{X}^T)^T + \mathbf{Z}(\mathbb{E}[W]\mathbf{Y}^T + \mathbf{Z}\boldsymbol{\gamma})\mathbf{X}^T] \\ & + \sum_{k=1}^{K}[\ln\frac{(N-m_k)!(m_k-1)!}{N!}] + \sum_{l=1}^{L}[\ln\frac{(D-m_l)!(m_l-1)!}{D!}] \\ & - \ln K_+! - \ln L_+! + \sum_{k=1}^{K}\sum_{l=1}^{L}\boldsymbol{\varphi}_{kl} + const\ ; \end{aligned} \tag{7}$$
$$\boldsymbol{\gamma} = \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{L}[\mathbb{E}[w_{kl}]^2 - \mathbb{E}[w_{kl}^2]]^T\ ,$$
$$\boldsymbol{\varphi}_{kl} = -\frac{KL}{2}\ln(\frac{\pi\sigma_W^2}{2}) - \frac{\mathbb{E}[w_{kl}^2]}{2\sigma_W^2} + \mathcal{H}(q(w_{kl}))\ .$$

Here $\mathbb{E}[\mathbf{W}]$ is a matrix with each element defined as $\mathbb{E}[w_{kl}]$.

### 3.4 Parameter Updates

The updates for the variational parameters of the non-negative $\mathbf{W}$ over the truncate Gaussian distribution are shown as follows:

$$q(\mathbf{W}) = \prod_{k=1}^{K}\prod_{l=1}^{L} TN(w_{kl}; \mu_{kl}, \sigma_{kl}^2) = \prod_{k=1}^{K}\prod_{l=1}^{L} \frac{N(\mu_{kl}, \sigma_{kl}^2)}{\mathbf{\Phi}(\infty) - \mathbf{\Phi}(\mathbf{0})} ,$$

where $t = -\frac{\mu_{kl}}{\sqrt{2}\sigma_{kl}}, \mathbf{\Phi}(\mathbf{a}) = \frac{1}{2}(1 + erf(\frac{a-\mu_{kl}}{\sqrt{2}\sigma_{kl}}))$, $\mathbf{\Phi}(\infty) = 1$, $erf(\cdot)$ is the Gaussian error function. According to the upper tail truncation, the parameters are updated as follows:

$$\mathbb{E}[w_{kl}] = \mu_{kl} + \sigma_{kl}\lambda(t) , \qquad \mathbb{E}[w_{kl}^2] = \mu_{kl}\mathbb{E}[w_{kl}] + \sigma_{kl}^2 ,$$
$$\lambda(t) = \frac{\sqrt{2}}{\sqrt{\pi}e^{t^2}(1-erf(t))} .$$

Meanwhile, the mean and variance of truncated Gaussian distributions can be updated as follows:

$$\mu_{kl} = \begin{cases} \tau^2 \sum_{n=1}^{N} z_{nk}^T (y_{nd} - \sum_{k'/k} z_{nk'}\mathbb{E}[w_{k'l}x_{dl}^T])x_{dl}, K \to \infty ; \\ \tau^2 \sum_{d=1}^{D} x_{dl}(y_{nd}^T - \sum_{l'/l} x_{dl}^T\mathbb{E}[w_{kl'}z_{nk'}])z_{nk}^T, L \to \infty . \end{cases} \tag{8}$$

$$\sigma_{kd} = \tau\sigma_Y , \tag{9}$$

where $\tau = (m_k^T m_l + \frac{\sigma_Y^2}{\sigma_W^2})^{-\frac{1}{2}}$. Then the entropy of truncated Gaussian distribution is given as

$$\mathcal{H}(q(w_{kl})) = \frac{1}{2\sigma_{kl}^2}\{\mathbb{E}[w_{kl}]^2 - \mathbb{E}[w_{kl}^2] - (\mathbb{E}[w_{kl}] - \mu_{kl})^2 - [\frac{1}{2}\ln\frac{2}{\pi\sigma_{kl}^2} - \ln(1 - erf(t))]\} .$$

The updates on $\mathbf{Z}$ and $\mathbf{X}$ are relatively straightforward by computing Eq. (3). Given $q(\mathbf{W})$, we compute MAP estimates of $\mathbf{X}, \mathbf{Z}$ by maximizing the evidence Eq. (7). Similar to variational IBP methods, we must split the expectation in Eq. (6) into terms depending on each of the latent variables [3], with the benefit that the binary variables updates are not affected by inactive features. Therefore, we decompose the relevant terms of $\mathbf{X}$ in Eq. (7). Similarly, we also decompose the terms depending on $\mathbf{Z}$ during updating. First, to decompose $\ln\frac{(D-m_l)!(m_l-1)!}{D!}$, we define a quadratic pseudo-Boolean function:

$$f(x_{dl}) = \begin{cases} 0, \ if \ m_{l\backslash d} = 0 \ and \ x_{dl} = 0 ; \\ \ln\frac{(D-m_{l\backslash d}-x_{dl})!(m_{l\backslash d}+x_{dl}-1)!}{D!}, otherwise . \end{cases}$$

Here the subscript "$\cdot$" indicates that the given variable is determined after removing the $d^{th}$ row from $L$. Therefore the terms $\sum^{K_+}[\ln\frac{(D-m_l)!(m_l-1)!}{D!}]$ is changed

to: $\sum_{l=1}^{L_+} f(x_{dl}) = \sum_{l=1}^{L_+} x_{dl}(f(x_{dl}=1) - f(x_{dl}=0)) + f(x_{dl}=0)$. Moreover, $\ln L!$ becomes $\ln L_+! = ln(L_{+\backslash n} + \sum_{l=1}^{L_+}[\mathbf{1}_{\{m_{l\backslash d}=0\}} x_{dl}])!$, where $\mathbf{1}_{\{\cdot\}}$ is the indicator function. Here we show that the evidence lower bound Eq. (7) is well-defined in the limit $L \to \infty$.

$$\begin{aligned}\mathcal{T}(\mathbf{X}_{d\cdot}) = & -\frac{1}{2\sigma_{\mathbf{Y}}^2}(\mathbf{A}_{n\cdot}\mathbf{X}d\cdot^T)(\mathbf{A}_{n\cdot}\mathbf{X}d\cdot^T)^T + \boldsymbol{\omega}_{n\cdot}\mathbf{X}_{d\cdot} \\ & + \sum_{k=1}^{K}[\frac{(N-m_k)!(m_k-1)!}{N!} + \mathbf{1}_{\{m_{l\backslash d}=0\}} x_{dl}\boldsymbol{\varphi}_{kl}] \\ & + [x_{dl}(f(x_{dl}=1) - f(x_{dl}=0)) + f(x_{dl}=0)] \\ & - \ln \mathbf{K}! - ln(L_{+\backslash l} + \sum_{l=1}^{L_+}[\mathbf{1}_{\{m_{l\backslash d}=0\}} l_{dl}])! + const\ ,\end{aligned}$$

where $\omega_{nk} = -\frac{1}{\sigma_{\mathbf{Y}}^2}(\mathbf{A}_{n\cdot}\mathbf{Y}_{nd}^T + \boldsymbol{\gamma})$ and $\mathbf{A}_{n\cdot} = \mathbf{Z}\mathbb{E}[W]$.

### 3.5 Complexity Analysis

In this part, we show that, under a linear-Gaussian likelihood model, the per-iteration complexity of our model outperforms another recently-proposed latent feature model via IBP [9]. The iNBMT model reduces many operations when updating the parameters of non-negative matrix. $q(\mathbf{W})$ is updated twice per iteration from Eq. (9). $O(K^2L)$ operations are involved when updating $\mathbf{ZW}$, while $O(L^2K)$ operations are needed in updating $\mathbf{WX}^T$. Hence it yields a per-iteration complexity of $O(N(K^2L) + D(L^2K))$ for the $p(\mathbf{W})$ updates. The latent feature model via IBP proposed in [9] uses similar ME inference over the latent factors. Its per-iteration complexity on $q(\mathbf{W})$ is easily checked as $O(NK^2D)$. Updating $p(Y|Z)$ and $p(Y|X)$ are independent of the remaining observations and only require the computation of $\mathcal{T}(\cdot)$. We can update $\mathcal{T}(\mathbf{Z})$ in $O(N(K^2 \ln K))$ operations and $O(D(L^2 \ln L))$ operations when updating $\mathbf{X}$. The total per-iteration complexity of iNBMF is then $O(NK^2(L + \ln K) + DL^2(K + \ln L))$. The traditional model just has an infinite variable $\mathbf{Z}$, therefore its total per-iteration complexity is $O(NK^2(D + \ln K))$. In practice, $N$ and $D$ are usually sufficiently larger than $K$ and $L$, hence, the per-iteration complexity of iNBMF can be written as a simple form: $O(\alpha N + \beta D)$, while that of ME-IBP model is simplified as $O(\gamma ND)$, where $\alpha, \beta$, and $\gamma$ are small coefficients. Clearly, our proposed iNBMF has the per-iteration complexity one order lower than that of the ME-IBP model.

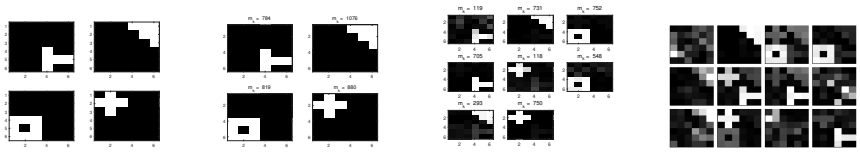
## 4 Experiments

In this section, we conduct experimental analysis of our proposed iNBMT. We study the latent features on a synthetic and a real digit dataset. We also compare the performance of iNBMT with two competitive algorithms: Maximization-Expectation-IBP (ME-IBP) and Correlated IBP-IBP (IBP-IBP).

### 4.1 Synthetic Dataset

The synthetic dataset has $4,500$ samples consisting of $6 \times 6$ grey images. Different from the dataset used in Griffiths [5], our dataset is added combination of three different luminance, illustrating as Fig. 3(b). Each row of the observations $\mathbf{Y}$ was a 36-dimension vector, which is generated by using $\mathbf{Z}$ to linearly combine a subset of the four binary factors $\mathbf{X}$. And $\mathbf{W}$ is loading different luminance combination (see Fig. 2(a)). The input datasets are shown in Fig. 3(a) by adding Gaussian noise $\sigma = 0.8$.

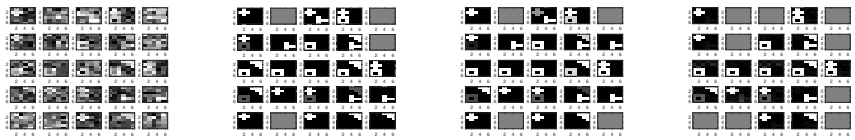
The first demonstration shown in Fig. 2 is used to evaluate various algorithms' ability to extract the latent features from the generated data. Fig. 2(c) shows the inferred features are closely match the truth features, however, each feature is repeated twice and have some noise. Compared with ME-IBP, the learning features of IBP-IBP shown in Fig. 2(c) also repeated and learning more noise. It is obvious that iNBMT outperforms other competitors by perfectly matching the truth features as well as identifying the feature number automatically.



(a) Ground-truth latent features (b) Features learned by iNBMT (c) Features learned by ME-IBP (d) Features learned by IBP-IBP

**Fig. 2.** Comparison of iNBMT, ME-IBP and IBP-IBP on synthetic dataset. iNBMT perfectly matches the truth features.

We also show the reconstruction power of our iNBMT model in Fig. 3.[1]



(a) Groundtruth (b) Groundtruth without noise (c) The reconstruction by NMBF (d) The reconstruction by ME-IBP

**Fig. 3.** Comparison of sample reconstruction on synthetic data. iNBMT best matches the groundtruth than ME-IBP.

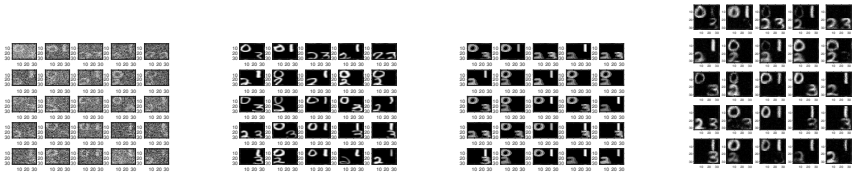
### 4.2 Digit Dataset

In this experiment, we further demonstrate the power of our iNBMT model on handwritten digit images. The digit dataset contains $2,000$ $64 \times 64$ samples

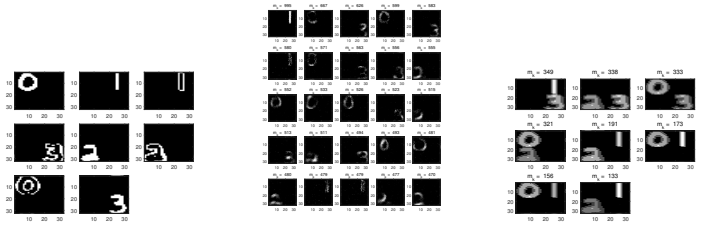
[1] Since IBP-IBP is mainly for clustering, we do not show its (almost messy) reconstruction results for fairness.

which are randomly combined with the digits $0, 1, 2, 3$ from the USPS dataset. We then corrupted the images with Gaussian noise $\sigma = 0.8$. Some examples of the randomly generated images and their corrupted version are shown in Fig. 4 (a) and (b).

It is interesting to see from Fig. 4(e), our proposed iNBMT not only captures the latent features, i.e., each of the clear digits, but also their image contours. Moreover, from the framework of iNBMT, $\mathbf{W} \times \mathbf{X}^T$ can be thought of as a set of basis images which can be added together with binary coefficients $\mathbf{Z}$ to recover images. In particular, Fig.4(g) shows the basis images which are captured by iNBMT. It is apparent that all digit combinations are detected. In terms of reconstruction, iNBMT almost perfectly recovers the images, as shown in Fig. 4(c). In comparison, ME-IBP extracts almost every different digit as the latent features, and their reconstruction results are also worse than our method.



(a) Corrupted groundtruth (b) Groundtruth w/o noise (c) Reconstruction by NMBF (d) Reconstruction by ME-IBP



(e) Features learned by iNBMT (f) Features learned by ME-IBP (g) Basis images learned by iNBMT

**Fig. 4.** Comparison of iNBMT and ME-IBP on Digits dataset. iNBMT clearly shows the best performance. We did not report IBP-IBP, since it is difficult to obtain reasonable results in this data set.

## 5 Conclusion

This paper proposes a new Bayesian model called infinite non-negative binary matrix tri-factorizations model (iNBMT), capable of learning automatically the latent binary features as well as feature number based on Indian Buffet Process (IBP). iNBMT engages a tri-factorization process that decomposes a nonnegative matrix into the product of three components including two binary matrices and a non-negative real matrix; this is also different from bi-factorization exploited

by many other NMF models. A series of experiments show that our proposed model outperforms the other competitive algorithms.

## Acknowledgement

The paper was supported by the National Basic Research Program of China (2012CB316301), National Science Foundation of China (NSFC 61473236), and Jiangsu University Natural Science Research Programme (14KJB520037).

## References

1. Chris Ding, Xiaofeng He, and Horst D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *SIAM International Conference on Data Mining*, 2005.
2. Chris Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 126–135. Press, 2006.
3. Finale Doshi-velez, Kurt T. Miller, Jurgen Van Gael, and Yee Whye Teh. Variational inference for the indian buffet process. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS*, pages 137–144, 2009.
4. Zoubin Ghahramani and Matthew J. Beal. Propagation algorithms for variational bayesian learning. In *Advances in Neural Information Processing Systems 13*, pages 507–513, 2000.
5. Thomas L. Griffiths and Zoubin Ghahramani. Infinite latent feature models and the indian buffet process. In *Advances in Neural Information Processing Systems 18*, pages 475–482, 2005.
6. David A. Knowles and Zoubin Ghahramani. Infinite sparse factor analysis and infinite independent components analysis. In *Independent Component Analysis and Signal Separation, 7th International Conference*, pages 381–388, 2007.
7. Kenichi Kurihara and Max Welling. Bayesian $k$-means as a "maximization-expectation" algorithm. In *Neural Computation*, volume 21(4), pages 1145–1172, 2009.
8. Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562. MIT Press, 2001.
9. Colorado Reed and Zoubin Ghahramani. Scaling the indian buffet process via submodular maximization. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1013–1021, 2013.
10. Zhongyuan Zhang, Tao Li, Chris H. Q. Ding, Xian-Wen Ren, and Xiang-Sun Zhang. Binary matrix factorization for analyzing gene expression data. *Data Min. Knowl. Discov.*, 20(1):28–52, 2010.