Probabilistic Semi-Nonnegative Matrix Factorization via Maximum-Likelihood and Variational Inference

Junbin Liu, Mingjie Shao and Wing-Kin Ma
Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR of China
liujunbin@link.cuhk.edu.hk, mjshao@link.cuhk.edu.hk, wkma@ee.cuhk.edu.hk

***Abstract*—Nonnegative matrix factorization (NMF) seeks a nonnegative low-rank factorization of a high-dimensional data matrix. It is a fundamental research area with various applications in signal processing and machine learning. This paper proposes a probabilistic semi-NMF formulation, motivated by a geometric semi-NMF method that has identifiability guarantee in theory. We reveal insight into how the proposed probabilistic formulation is related to the volume-minimization semi-NMF formulation, which is known to have powerful identifiability guarantee in the noiseless case, and how the proposed probabilistic formulation accommodates noise from the perspective of geometric semi-NMF. We build an algorithm for the proposed probabilistic formulation by the variational inference technique, wherein we employ a specific Dirichlet variational scheme to tackle a technical challenge, namely, an intractable integral arising from the proposed formulation. We test the proposed method on both simulated data and a real-world dataset. The proposed method shows noise robustness.**

***Index Terms*—Semi-nonnegative matrix factorization, maximum likelihood, variational inference**

I. Introduction

Non-negative matrix factorization (NMF) is a basic research problem in data analysis and signal processing. By postulating a low-rank structure underlying the data matrix, NMF aims to discover the two nonnegative factors and use them for different purposes such as clustering, low-dimensional representation, and source unmixing. There are a wide variety of forms with NMF. For example, in hyperspectral unmixing of remote sensing images, a simplex structure is imposed on the factor to represent the weights of different materials in every image pixel [1]. In blind source separation, one factor may take negative values and one may drop the non-negative constraint on one matrix factor [2], [3]. Also, the study of NMF and its variants lay foundation for recent advanced NMF designs that handle non-linearity and multilayer factors [4], [5].

This paper considers semi-NMF that relaxes the non-negative constraint on one factor. Semi-NMF can be tackled by different approaches. It can be handled by adapting the multiplicative update algorithm proposed for NMF [6]. It can also be tackled by the volume minimization (VolMin) approach for simplex-structured NMF [7], [8]. In simplex-structured NMF, VolMin aims to find the smallest data-encompassing simplex, and the vertices of this simplex form the NMF factors. Fu *et al.* [9] found that this idea can be repurposed for the more general semi-NMF problem without the simplex structure. It is shown that, in the noiseless case, VolMin provides powerful theoretical guarantee on the identifiability of the matrix factors.

Recently, Wu *et al.* [10] proposed a probabilistic method for simplex-structured semi-NMF in the presence of additive Gaussian noise. It is shown that the probabilistic method has a close connection with the VolMin approach. More specifically, VolMin may be seen as the probabilistic method when there is no noise. It is worth noting that similar connections are drawn in some later works [11], [12].

Inspired by this, in this paper, we extend the idea of the probabilistic method in [10] to tackle a general semi-NMF problem in the presence of additive Gaussian noise. In the same spirit as our previous work, we present an connection between our probabilistic formulation and the VolMin approach for semi-NMF. The challenge with the probabilistic semi-NMF approach (both our formulation and some formulations in prior studies) is that we need to maximize a likelihood function that appears as an intractable integral. We employ variational inference (VI) [13] to deal with this challenge. Both synthetic and real-data experiments are carried out to illustrate the performance of the proposed method.

Some related prior works should be mentioned. In the literature, there are various probabilistic formulations and inference methods for NMF or semi-NMF. In particular, we have seen a rich variety of the generative models, non-negative distribution models for the matrix factors, and the noise models [14]–[21]. None of the existing studies puts a link between the probabilistic method and the geometric VolMin approach. Our study seeks to draw a connection between the two, and leverage on that to develop a semi-NMF algorithm that may inherit the fundamental merits of VolMin.

II. Problem Formulation

We begin with an NMF model:

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{Z}^\top + \boldsymbol{W}, \tag{1}$$

where $\boldsymbol{Y} \in \mathbb{R}^{M\times N}$ is the observed data; $\boldsymbol{A} \in \mathbb{R}^{M\times K}$ and $\boldsymbol{Z} \in \mathbb{R}_+^{N\times K}$ are the underlying matrix factors, with $K \leq \min\{M, N\}$ and with $\mathbb{R}_+$ denoting the set of non-negative numbers; $\boldsymbol{W} \in \mathbb{R}^{M\times N}$ is the noise matrix whose elements independently and identically follow Gaussian distribution with mean zero and variance σ^2. Note that we do not restrict $\boldsymbol{A}$ to be nonnegative, and hence we are considering a

This work was supported by a General Research Fund (GRF) of Hong Kong Research Grant Council (RGC) under Project ID CUHK 14203721.

semi-NMF problem. Given the data $\boldsymbol{Y}$, we want to retrieve $\boldsymbol{A}$ and $\boldsymbol{Z}$.

Without loss of generality, we assume that each column of $\boldsymbol{Z}$ has unit ℓ_1-norm. That, together with the nonnegativity of $\boldsymbol{Z}$, means that

$$\boldsymbol{z}_i \in \Delta := \left\{\boldsymbol{z} \in \mathbb{R}_+^N \mid \mathbf{1}^\top \boldsymbol{z} = 1\right\}, \quad i = 1, 2, ..., K.$$

Assuming no noise, i.e., $\boldsymbol{W} = \mathbf{0}$, the authors in [9] propose a volume minimization (VolMin) approach

$$\min_{\boldsymbol{A},\boldsymbol{Z}} \det\left(\boldsymbol{A}^\top \boldsymbol{A}\right), \text{ s.t. } \boldsymbol{Y} = \boldsymbol{A}\boldsymbol{Z}^\top, \ \boldsymbol{Z}^\top \mathbf{1} = \mathbf{1}, \ \boldsymbol{Z} \geq \mathbf{0}, \quad (2)$$

where $\det(\cdot)$ denotes the matrix determinant. VolMin is known to be a powerful approach in the noiseless case. Specifically, under some fairly mild assumptions with the true $\boldsymbol{A}$ and $\boldsymbol{Z}$, it is shown that VolMin can uniquely identify the true $\boldsymbol{A}$ and $\boldsymbol{Z}$ (subject to the minor effects of column and row permutations with $\boldsymbol{A}$ and $\boldsymbol{Z}$, respectively). We refer the readers to the literature [22], [23] for details. In the noisy case, it is typical to consider the following penalty variation for (2)

$$\begin{aligned} &\min_{\boldsymbol{A},\boldsymbol{Z}} \det\left(\boldsymbol{A}^\top \boldsymbol{A}\right) + \lambda \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{Z}^\top\|_F^2 \\ &\text{s. t. } \boldsymbol{Y} = \boldsymbol{A}\boldsymbol{Z}^\top, \ \boldsymbol{Z}^\top \mathbf{1} = \mathbf{1}, \ \boldsymbol{Z} \geq \mathbf{0}, \end{aligned} \quad (3)$$

where $\lambda \geq 0$ is a parameter that balances volume minimization and data fitting. In practice, λ is often manually tuned.

In this study, we focus on a probabilistic formulation for NMF. To handle the constraints placed on $\boldsymbol{Z}$, we assume that the columns of $\boldsymbol{Z}$ are independently and uniformly distributed on the unit simplex Δ, i.e., the columns of $\boldsymbol{Z}$ follow a uniform Dirichlet distribution

$$p(\boldsymbol{Z}) = \prod_{k=1}^K p(\boldsymbol{z}_k) = \prod_{k=1}^K \mathcal{D}(\boldsymbol{z}_k; \mathbf{1}), \quad (4)$$

where

$$\mathcal{D}(\boldsymbol{z}; \boldsymbol{\beta}) = \frac{1}{\mathcal{B}(\boldsymbol{\beta})} \prod_{i=1}^K z_i^{\beta_i - 1}$$

denotes the Dirichlet distribution parameterized by $\boldsymbol{\beta} > \mathbf{0}$; $\mathcal{B}(\boldsymbol{\beta}) = \left(\prod_{i=1}^N \Gamma(\beta_i)\right) / \Gamma\left(\sum_{i=1}^N \beta_i\right)$ is the multivariate Beta function with $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \, \mathrm{d}t$ being the Gamma function. We view $\boldsymbol{A}$ as a deterministic parameter, and estimate it via maximizing the likelihood of $\boldsymbol{Y}$,

$$\max_{\boldsymbol{\theta}} \log p(\boldsymbol{Y}; \boldsymbol{\theta}) = \log\left(\int p(\boldsymbol{Y}|\boldsymbol{Z}; \boldsymbol{\theta}) p(\boldsymbol{Z}) d\boldsymbol{Z}\right), \quad (5)$$

where $\boldsymbol{\theta} = \{\boldsymbol{A}, \sigma^2\}$ is the parameter to be estimated (we also estimate the noise variance); $p(\boldsymbol{Y}|\boldsymbol{Z}; \boldsymbol{\theta})$ takes a Gaussian form

$$p(\boldsymbol{Y}|\boldsymbol{Z}; \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{Y}; \boldsymbol{A}\boldsymbol{Z}^\top, \ \sigma^2 \boldsymbol{I}). \quad (6)$$

III. Connections between Maximizing the Likelihood and Volume Minimization

In this section, we reveal a connection between the probabilistic maximum-likelihood (ML) formulation (5) and the geometric VolMin formulation (2). For simplicity, we assume $M = K$. Assuming that $\boldsymbol{A}$ has full column rank and changing the variable $\boldsymbol{Z}$ by $\boldsymbol{A}^{-1}\boldsymbol{Y} = \boldsymbol{Z}^\top$, we can equivalently transform (2) into

$$\min_{\boldsymbol{A}} \det\left(\boldsymbol{A}^\top \boldsymbol{A}\right), \text{ s. t. } \boldsymbol{A}^{-1}\boldsymbol{Y}\mathbf{1} = \mathbf{1}, \ \boldsymbol{A}^{-1}\boldsymbol{Y} \geq \mathbf{0}. \quad (7)$$

The proof of (7) is similar to that in [24]. We then turn to the probabilistic formulation (5). Assuming that σ^2 is known, the likelihood is given by

$$p(\boldsymbol{Y}; \boldsymbol{A}) = \int \mathcal{N}(\boldsymbol{Y}; \boldsymbol{A}\boldsymbol{Z}^\top, \sigma^2 \boldsymbol{I}) \prod_{k=1}^K \mathbb{I}_\Delta(\boldsymbol{z}_k) d\boldsymbol{Z}, \quad (8)$$

where $\mathbb{I}_{\mathcal{S}}(\boldsymbol{z})$ is the indicator function, i.e., $\mathbb{I}_{\mathcal{S}}(\boldsymbol{z}) = 1$ if $x \in \mathcal{S}$ and $\mathbb{I}_{\mathcal{S}}(\boldsymbol{z}) = 0$ if $x \notin \mathcal{S}$. By closely following the approximation presented in [12] with some subtle modifications, one can show the following,

$$-\log p(\boldsymbol{Y}; \boldsymbol{A}) \propto \log\det\left(\boldsymbol{A}^\top \boldsymbol{A}\right) + f(\boldsymbol{A}) + g(\boldsymbol{A}), \quad (9)$$

where

$$f(\boldsymbol{A}) = -\frac{2}{N} \sum_{k=1}^K \sum_{i=1}^N \log \Phi\left(\frac{\hat{\boldsymbol{a}}_k^\top \boldsymbol{y}_i}{\sigma \|\hat{\boldsymbol{a}}_k\|}\right),$$

$$g(\boldsymbol{A}) = \frac{1}{N} \sum_{k=1}^K \left[\log N\sigma^2 \|\hat{\boldsymbol{a}}_k\|^2 + \frac{\left(1 - \hat{\boldsymbol{a}}_k^\top \boldsymbol{Y}\mathbf{1}\right)^2}{N\sigma^2 \|\hat{\boldsymbol{a}}_k\|^2}\right],$$

with $\hat{\boldsymbol{a}}_k^\top$ being the kth row of $\boldsymbol{A}^{-1}$; $\Phi(x) = \int_{-\infty}^x e^{-z^2/2} dz / \sqrt{2\pi}$. We shall omit the derivation and focus on the revelations. Comparing (7) and (9) we see that they both minimize the volume of $\boldsymbol{A}$. By noticing the function $-\log \Phi(x)$ has large value when x is negative, we can regard $f(\boldsymbol{A})$ as a soft constraint for $\boldsymbol{A}^{-1}\boldsymbol{Y} \geq \mathbf{0}$ in (9). Similarly, $g(\boldsymbol{A})$ serves as a soft constraint to enforce $\boldsymbol{A}^{-1}\boldsymbol{Y}\mathbf{1} = \mathbf{1}$. In this regard, we may view the ML problem (5) as minimizing the volume of $\boldsymbol{A}$ with penalty terms tailored to accommodate noise.

IV. Maximizing the Likelihood via Variational Inference

Now, we focus on algorithm design for the ML problem (5). The crux of optimizing (5) lies in the integral, which is intractable in general. We utilize the variational inference (VI) technique in [10] to deal with the problem. The idea of VI is to find a tractable lower bound of $\log p(\boldsymbol{Y}; \boldsymbol{\theta})$ and maximize the lower bound instead. To start with, we introduce an arbitrary distribution $q(\boldsymbol{Z})$ that has the same support as $p(\boldsymbol{Z})$. By Jensen's inequality, we have

$$\begin{aligned} \log p(\boldsymbol{Y}; \boldsymbol{\theta}) &= \log\left(\int p(\boldsymbol{Y}|\boldsymbol{Z}; \boldsymbol{\theta}) p(\boldsymbol{Z}) \frac{q(\boldsymbol{Z})}{q(\boldsymbol{Z})} d\boldsymbol{Z}\right) \\ &\geq \mathbb{E}_{q(\boldsymbol{Z})}\left[\log \frac{p(\boldsymbol{Y}|\boldsymbol{Z}; \boldsymbol{\theta}) p(\boldsymbol{Z})}{q(\boldsymbol{Z})}\right] \end{aligned}, \quad (10)$$

where equality is achieved when $q(\boldsymbol{Z}) = p(\boldsymbol{Z} \mid \boldsymbol{Y}; \boldsymbol{\theta})$. The idea of VI is to restrict $q(\boldsymbol{Z})$ such that the above lower bound is tractable. Specifically, we restrict $q(\boldsymbol{Z})$ to take a factored form across the columns of $\boldsymbol{Z}$,

$$q(\boldsymbol{Z}) = \prod_{k=1}^K q(\boldsymbol{z}_k) = \prod_{k=1}^K \mathcal{D}(\boldsymbol{z}_k; \boldsymbol{\beta}_k), \quad (11)$$

where we choose $q(\boldsymbol{z}_k)$ as Dirichlet distribution $\mathcal{D}(\cdot;\boldsymbol{\beta}_k)$ with parameters $\boldsymbol{\beta}_k > 0,\ k = 1, 2, ..., K$. Substituting $q(\boldsymbol{Z})$ back, we obtain the following variational approximation of the ML problem:

$$\max_{\boldsymbol{B}\in\mathbb{R}_{++}^{N\times K},\ \boldsymbol{\theta}} \mathbb{E}_q\left[\log p\left(\boldsymbol{Y},\boldsymbol{Z};\boldsymbol{\theta}\right) - \sum_{k=1}^{K}\log\mathcal{D}(\boldsymbol{z}_k;\boldsymbol{\beta}_k)\right], \quad (12)$$

where the columns of $\boldsymbol{B}$ collects the variational parameters, i.e., the $\boldsymbol{\beta}_k$'s; the optimization is with respect to $\{\boldsymbol{B},\boldsymbol{\theta}\}$. Our strategy is to apply alternating maximization to tackle the problem. Specifically, in the tth iteration, we perform

$$\boldsymbol{A}^{(t+1)} = \arg\max_{\boldsymbol{A}} \mathbb{E}_{q^{(t)}}\left[\log p\left(\boldsymbol{Y},\boldsymbol{Z};\boldsymbol{A},(\sigma^2)^{(t)}\right)\right], \quad \text{(P1)}$$

$$(\sigma^2)^{(t+1)} = \arg\max_{\sigma^2>0} \mathbb{E}_{q^{(t)}}\left[\log p\left(\boldsymbol{Y},\boldsymbol{Z};\boldsymbol{A}^{(t+1)},\sigma^2\right)\right], \quad \text{(P2)}$$

$$\boldsymbol{B}^{(t+1)} = \arg\max_{\boldsymbol{B}\in\mathbb{R}_{++}^{N\times K}} \mathbb{E}_q\left[\log p\left(\boldsymbol{Y},\boldsymbol{Z};\boldsymbol{\theta}^{(t+1)}\right) - \sum_{k=1}^{K}\log\mathcal{D}(\boldsymbol{z}_k;\boldsymbol{\beta}_k)\right]. \quad \text{(P3)}$$

A. Update the model parameters $\boldsymbol{A}$ and σ^2

Given the variational parameter $\boldsymbol{B}$, the model parameters $\boldsymbol{A}$ and σ^2 can be updated as follows. It can be shown that the objective of the sub-problem (P1) has a quadratic form w.r.t. $\boldsymbol{A}$:

$$\begin{aligned}&\mathbb{E}_q\left[\log p\left(\boldsymbol{Y},\boldsymbol{Z};\boldsymbol{A},\sigma^2\right)\right]\\ \propto&\frac{1}{2\sigma^2}\left[2\mathrm{Tr}\left(\boldsymbol{Y}^\top\boldsymbol{A}\mathbb{E}_q(\boldsymbol{Z})^\top\right) - \mathrm{Tr}\left(\boldsymbol{A}^\top\boldsymbol{A}\mathbb{E}_q\left(\boldsymbol{Z}^\top\boldsymbol{Z}\right)\right)\right].\end{aligned} \quad (13)$$

Hence, given other variables, the optimal $\boldsymbol{A}$ admits a closed-form expression

$$\boldsymbol{A} = \boldsymbol{Y}\mathbb{E}_q(\boldsymbol{Z})\left(\mathbb{E}_q\left(\boldsymbol{Z}^\top\boldsymbol{Z}\right)\right)^{-1}. \quad (14)$$

Using the moment results for Dirichlet distributions, it can be shown that

$$\mathbb{E}_q(\boldsymbol{Z}) = \boldsymbol{B}\mathrm{Diag}^{-1}\left(\boldsymbol{B}^\top\mathbf{1}\right), \quad (15)$$

with $\mathrm{Diag}(\boldsymbol{x})$ denoting a diagonal matrix with the diagonal elements given by $\boldsymbol{x}$. Also we have,

$$\left[\mathbb{E}_q\left(\boldsymbol{Z}^\top\boldsymbol{Z}\right)\right]_{ij} = \begin{cases}\frac{\boldsymbol{\beta}_i^\top\boldsymbol{\beta}_j}{\mathbf{1}^\top\boldsymbol{\beta}_i\mathbf{1}^\top\boldsymbol{\beta}_j}, & i\neq j\\ \frac{\mathbf{1}^\top\left(\boldsymbol{\beta}_i^2+\boldsymbol{\beta}_i\right)}{\mathbf{1}^\top\boldsymbol{\beta}_i(\mathbf{1}^\top\boldsymbol{\beta}_i+1)}, & i=j\end{cases}, \quad (16)$$

where the square on a vector is element-wise.

For the objective function in (P2), we can write,

$$\begin{aligned}&\mathbb{E}_q\left[\log p\left(\boldsymbol{Y},\boldsymbol{Z};\boldsymbol{A},\sigma^2\right)\right]\\ \propto&-\frac{MN}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\mathbb{E}_q\left[\left\|\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{Z}^\top\right\|_F^2\right].\end{aligned} \quad (17)$$

It can be shown that (17) is maximized w.r.t. σ^2 if

$$\sigma^2 = \frac{1}{MN}\mathbb{E}_q\left[\left\|\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{Z}^\top\right\|_F^2\right], \quad (18)$$

where the expectation can be evaluated based on (15) and (16).

B. Update the variational parameters

We write the objective of (P3) as

$$\begin{aligned}&\mathbb{E}_q\left[\log p\left(\boldsymbol{Y},\boldsymbol{Z};\boldsymbol{A},\sigma^2\right) - \sum_{k=1}^{K}\log\mathcal{D}(\boldsymbol{z}_k;\boldsymbol{\beta}_k)\right]\\ \propto&-\frac{1}{2\sigma^2}\mathbb{E}_q\left[\left\|\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{Z}^\top\right\|_F^2\right] + \sum_{k=1}^{K}H(\boldsymbol{\beta}_k),\end{aligned} \quad (19)$$

where the expectation is evaluated similarly to that in (18); $H(\boldsymbol{\beta})$ is the entropy of the Dirichlet distribution which takes the form

$$H(\boldsymbol{\beta}) = \log\mathcal{B}(\boldsymbol{\beta}) - \sum_{i=1}^{N}(\beta_i-1)\left(\psi\left(\beta_i\right) - \psi\left(\boldsymbol{\beta}^\top\mathbf{1}\right)\right), \quad (20)$$

with $\psi(x) = \frac{\mathrm{d}\log\Gamma(x)}{\mathrm{d}x}$ referring to the digamma function.

The sub-problem (P3) is non-convex and does not have a closed-form solution. We use an accelerated gradient ascent method [25] to handle this sub-problem. The accelerated gradient ascent method was developed for convex problems, but it has been found to work well empirically in a number of applications, see, e.g., [26]–[28]. The gradient of the objective w.r.t. $\boldsymbol{\beta}_k$ is given by

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}_k}\left(-\frac{1}{2\sigma^2}\mathbb{E}_q\left[\left\|\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{Z}^\top\right\|_F^2\right] + \sum_{k=1}^{K}H(\boldsymbol{\beta}_k)\right)\\ =&\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}_k}H(\boldsymbol{\beta}_k) + \frac{1}{\sigma^2}\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}_k}\left(\frac{\boldsymbol{\beta}_k}{\mathbf{1}^\top\boldsymbol{\beta}_k}\right)^\top\overline{\boldsymbol{Y}}_k^\top\boldsymbol{a}_k\\ &-\frac{\|\boldsymbol{a}_k\|^2}{2\sigma^2}\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}_k}\mathbb{E}_{q(\boldsymbol{z}_k)}\left(\boldsymbol{z}_k^\top\boldsymbol{z}_k\right)\end{aligned}, \quad (21)$$

with

$$\overline{\boldsymbol{Y}}_k = \boldsymbol{Y} - \sum_{i\neq k}\boldsymbol{a}_i\left(\frac{\boldsymbol{\beta}_i}{\mathbf{1}^\top\boldsymbol{\beta}_i}\right)^\top, \quad (22)$$

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}_k}H(\boldsymbol{\beta}_k) = (\mathbf{1}^\top\boldsymbol{\beta}_k - N)\psi'(\mathbf{1}^\top\boldsymbol{\beta}_k)\mathbf{1} - \psi'(\boldsymbol{\beta}_k)\odot(\boldsymbol{\beta}_k-\mathbf{1}), \quad (23)$$

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}_k}\left(\frac{\boldsymbol{\beta}_k}{\mathbf{1}^\top\boldsymbol{\beta}_k}\right)^\top\overline{\boldsymbol{Y}}_k^\top\boldsymbol{a}_k = \frac{\overline{\boldsymbol{Y}}_k^\top\boldsymbol{a}_k}{\mathbf{1}^\top\boldsymbol{\beta}_k} - \frac{\boldsymbol{\beta}_k^\top\overline{\boldsymbol{Y}}_k^\top\boldsymbol{a}_k}{(\mathbf{1}^\top\boldsymbol{\beta}_k)^2}\mathbf{1}, \quad (24)$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\beta}_k}\mathbb{E}_{q(\boldsymbol{z}_k)}\left(\boldsymbol{z}_k^\top\boldsymbol{z}_k\right) =& \frac{1}{\mathbf{1}^\top\boldsymbol{\beta}_k(\mathbf{1}^\top\boldsymbol{\beta}_k+1)}\left(2\boldsymbol{\beta}_k+\mathbf{1}\right)\\ &-\frac{\mathbf{1}^\top(\boldsymbol{\beta}_k^2+\boldsymbol{\beta}_k)(2\mathbf{1}^\top\boldsymbol{\beta}_k+1)}{\left(\mathbf{1}^\top\boldsymbol{\beta}_k\right)^2\left(\mathbf{1}^\top\boldsymbol{\beta}_k+1\right)^2}\mathbf{1}\end{aligned}, \quad (25)$$

where $\odot$ means element-wise multiplication. Putting the above pieces together, we obtain the entire algorithm.

V. Simulations

In this section, we provide simulation results of testing the proposed algorithm. The main benchmarking algorithms are the SNMF in [6] and the VolMin in [9] with open-source codes. We also compare the proposed method with two probabilistic methods, a variational inference based NMF

method denoted as VBNMF [14] and a maximum *a posteriori* NMF method [19] denoted as MAPNMF, in the real-world dataset experiment. We denote the proposed method as VISNMF. We stop the algorithms, SNMF and VISNMF, when $\left\|\boldsymbol{A}^{(t+1)}-\boldsymbol{A}^{(t)}\right\|_{\mathrm{F}}<10^{-10}$ or the maximum number of iterations 1000 is achieved. For VolMin, we use the default settings in its open-source codes.

A. Simulated data

We evaluate algorithms' performance based on the mean square error (MSE) of estimating factor $\boldsymbol{A}$:

$$\mathrm{MSE}(\boldsymbol{A},\overline{\boldsymbol{A}}):=\frac{\|\boldsymbol{A}\boldsymbol{D}_1-\overline{\boldsymbol{A}}\boldsymbol{\Pi}\boldsymbol{D}_2\|_F^2}{\|\boldsymbol{A}\boldsymbol{D}_1\|_F^2}, \tag{26}$$

where $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are positive diagonal matrices normalizing the columns of $\boldsymbol{A}$ and $\overline{\boldsymbol{A}}$ respectively; $\boldsymbol{\Pi}$ is a proper permutation matrix to align the columns. The signal-to-noise ratio (SNR) is defined as $\|\boldsymbol{A}\boldsymbol{Z}^\top\|_F^2/(\sigma^2MN)$. All the results are obtained by averaging over 50 independent trials.

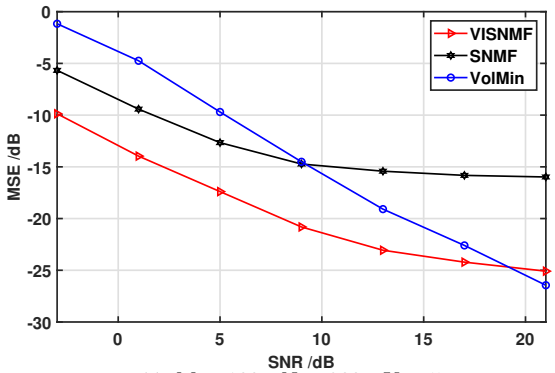
We generate the factor $\boldsymbol{A}$ with elements independently drawn from $\mathcal{N}(0,1)$; $\boldsymbol{Z}$ is simulated according to the model assumption, i.e., columns of $\boldsymbol{Z}$ independently follow $\mathcal{D}(\cdot;\mathbf{1})$. The results are shown in Fig. 1. From the results we can see that VISNMF performs better than the other algorithms in the low SNR region. As the SNR increases, the performance of VolMin improves and is comparable to that of VISNMF. The running times are shown in Table I. SNMF runs very fast. VISNMF and VolMin have relatively slow computation speeds, with VISNMF being slightly faster than VolMin.

	M=100, N=200	M=200, N=200	M=200, N=100
SNMF	0.0099s	0.0125s	0.0681s
VolMin	2.7585s	2.7321s	2.7096s
VISNMF	1.8881s	2.3109s	1.4156s

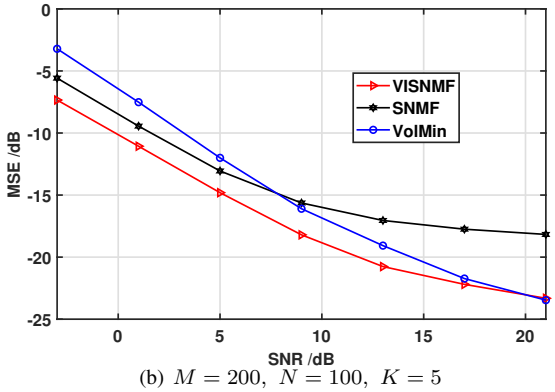
TABLE I
RUNNING TIME. $K=5$, SNR $=10dB$.

In Fig. 2, we present the impact of the number of the factors, i.e., K, when the size of $\boldsymbol{Y}$ is fixed. The SNR is set as 10 dB. From the result we see that the performances of all the considered algorithms become worse as K increases. SNMF and VISNMF deliver reasonable performance for all the tested K's, but VolMin does not perform satisfactorily for large K's.

We then test two cases where the columns of $\boldsymbol{Z}$ are not independently and uniformly drawn from the unit simplex. We let $\boldsymbol{Z}$ be sufficiently scattered [29], under which VolMin has theoretical guarantee of perfect retrieval of $\boldsymbol{A}$ if the noise is absent. We generate the elements of $\boldsymbol{Z}$ independently and uniformly from the interval $[0,1]$ and randomly zeroing out 35% elements, such that $\boldsymbol{Z}$ is sufficiently scattered with high probability [30]. The results are shown in Fig. 3. We see that VISNMF delivers reasonably good results while VolMin delivers good performance for high SNRs. As a summary, the above numerical results suggest that VISNMF shows good robustness to noise.



(a) $M=100,\ N=200,\ K=5$



(b) $M=200,\ N=100,\ K=5$

Fig. 1. MSE vs. SNR under different problem sizes.

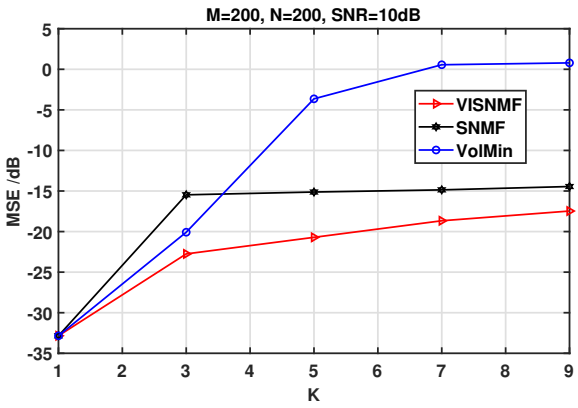


Fig. 2. MSE vs. K.

B. Noisy image representation learning

We test the algorithms on a face factor learning task. We use the Frey Faces data set, which contains about 2000 images of Brendan's face with size 20×28 taken from sequential frames of a small video. We add additive white Gaussian noise on the original images (SNR=10 dB). Fig. 4 shows the learned factors, i.e., the columns of $\boldsymbol{A}$ reshaped to the image size. From the results, the five algorithms all learn meaningful factors from the noisy data. Arguably, VISNMF and MAPNMF give more sensible factors.

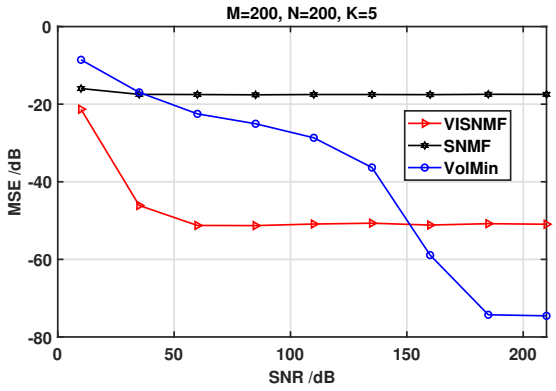


Fig. 3. MSE vs. SNR with $\boldsymbol{Z}$ being sufficiently scattered.

Fig. 4. Learned factors. Left: $K = 4$. Right: $K = 5$. From the top row to the bottom: VolMin, SNMF, VISNMF, VBNMF, and MAPNMF. SNR=10dB.

VI. Conclusion

To conclude, we studied a probabilistic formulation of semi-NMF. We drew connections between our probabilistic formulation and the geometric method presented in [9], and we derived a variational inference-based algorithm for our probabilistic formulation. By simulations, the proposed algorithm exhibits effectiveness in noisy cases.

References

[1] J. M. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot, "Hyperspectral remote sensing data analysis and future challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 2, pp. 6–36, 2013.

[2] X. Fu, W.-K. Ma, K. Huang, and N. D. Sidiropoulos, "Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain," *IEEE Trans. Signal Process.*, vol. 63, no. 9, pp. 2306–2320, 2015.

[3] M. D. Plumbley, "Algorithms for nonnegative independent component analysis," *IEEE Trans. Neural Netw.*, vol. 14, no. 3, pp. 534–543, 2003.

[4] Q. Lyu and X. Fu, "Identifiability-guaranteed simplex-structured post-nonlinear mixture learning via autoencoder," *IEEE Trans. Signal Process.*, vol. 69, pp. 4921–4936, 2021.

[5] B. Yang, X. Fu, N. D. Sidiropoulos, and K. Huang, "Learning nonlinear mixtures: Identifiability and algorithm," *IEEE Trans. Signal Process.*, vol. 68, pp. 2857–2869, 2020.

[6] C. H. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 45–55, 2008.

[7] M. D. Craig, "Minimum-volume transforms for remotely sensed data," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 3, pp. 542–552, 1994.

[8] J. Li and J. M. Bioucas-Dias, "Minimum volume simplex analysis: A fast algorithm to unmix hyperspectral data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp*, vol. 3, pp. 250–253, 2008.

[9] X. Fu, K. Huang, and N. D. Sidiropoulos, "On identifiability of nonnegative matrix factorization," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 328–332, 2018.

[10] R. Wu, W.-K. Ma, Y. Li, A. M.-C. So, and N. D. Sidiropoulos, "Probabilistic simplex component analysis," *IEEE Trans. Signal Process.*, vol. 70, pp. 582–599, 2021.

[11] Y. Chen, S. He, Y. Yang, and F. Liang, "Learning topic models: Identifiability and finite-sample analysis," *J. Am. Stat. Assoc.*, pp. 1–16, 2022.

[12] C. Huang, M. Shao, W.-K. Ma, and A. M.-C. So, "SISAL revisited," *SIAM J. Imag. Sci.*, vol. 15, no. 2, pp. 591–624, 2022.

[13] 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.

[14] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computat. Intell. Neurosci.*, vol. 2009, 2009.

[15] T. Virtanen, A. T. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 1825–1828, Las Vegas, Nevada, USA, 2008.

[16] D. Liang, M. D. Hoffman, and D. P. Ellis, "Beta process sparse nonnegative matrix factorization for music.," in *ISMIR*, pp. 375–380, 2013.

[17] A. Vilamala, L. A. Belanche, and A. Vellido, "A map approach for convex non-negative matrix factorization in the diagnosis of brain tumors," in *2014 Int. Workshop Pattern Recognit. Neuroimag.*, pp. 1–4, IEEE, 2014.

[18] A. Vilamala Muñoz, A. Vellido Alcacena, and L. A. Belanche Muñoz, "Bayesian semi non-negative matrix factorisation," in *Proc. Eur. Symp. Artif. Neural Networks, Comput. Intell. Mach. Learn.*, pp. 195–200, 2016.

[19] I. Psorakis, S. Roberts, M. Ebden, and B. Sheldon, "Overlapping community detection using Bayesian non-negative matrix factorization," *Phys. Rev. E.*, vol. 83, no. 6, p. 066114, 2011.

[20] M. N. Schmidt, O. Winther, and L. K. Hansen, "Bayesian non-negative matrix factorization," in *Proc. 8th Int. Conf. Independent Component Anal. Signal Separation*, pp. 540–547, Springer, 2009.

[21] Z. Ma, A. E. Teschendorff, A. Leijon, Y. Qiao, H. Zhang, and J. Guo, "Variational Bayesian matrix factorization for bounded support data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 876–889, 2014.

[22] X. Fu, K. Huang, N. D. Sidiropoulos, and W.-K. Ma, "Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications.," *IEEE Signal Process. Mag.*, vol. 36, no. 2, pp. 59–80, 2019.

[23] N. Gillis, *Nonnegative matrix factorization*. SIAM, 2020.

[24] W.-K. Ma, "On hyperspectral unmixing," in *IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, pp. 17–20, 2021.

[25] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[26] Y. Xu and W. Yin, "A globally convergent algorithm for nonconvex optimization based on block coordinate update," *J. Sci. Comput.*, vol. 72, no. 2, pp. 700–734, 2017.

[27] J. Tranter, N. D. Sidiropoulos, X. Fu, and A. Swami, "Fast unit-modulus least squares with applications in beamforming," *IEEE Trans. Signal Process.*, vol. 65, no. 11, pp. 2875–2887, 2017.

[28] M. Shao, Q. Li, W.-K. Ma, and A. M.-C. So, "A framework for one-bit and constant-envelope precoding over multiuser massive miso channels," *IEEE Trans. Signal Process.*, vol. 67, no. 20, pp. 5309–5324, 2019.

[29] K. Huang, N. D. Sidiropoulos, and A. Swami, "Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition," *IEEE Trans. Signal Process.*, vol. 62, no. 1, pp. 211–224, 2013.

[30] H. Kim and H. Park, "Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 2, pp. 713–730, 2008.