


Article

# Survey on Probabilistic Models of Low-Rank Matrix Factorizations

Jiarong Shi <sup>1,2,\*</sup> , Xiuyun Zheng <sup>1</sup> and Wei Yang <sup>1</sup>

<sup>1</sup> School of Science, Xi'an University of Architecture and Technology, Xi'an 710055, China; xyzhengzzf@sohu.com (X.Z.); yangweipyf@163.com (W.Y.)

<sup>2</sup> School of Architecture, Xi'an University of Architecture and Technology, Xi'an 710055, China

\* Correspondence: shijiarong@xauat.edu.cn; Tel.: +86-29-8220-5670

Received: 12 June 2017; Accepted: 16 August 2017; Published: 19 August 2017

**Abstract:** Low-rank matrix factorizations such as Principal Component Analysis (PCA), Singular Value Decomposition (SVD) and Non-negative Matrix Factorization (NMF) are a large class of methods for pursuing the low-rank approximation of a given data matrix. The conventional factorization models are based on the assumption that the data matrices are contaminated stochastically by some type of noise. Thus the point estimations of low-rank components can be obtained by Maximum Likelihood (ML) estimation or Maximum a posteriori (MAP). In the past decade, a variety of probabilistic models of low-rank matrix factorizations have emerged. The most significant difference between low-rank matrix factorizations and their corresponding probabilistic models is that the latter treat the low-rank components as random variables. This paper makes a survey of the probabilistic models of low-rank matrix factorizations. Firstly, we review some probability distributions commonly-used in probabilistic models of low-rank matrix factorizations and introduce the conjugate priors of some probability distributions to simplify the Bayesian inference. Then we provide two main inference methods for probabilistic low-rank matrix factorizations, i.e., Gibbs sampling and variational Bayesian inference. Next, we classify roughly the important probabilistic models of low-rank matrix factorizations into several categories and review them respectively. The categories are performed via different matrix factorizations formulations, which mainly include PCA, matrix factorizations, robust PCA, NMF and tensor factorizations. Finally, we discuss the research issues needed to be studied in the future.

**Keywords:** matrix factorizations; low-rank; variational Bayesian inference; Gibbs sampling; probabilistic principal component analysis; probabilistic matrix factorizations; probabilistic tensor factorizations

## 1. Introduction

In many practical applications, a commonly-used assumption is that the dataset approximately lies in a low-dimensional linear subspace. Low-rank matrix factorizations (or decompositions, the two terms are used interchangeably) are just a type of method for recovering the low-rank structure, removing noise and completing missing values. Principal Component Analysis (PCA) [1], a traditional matrix factorization method, copes effectively with the situation that the dataset is contaminated by Gaussian noise. If the mean vector is set to be a zero vector, then PCA is transformed into Singular Value Decomposition (SVD) [2]. Classical PCA approximates the low-rank representation according to eigen decomposition of the covariance matrix of a dataset to be investigated. When there are outliers or large sparse errors, the original version of PCA does not work well for obtaining the low-rank representations. In this case, many robust methods of PCA have been extensively studied such as L1-PCA [3], L1-norm maximization PCA [4], L21-norm maximization PCA [5] and so on.

In this paper, low-rank matrix factorizations refer to a general formulation of matrix factorizations and they mainly consist of PCA, matrix factorizations, robust PCA [6], Non-negative Matrix

Factorization (NMF) [7] and tensor decompositions [8]. As a matter of fact, matrix factorizations are usually a special case of PCA and they directly factorize a data matrix into the product of two low-rank matrices without consideration of the mean vector. Matrix completion aims to complete all missing values according to the approximate low-rank structure and it is mathematically described as a nuclear norm minimization problem [9]. To this end, we regard matrix completion as a special case of matrix factorizations. In addition, robust PCA decomposes the data matrix into the superposition of a low-rank matrix and a sparse component, and recovers both the low-rank and the sparse matrices via principal component pursuit [6].

Low-rank matrix factorizations suppose that the low-rank components are corrupted by certain random noise and the low-rank matrices are deterministic unknown parameters. Maximum Likelihood (ML) estimation and Maximum a posteriori (MAP) are two most popular methodologies used in estimating the low-rank matrices. A prominent advantage of the aforementioned point estimate methods is that they are simple and easy to implement. However, we can not obtain the probability distributions of the low-rank matrices that are pre-requisite in exploring the generative models. Probabilistic models of low-rank matrix factorizations can approximate the true probability distributions of the low-rank matrices and they have attracted a great deal of attention in the past two decades. These models have been widely applied in the fields of signal and image processing, computer vision, machine learning and so on.

Tipping and Bishop [10] originally presented probabilistic PCA in which the latent variables are assumed to be a unit isotropic Gaussian distribution. Subsequently, several other probabilistic models of PCA were proposed successively, such as Bayesian PCA [11], Robust L1 PCA [12], Bayesian robust PCA [13] and so on. As a special case of probabilistic models of PCA, Bayesian matrix factorization [14] placed zero-mean spherical Gaussian priors on two low-rank matrices and it was applied to collaborative filtering. Probabilistic models of matrix factorizations also include probabilistic matrix factorization [15], Bayesian probabilistic matrix factorization [16], robust Bayesian matrix factorization [17], probabilistic robust matrix factorization [18], sparse Bayesian matrix completion [19], Bayesian robust matrix factorization [20] and Bayesian L1-norm low-rank matrix factorizations [21].

As to robust PCA, we take the small dense noise into account. In other words, the data matrix is decomposed into the sum of a low-rank matrix, a sparse noise matrix and a dense Gaussian noise matrix. Hence, the corresponding probabilistic models are robust to outliers and large sparse noise, and they are mainly composed of Bayesian robust PCA [22], variational Bayesian robust PCA [23] and sparse Bayesian robust PCA [19]. As another type of low-rank matrix factorizations, NMF decomposes a non-negative data matrix into the product of two non-negative low-rank matrices. Compared with PCA, NMF is a technique which learns holistic, not parts-based representations [7]. Different probabilistic models of NMF were proposed in [24–28]. Recently, probabilistic models of low-rank matrix factorizations are also extended to the case of tensor decompositions. Tucker decomposition and CANDECOMP/PARAFAC (CP) decomposition are two most important tensor decompositions. By generalizing the subspace approximation, some new low rank tensor decompositions have emerged, such as the hierarchical Tucker (HT) format [29,30], the tensor tree structure [31] and the tensor train (TT) format [32]. Among them, the TT format is a special case of the HT and the tensor tree structure [33]. The probabilistic models of the Tucker were presented in [34–36] and that of the CP were developed in [37–48].

For probabilistic models of low-rank matrix factorizations, there are three most frequently used statistical approaches for inferring the corresponding probability distributions or parameters, i.e., Expectation Maximization (EM) [49–54], Gibbs sampling (or a Gibbs sampler) [54–62] and variational Bayesian (VB) inference [54,62–71]. EM is an iterative algorithm with guaranteed the local convergence for ML estimation and does not require explicit evaluation of the likelihood function [70]. Although it has many advantages over ML, EM tends to be limited in applications because of its serious requirements for the posterior of the hidden variables and can not be used to solve complex Bayesian models [70]. However, VB inference relaxes some limitations of the EM algorithm and ameliorates

its shortcomings. As a means of VB, Gibbs sampling is another method used to infer the probability distributions of all parameters and hidden variables.

This paper provides a survey on probabilistic models of low-rank matrix factorizations. Firstly, we review the significant probability distributions commonly-used in probabilistic low-rank matrix factorizations and introduce the conjugate priors that are essential to Bayesian statistics. Then we present two most important methods for inferring the probability distributions, that is, Gibbs sampling and VB inference. Next, we roughly divide the available low-rank matrix factorization models into five categories: PCA, matrix factorizations, robust PCA, NMF and tensor decompositions. For each category, we provide a detailed overview of the corresponding probabilistic models.

A central task for probabilistic low-rank matrix factorizations is to predict the missing or incomplete data. For the sake of concise descriptions, we do not consider the missing entries in all models except the sparse Bayesian model of matrix completion. The remainder of this paper is listed as below. Section 2 introduces the commonly-used probability distributions and the conjugate priors. Section 3 presents two frequently used inferring methods: Gibbs sampling and VB inference. Probabilistic models of PCA and matrix factorizations are reviewed respectively in Sections 4 and 5. Sections 6 and 7 survey probabilistic models of robust PCA and NMF, respectively. Section 8 provides other probabilistic models of low-rank matrix factorizations and probabilistic tensor factorizations. The conclusions and future research directions are drawn in Section 9.

*Notation:* Let  $\mathbb{R}$  be the set of real numbers and  $\mathbb{R}_+$  the set of non-negative real numbers. We denote scalars with italic letters (e.g.,  $x$ ), vectors with bold letters (e.g.,  $\mathbf{x}$ ), matrices with bold capital letters (e.g.,  $\mathbf{X}$ ) and sets with italic capital letters (e.g.,  $X$ ). Given a matrix  $\mathbf{X}$ , its  $i$ th row,  $j$ th column and  $(i, j)$ th element are expressed as  $\mathbf{x}_i$ ,  $\mathbf{x}_j$  and  $X_{ij}$ , respectively. If  $\mathbf{X}$  is square, let  $\text{Tr}(\mathbf{X})$  and  $|\mathbf{X}|$  be the trace and the determinant of  $\mathbf{X}$ , respectively.

## 2. Probability Distributions and Conjugate Priors

This section will introduce some probability distributions commonly adopted in probabilistic models of low-rank matrix factorizations and discuss the conjugate priors for algebraic and intuitive convenience in Bayesian statistics.

### 2.1. Probability Distributions

We consider some significant probability distributions that will be needed in the following sections. Given a univariate random variable  $x$  or a random vector  $\mathbf{x}$ , we denote its probability density/mass function by  $p(x)$  or  $p(\mathbf{x})$ . Let  $\mathbb{E}(\mathbf{x})$  be the expectation of  $\mathbf{x}$  and  $\text{Cov}(\mathbf{x}, \mathbf{x})$  the covariance matrix of  $\mathbf{x}$ .

Several probability distributions such as Gamma distribution and Beta distribution deal with the gamma function defined by

$$\Gamma(x) = \int_0^{+\infty} u^{x-1} \exp(-u) du. \quad (1)$$

We summarize the probability distributions used frequently in probabilistic models of low-rank matrix factorizations, as shown in Table 1. In this table, we list the notation for each probability distribution, the probability density/mass function, the expectation and the variance/covariance respectively. For the Wishart distribution  $\mathcal{W}(\mathbf{\Lambda}|\mathbf{W}, v)$ , the term  $B(\mathbf{W}, v)$  is given by

$$|\mathbf{W}|^{-v/2} \left( 2^{vd/2} \pi^{d(d-1)/4} \prod_{i=1}^d \Gamma((v+1-i)/2) \right)^{-1} \quad (2)$$

where the positive integer  $v \geq d - 1$  is named the degrees of freedom. For the generalized inverse Gaussian distribution  $\text{GIG}(x|p, a, b)$ ,  $K_p(c)$  is a modified Bessel function of the second kind. For brevity of notation, we stipulate a few simple representations of a random variable or vector. For instance, if  $\mathbf{x}$  follows a multivariate Gaussian distribution with mean  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$ , we can write this distribution as  $\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , or  $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , or  $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . In addition, some probability

distributions can be extended to the multivariate case under identical conditions. The following cites an example: if random variables  $x_i \sim \text{St}(\mu_i, \lambda_i, v_i)$  and  $x_1, x_2, \dots, x_N$  are independent, then we have  $\mathbf{x} = (x_1, x_2, \dots, x_N)^T \sim \text{St}(\boldsymbol{\mu}, \boldsymbol{\lambda}, \mathbf{v})$  with the probability density function  $p(\mathbf{x}) = \prod_{i=1}^N \text{St}(x_i | \mu_i, \lambda_i, v_i)$ , where  $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_N)^T$ ,  $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_N)^T$  and  $\mathbf{v} = (v_1, v_2, \dots, v_N)^T$ .

There exist close relationships among probability distributions listed in Table 1, as shown partially in Figure 1. Refer to [54,62,66] for more probability distributions. Moreover, some continuous random variables can be reformulated as Gaussian scale mixtures with other distributions from a hierarchical viewpoint. Now, we give two most frequently used examples as below.

**Example 1.** For given parameter  $\mu$ , if the conditional probability distribution of  $x$  is  $p(x|\lambda) = \mathcal{N}(x|\mu, \lambda^{-1})$  and the prior distribution of  $\lambda$  is  $p(\lambda) = \text{IGam}(\lambda|1, 1)$ , then we have  $p(x) = \text{Lap}(x|\mu, \sqrt{2}/2)$ .

The probability density function of  $x$  is derived as:

$$\begin{aligned} p(x) &= \int_0^{+\infty} p(\lambda)p(x|\lambda)d\lambda \\ &= \int_0^{+\infty} \frac{1}{\sqrt{2\pi}}\lambda^{-3/2} \exp\left(-\frac{1}{\lambda}\right) \exp\left(-\frac{1}{2}\lambda(x-\mu)^2\right)d\lambda \\ &\stackrel{\lambda=2t}{=} \int_0^{+\infty} \frac{1}{2\sqrt{\pi}}t^{-3/2} \exp\left(-\frac{1}{2t}\right) \exp\left(-t(x-\mu)^2\right)dt \\ &= \mathcal{L}\left\{\frac{1}{2\sqrt{\pi}}t^{-3/2} \exp\left(-\frac{1}{2t}\right)\right\}(x-\mu)^2. \end{aligned} \quad (3)$$

Meanwhile, it holds that

$$\mathcal{L}^{-1}\left\{\frac{\sqrt{2}}{2} \exp\left(-\sqrt{2(x-\mu)^2}\right)\right\}(t) = \frac{1}{2\sqrt{\pi}}t^{-3/2} \exp\left(-\frac{1}{2t}\right) \quad (4)$$

where  $\mathcal{L}\{\cdot\}$  is the Laplace transform and  $\mathcal{L}^{-1}\{\cdot\}$  is the inverse Laplace transform. Hence, we get  $x \sim \text{Lap}(\mu, \sqrt{2}/2)$ .

**Example 2.** For given parameters  $\mu$  and  $\tau$ , if the conditional probability distribution of  $x$  is  $p(x|u) = \mathcal{N}(x|\mu, (\tau u)^{-1})$  and the prior distribution of  $u$  is  $p(u|v) = \text{Gam}(u|v/2, v/2)$ , then it holds that  $p(x|v) = \text{St}(x|\mu, \tau, v)$ , where  $v$  is a fixed parameter.

The derivation process for  $p(x|v)$  is described as follows:

$$\begin{aligned} p(x|v) &= \int_0^{+\infty} p(x, u|v)du = \int_0^{+\infty} p(u|v)p(x|u)du \\ &\propto \int_0^{+\infty} u^{(v+1)/2-1} \exp\left(-\frac{u(v+(x-\mu)^2\tau)}{2}\right)du \\ &\propto \frac{1}{(v+(x-\mu)^2\tau)^{(v+1)/2}}. \end{aligned} \quad (5)$$

So,  $p(x|v) = \text{St}(x|\mu, \tau, v)$ .

**Table 1.** Commonly-used probability distributions.

Probability Distribution	Notation	Probability Density/Mass Function	Expectation	Variance/Covariance
Bernoulli distribution	$\text{Bern}(x \mu)$	$\mu^x(1-\mu)^{1-x}, x \in \{0, 1\}$	$\mu$	$\mu(1-\mu)$
Poisson distribution	$\text{Poiss}(x \lambda)$	$\frac{\lambda^x}{x!} \exp(-\lambda), x = 0, 1, \dots$	$\lambda$	$\lambda$
Uniform distribution	$\text{U}(x a, b)$	$\frac{1}{b-a}, x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Multivariate Gaussian distribution	$\mathcal{N}(\mathbf{x} \boldsymbol{\mu}, \boldsymbol{\Sigma})$	$\frac{1}{(2\pi)^{d/2} \boldsymbol{\Sigma} ^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \boldsymbol{\Sigma}$ is a $d \times d$ symmetric, positive definite matrix	$\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$
Exponential distribution	$\text{Exp}(x \lambda)$	$\lambda \exp(-\lambda x), x > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Laplace distribution	$\text{Lap}(x \mu, \sigma)$	$\frac{1}{2\sigma} \exp\left(-\frac{ x-\mu }{\sigma}\right)$	$\mu$	$\sigma^2$
Gamma distribution	$\text{Gam}(x a, b)$	$\frac{1}{\Gamma(a)} b^a x^{a-1} \exp(-bx), x > 0$	$\frac{a}{b}$	$\frac{a}{b^2}$
Inverse-Gamma distribution	$\text{IGam}(x a, b)$	$\frac{1}{\Gamma(a)} b^a x^{-a-1} \exp\left(-\frac{b}{x}\right), x > 0$	$\frac{b}{a-1}$ for $a > 1$	$\frac{b^2}{(a-1)^2(a-2)}$ for $a > 2$
Student's $t$ -distribution	$\text{St}(x \mu, \lambda, \nu)$	$\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \left(\frac{\lambda}{\pi\nu}\right)^{1/2} \left(1 + \frac{\lambda(x-\mu)^2}{\nu}\right)^{-(\nu+1)/2}$	$\mu$	$\frac{\nu}{\lambda(\nu-2)}$ for $\nu > 2$
Beta distribution	$\text{Beta}(x a, b)$	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, x \in [0, 1]$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$
Wishart distribution	$\mathcal{W}(\boldsymbol{\Lambda} \mathbf{W}, \nu)$	$B(\mathbf{W}, \nu) \boldsymbol{\Lambda} ^{(\nu-d-1)/2} \exp\left(-\frac{1}{2}\text{Tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right), \boldsymbol{\Lambda}$ is a $d \times d$ symmetric, positive definite matrix	$\nu\mathbf{W}$	$\nu(W_{ij}^2 + W_{ii}W_{jj})$ for $\Lambda_{ij}$
Inverse Gaussian distribution	$\text{IG}(x \mu, \lambda)$	$\sqrt{\frac{\lambda}{2\pi x^3}} \exp\left(-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right), x > 0$	$\mu$	$\frac{\mu^3}{\lambda}$
Generalized inverse Gaussian distribution	$\text{GIG}(x p, a, b)$	$\frac{(a/b)^{p/2} x^{p-1}}{2K_p(\sqrt{ab})} \exp\left(-\frac{1}{2}\left(ax + \frac{b}{x}\right)\right), x > 0$	$\frac{\sqrt{b}K_{p+1}(c)}{\sqrt{a}K_p(c)}, c = \sqrt{ab}$	$\frac{bK_{p+2}(c)}{aK_p(c)} - \frac{bK_{p+1}^2(c)}{aK_p^2(c)}$

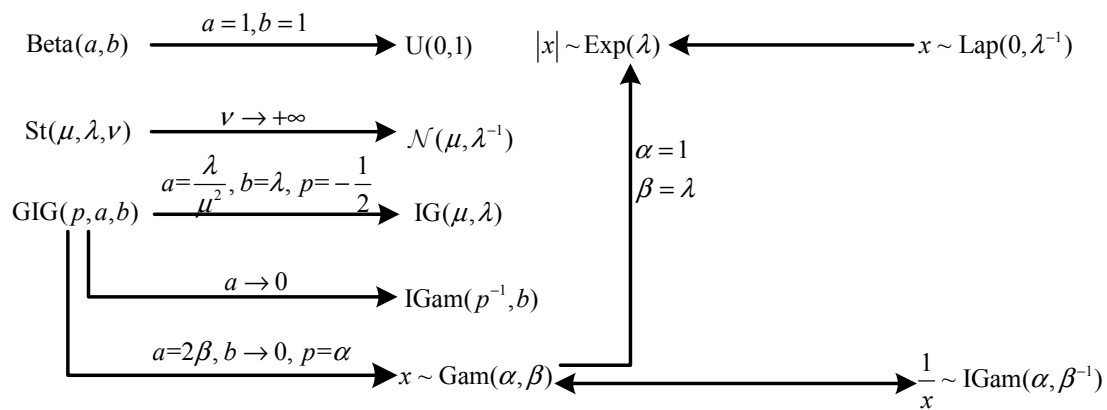


Figure 1. Relationships among several probability distributions.

### 2.2. Conjugate Priors

Let  $\mathbf{x}$  be a random vector with the parameter vector  $\mathbf{z}$  and  $X = \{x_1, x_2, \dots, x_N\}$  a collection of  $N$  observed samples. In the presence of latent variables, they are also absorbed into  $\mathbf{z}$ . For given  $\mathbf{z}$ , the conditional probability density/mass function of  $\mathbf{x}$  is denoted by  $p(\mathbf{x}|\mathbf{z})$ . Thus, we can construct the likelihood function:

$$L(\mathbf{z}|X) = p(X|\mathbf{z}) = \prod_{i=1}^N p(x_i|\mathbf{z}). \tag{6}$$

As for variational Bayesian methods, the parameter vector  $\mathbf{z}$  is usually assumed to be stochastic. Here, the prior distribution of  $\mathbf{z}$  is expressed as  $p(\mathbf{z})$ .

To simplify Bayesian analysis, we hope that the posterior distribution  $p(\mathbf{z}|X)$  is in the same functional form as the prior  $p(\mathbf{z})$ . Under this circumstance, the prior and the posterior are called conjugate distributions and the prior is also called a conjugate prior for the likelihood function  $L(\mathbf{z}|X)$  [54,66]. In the following, we provide three most commonly-used examples of conjugate priors.

**Example 3.** Assume that random variable  $x$  obeys the Bernoulli distribution with parameter  $\mu$ . We have the likelihood function for  $x$ :

$$L(\mu|X) = \prod_{i=1}^N \text{Bern}(x_i|\mu) = \prod_{i=1}^N \mu^{x_i} (1 - \mu)^{1-x_i} = \mu^{\sum_{i=1}^N x_i} (1 - \mu)^{N - \sum_{i=1}^N x_i} \tag{7}$$

where the observations  $x_i \in \{0, 1\}$ . In consideration of the form of  $L(\mu|X)$ , we stipulate the prior distribution of  $\mu$  as the Beta distribution with parameters  $a$  and  $b$ :

$$p(\mu) = \text{Beta}(\mu|a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1 - \mu)^{b-1}. \tag{8}$$

At this moment, we get the posterior distribution of  $\mu$  via the Bayes' rule:

$$p(\mu|X) = \frac{p(\mu)p(X|\mu)}{p(X)} \propto p(\mu)p(X|\mu). \tag{9}$$

Because

$$\begin{aligned} p(\mu)p(X|\mu) &= p(\mu)L(\mu|X) \\ &\propto \mu^{a-1} (1 - \mu)^{b-1} \mu^{\sum_{i=1}^N x_i} (1 - \mu)^{N - \sum_{i=1}^N x_i} \\ &\propto \mu^{a + \sum_{i=1}^N x_i - 1} (1 - \mu)^{b + N - \sum_{i=1}^N x_i - 1} \end{aligned} \tag{10}$$

we have  $p(\mu|X) \sim \text{Beta}(a + \sum_{i=1}^N x_i, b + N - \sum_{i=1}^N x_i)$ . The conclusion means that the Beta distribution is the conjugate prior for the Bernoulli likelihood.

**Example 4.** Assume that random variable  $x$  obeys a univariate Gaussian distribution  $\mathcal{N}(\mu, \lambda^{-1})$ , where  $\lambda$  is also named the precision. The likelihood function of  $x$  for given  $\mu$  is

$$\begin{aligned} L(\lambda|\mu, X) &= \prod_{i=1}^N \mathcal{N}(x_i|\mu, \lambda^{-1}) \\ &= \prod_{i=1}^N \frac{\sqrt{\lambda}}{\sqrt{2\pi}} \exp\left(-\frac{\lambda}{2}(x_i - \mu)^2\right) \\ &= \lambda^{N/2} (2\pi)^{-N/2} \exp\left(-\frac{\lambda}{2} \sum_{i=1}^N (x_i - \mu)^2\right) \end{aligned} \tag{11}$$

We further suppose that the prior distribution of  $\lambda$  is the Gamma distribution with parameters  $a$  and  $b$ . Let  $p(\lambda|\mu, X)$  be the posterior distribution of  $\lambda$ . Then we have

$$p(\lambda|\mu, X) \propto p(\lambda)p(X, \mu|\lambda). \tag{12}$$

Because

$$\begin{aligned} &p(\lambda)p(X, \mu|\lambda) \\ &= \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \lambda^{N/2} (2\pi)^{-N/2} \exp\left(-\frac{\lambda}{2} \sum_{i=1}^N (x_i - \mu)^2\right) \\ &\propto \lambda^{a+N/2-1} \exp\left(-\lambda\left(b + \sum_{i=1}^N (x_i - \mu)^2/2\right)\right) \end{aligned} \tag{13}$$

we get  $p(\lambda|\mu, X) \sim \text{Gam}(a + N/2, b + \sum_{i=1}^N (x_i - \mu)^2/2)$ . Therefore, the conjugate prior of the precision for a Gaussian likelihood is a Gamma distribution.

**Example 5.** Assume that random vector  $\mathbf{x}$  obeys a  $d$ -dimensional Gaussian distribution  $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ , where  $\boldsymbol{\Lambda}$ , the inverse of covariance matrix, is called the precision matrix. We consider the case that both  $\boldsymbol{\mu}$  and  $\boldsymbol{\Lambda}$  are unknown. Thus, the likelihood function for  $\mathbf{x}$  is

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathbf{X}) &= \prod_{i=1}^N \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \\ &= \prod_{i=1}^N \frac{|\boldsymbol{\Lambda}|^{1/2}}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Lambda}(\mathbf{x}_i - \boldsymbol{\mu})\right) \\ &= \frac{|\boldsymbol{\Lambda}|^{N/2}}{(2\pi)^{dN/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Lambda}(\mathbf{x}_i - \boldsymbol{\mu})\right). \end{aligned} \tag{14}$$

The prior distribution of  $(\boldsymbol{\mu}, \boldsymbol{\Lambda})$  is given by a Gaussian–Wishart distribution  $\mathcal{NW}(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{\mu}_0, \boldsymbol{\beta}, \mathbf{W}, v)$  with the joint probability density function:

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{\mu}_0, \boldsymbol{\beta}, \mathbf{W}, v) = \mathcal{N}\left(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (\boldsymbol{\beta}\boldsymbol{\Lambda})^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}, v) \tag{15}$$

where  $\boldsymbol{\mu}_0, \boldsymbol{\beta}, \mathbf{W}$  and  $v$  are the fixed hyperparameters. Hence, it holds that

$$\begin{aligned} &p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{\mu}_0, \boldsymbol{\beta}, \mathbf{W}, v)p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Lambda}) \\ &= \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (\boldsymbol{\beta}\boldsymbol{\Lambda})^{-1}) \mathcal{W}(\boldsymbol{\Lambda}|\mathbf{W}, v) L(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathbf{X}) \\ &\propto |\boldsymbol{\Lambda}|^{(N+v-d)/2} \exp\left(-\frac{1}{2}\left((\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T (\boldsymbol{\beta}\boldsymbol{\Lambda})(\boldsymbol{\mu} - \boldsymbol{\mu}_0) + \sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Lambda}(\mathbf{x}_i - \boldsymbol{\mu}) + \text{Tr}(\mathbf{W}^{-1}\boldsymbol{\Lambda})\right)\right). \end{aligned} \tag{16}$$

Denote  $\bar{\mathbf{x}} = \sum_{i=1}^N \mathbf{x}_i / N$  and  $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$ . To obtain the above probability density function, we first derive the following formula:

$$\begin{aligned} & (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T (\beta \boldsymbol{\Lambda}) (\boldsymbol{\mu} - \boldsymbol{\mu}_0) + \sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Lambda} (\mathbf{x}_i - \boldsymbol{\mu}) \\ &= \boldsymbol{\mu}^T (\beta \boldsymbol{\Lambda}) \boldsymbol{\mu} - 2\boldsymbol{\mu}_0^T (\beta \boldsymbol{\Lambda}) \boldsymbol{\mu} + \boldsymbol{\mu}_0^T (\beta \boldsymbol{\Lambda}) \boldsymbol{\mu}_0 + \sum_{i=1}^N (\boldsymbol{\mu}^T \boldsymbol{\Lambda} \boldsymbol{\mu} - 2\mathbf{x}_i^T \boldsymbol{\Lambda} \boldsymbol{\mu} + \mathbf{x}_i^T \boldsymbol{\Lambda} \mathbf{x}_i) \\ &= (\beta + N) \boldsymbol{\mu}^T \boldsymbol{\Lambda} \boldsymbol{\mu} - 2(\boldsymbol{\Lambda}(\beta \boldsymbol{\mu}_0 + N\bar{\mathbf{x}}))^T \boldsymbol{\mu} + \text{Tr}((\beta \boldsymbol{\mu}_0 \boldsymbol{\mu}_0^T + \mathbf{X}\mathbf{X}^T) \boldsymbol{\Lambda}) \\ &= (\boldsymbol{\mu} - (\beta \boldsymbol{\mu}_0 + N\bar{\mathbf{x}}) / (\beta + N))^T ((\beta + N) \boldsymbol{\Lambda}) (\boldsymbol{\mu} - (\beta \boldsymbol{\mu}_0 + N\bar{\mathbf{x}}) / (\beta + N)) \\ & \quad + \text{Tr}((\beta \boldsymbol{\mu}_0 \boldsymbol{\mu}_0^T + \mathbf{X}\mathbf{X}^T - (\beta \boldsymbol{\mu}_0 + N\bar{\mathbf{x}})(\beta \boldsymbol{\mu}_0 + N\bar{\mathbf{x}})^T / (\beta + N)) \boldsymbol{\Lambda}). \end{aligned} \quad (17)$$

Then we get

$$\begin{aligned} & p(\boldsymbol{\mu}, \boldsymbol{\Lambda} | \boldsymbol{\mu}_0, \beta, \mathbf{W}, v) p(X | \boldsymbol{\mu}, \boldsymbol{\Lambda}) \\ & \propto |\boldsymbol{\Lambda}|^{1/2} \exp\left(-\frac{1}{2}(\boldsymbol{\mu} - (\beta \boldsymbol{\mu}_0 + N\bar{\mathbf{x}}) / (\beta + N))^T ((\beta + N) \boldsymbol{\Lambda}) (\boldsymbol{\mu} - (\beta \boldsymbol{\mu}_0 + N\bar{\mathbf{x}}) / (\beta + N))\right) \\ & \quad |\boldsymbol{\Lambda}|^{(N+v-d-1)/2} \exp\left(-\frac{1}{2} \text{Tr}\left(\left(\mathbf{W}^{-1} + \beta \boldsymbol{\mu}_0 \boldsymbol{\mu}_0^T + \mathbf{X}\mathbf{X}^T - \frac{(\beta \boldsymbol{\mu}_0 + N\bar{\mathbf{x}})(\beta \boldsymbol{\mu}_0 + N\bar{\mathbf{x}})^T}{(\beta + N)}\right) \boldsymbol{\Lambda}\right)\right). \end{aligned} \quad (18)$$

Because  $p(\boldsymbol{\mu}, \boldsymbol{\Lambda} | X) \propto p(\boldsymbol{\mu}, \boldsymbol{\Lambda} | \boldsymbol{\mu}_0, \beta, \mathbf{W}, v) p(X | \boldsymbol{\mu}, \boldsymbol{\Lambda})$ , we have

$$\begin{aligned} & p(\boldsymbol{\mu}, \boldsymbol{\Lambda} | X) \sim \\ & \mathcal{NW}\left(\frac{\beta \boldsymbol{\mu}_0 + N\bar{\mathbf{x}}}{\beta + N}, \beta + N, \left(\mathbf{W}^{-1} + \beta \boldsymbol{\mu}_0 \boldsymbol{\mu}_0^T + \mathbf{X}\mathbf{X}^T - \frac{(\beta \boldsymbol{\mu}_0 + N\bar{\mathbf{x}})(\beta \boldsymbol{\mu}_0 + N\bar{\mathbf{x}})^T}{(\beta + N)}\right)^{-1}, N + v\right). \end{aligned} \quad (19)$$

This example shows that the conjugate prior of  $(\boldsymbol{\mu}, \boldsymbol{\Lambda})$  for a multivariate Gaussian  $\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Lambda})$  is a Gaussian–Wishart distribution.

There are some other conclusions on the conjugate priors. For instance, given Gaussian likelihood, the conjugate prior for the mean is another Gaussian distribution, the conjugate prior for the precision matrix is a Wishart distribution. All probability distributions discussed in Section 2.1 belong to a broad class of distributions, that is, the exponential family. It is shown that the exponential family distributions have conjugate priors [54,66,72].

### 3. Gibbs Sampling and Variational Bayesian Inference

Due to the existence of the latent variables or unknown parameters, computing the posterior distribution is analytically intractable in general. In this section, we will provide two approximation inference methods: Gibbs sampling and variational Bayesian inference.

#### 3.1. Gibbs Sampling

As a powerful framework for sampling from a probability distribution, Markov chain Monte Carlo (MCMC) methods, also called Markov chain simulation, construct a Markov chain such that its equilibrium distribution is the desired distribution [73–80]. Random walk Monte Carlo methods are a large subclass of MCMC and they include mainly Metropolis–Hastings [54,62,66,81], Gibbs sampling, Slice sampling [54,62,66], Multiple-try Metropolis [82,83].

Gibbs sampling or Gibbs sampler, a simple MCMC algorithm, is especially applicable for approximating a sequence of observations by a specified probability distribution when direct sampling is intractable. In addition, the basic version of Gibbs sampling is also a special case of the Metropolis–Hastings algorithm [54,62]. In detailed implementation, Gibbs sampling adopts the strategy of sampling from a conditional probability distribution instead of marginalizing the joint probability distribution by integrating over other variables. In other words, Gibbs sampling generates alternatively an instance from its corresponding conditional probability distribution by fixing other variables.



Let  $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_M$  be  $M$  blocks of random variables and set  $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_M)$ . The joint probability distribution of  $\mathbf{z}$  is written as  $p(\mathbf{z}) = p(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_M)$ . In this subsection, we only consider the case that it is difficult to directly sampling from the joint probability distribution  $p(\mathbf{z})$ . To this end, we can sample each component  $\mathbf{z}_i$  from marginal distribution in order. The marginal distribution of  $\mathbf{z}_i$  can be theoretically obtained by the following formulation:

$$p(\mathbf{z}_i) = \int \int \cdots \int p(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_M) d\mathbf{z}_1 \cdots d\mathbf{z}_{i-1} d\mathbf{z}_{i+1} \cdots d\mathbf{z}_M. \quad (20)$$

Generally speaking, the integrating term in the above formula is not tractable. A wise sampling strategy is that we generate  $\mathbf{z}_i$  according to the conditional probability distribution  $p(\mathbf{z}_i | \mathbf{z} \setminus \mathbf{z}_i)$ , where  $\mathbf{z} \setminus \mathbf{z}_i = (\mathbf{z}_1, \dots, \mathbf{z}_{i-1}, \mathbf{z}_{i+1}, \dots, \mathbf{z}_M)$ . By using Bayes' rule, the relationship between the conditional probability distribution and the joint probability distribution is given as follows:

$$p(\mathbf{z}_i | \mathbf{z} \setminus \mathbf{z}_i) = \frac{p(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_M)}{p(\mathbf{z}_1, \dots, \mathbf{z}_{i-1}, \mathbf{z}_{i+1}, \dots, \mathbf{z}_M)} \propto p(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_M). \quad (21)$$

The sampling procedure is repeated by cycling through all variables. We summarize the outline of Gibbs sampling as below.

1. Initialize  $\mathbf{z}_i = \mathbf{z}_i^{(0)}, i = 2, \dots, M$ .
2. For  $k = 1, 2, \dots, T$ 
  - For  $i = 1, 2, \dots, M$ 
    - Generate  $\mathbf{z}_i^{(k)}$  from the conditional probability distribution  $p(\mathbf{z}_i | \mathbf{z}_1^{(k)}, \dots, \mathbf{z}_{i-1}^{(k)}, \mathbf{z}_{i+1}^{(k-1)}, \dots, \mathbf{z}_M^{(k-1)})$ .
    - End
  - End

In the above sampling procedure,  $T$  is the maximum number of iterations. Therefore, we have  $T$  samples  $\mathbf{z}^{(k)} = (\mathbf{z}_1^{(k)}, \mathbf{z}_2^{(k)}, \dots, \mathbf{z}_M^{(k)}), k = 1, 2, \dots, T$ . To alleviate the influence of random initializations, we can ignore the samples within the burn-in period. Although Gibbs sampling is commonly efficient in obtaining marginal distributions from conditional distributions, it can be highly inefficient for some cases such as sampling from Mexican hat like distribution [53].

The Metropolis algorithm is an instance of MCMC algorithms and walks randomly with an acceptance/rejection rule until the convergence is achieved. As the generalization of Metropolis algorithm, the Metropolis–Hastings algorithm modifies the jumping rules and its convergence speed is improved [66]. Broadly speaking, the advanced Gibbs sampling algorithm can be regarded as a special case of the Metropolis–Hastings algorithm. The Gibbs sampler is applicable for models that are conditionally conjugate, while the Metropolis algorithm can be used for not conditionally conjugate models. Hence, we can combine both the Gibbs sampler and the Metropolis algorithm to sample from complex distributions [66].

### 3.2. Variational Bayesian Inference

Another widely used class of approximating the marginal probability distribution  $p(\mathbf{z}_i)$  is the variational Bayesian (VB) inference [54]. We still use the notation  $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$  to represent a matrix composed of  $N$  observed datum and  $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_M)$  to represent a vector constructed by  $M$  blocks of latent variables. In VB methods, both  $\mathbf{X}$  and  $\mathbf{z}$  are assumed to be stochastic.

Given the prior distribution  $p(\mathbf{z})$  and the data likelihood  $p(\mathbf{X} | \mathbf{z})$ , the probability distribution of data matrix  $\mathbf{X}$ , also called the evidence, can be calculated as  $p(\mathbf{X}) = \int p(\mathbf{z}) p(\mathbf{X} | \mathbf{z}) d\mathbf{z}$ . Then we derive the conditional probability distribution of  $\mathbf{z}$  via Bayes' rule:

$$p(\mathbf{z} | \mathbf{X}) = \frac{p(\mathbf{z}) p(\mathbf{X} | \mathbf{z})}{\int p(\mathbf{z}) p(\mathbf{X} | \mathbf{z}) d\mathbf{z}}. \quad (22)$$

However, the integrating term  $\int p(\mathbf{z})p(\mathbf{X}|\mathbf{z})d\mathbf{z}$  is analytically intractable under normal conditions. Now, we harness VB methods to approximate the posterior distribution  $p(\mathbf{z}|\mathbf{X})$  as well as  $p(\mathbf{X})$ .

Let  $q(\mathbf{z})$  be the trial probability distribution of  $p(\mathbf{z}|\mathbf{X})$ . The log of the probability distribution  $p(\mathbf{X})$  is decomposed as below:

$$\begin{aligned}\ln p(\mathbf{X}) &= \int q(\mathbf{z}) \ln p(\mathbf{X})d\mathbf{z} = \int q(\mathbf{z}) \ln \frac{p(\mathbf{X},\mathbf{z})}{p(\mathbf{z}|\mathbf{X})}d\mathbf{z} \\ &= \int q(\mathbf{z}) \ln p(\mathbf{X},\mathbf{z})d\mathbf{z} - \int q(\mathbf{z}) \ln p(\mathbf{z}|\mathbf{X})d\mathbf{z} \\ &= \int q(\mathbf{z}) \ln p(\mathbf{X},\mathbf{z})d\mathbf{z} - \left( \int q(\mathbf{z}) \ln \frac{p(\mathbf{z}|\mathbf{X})}{q(\mathbf{z})}d\mathbf{z} + \int q(\mathbf{z}) \ln q(\mathbf{z})d\mathbf{z} \right) \\ &= L(q) + KL(q||p) \geq L(q)\end{aligned}\tag{23}$$

where  $L(q) = \int q(\mathbf{z}) \ln \frac{p(\mathbf{X},\mathbf{z})}{q(\mathbf{z})}d\mathbf{z}$  and  $KL(q||p) = -\int q(\mathbf{z}) \ln \frac{p(\mathbf{z}|\mathbf{X})}{q(\mathbf{z})}d\mathbf{z}$ . The term  $KL(q||p)$  is called the Kullback–Leibler (KL) divergence [54,84–87] between  $q(\mathbf{z})$  and  $p(\mathbf{z}|\mathbf{X})$ , and  $L(q)$  is the lower bound of  $\ln p(\mathbf{X})$  that achieves its lower bound if and only if  $KL(q||p) = 0$  (or equivalently  $q = p$ ). The above divergence can also be explained as the information gain changing from the prior  $q(\mathbf{z})$  to the posterior  $p(\mathbf{z}|\mathbf{X})$  [85].

Another perspective for Equation (23) is that,

$$\ln p(\mathbf{X}) = \ln \int p(\mathbf{X},\mathbf{z})d\mathbf{z} = \ln \int q(\mathbf{z}) \frac{p(\mathbf{X},\mathbf{z})}{q(\mathbf{z})}d\mathbf{z} \geq \int q(\mathbf{z}) \ln \frac{p(\mathbf{X},\mathbf{z})}{q(\mathbf{z})}d\mathbf{z} = L(q).\tag{24}$$

The above inequality is obtained by Jensen’s inequality. The negative lower bound  $-L(q)$  is called the free energy [88,89].

The KL divergence  $KL(q||p)$  can be regarded as a metric for evaluating the approximation performance of the prior distribution  $q(\mathbf{z})$  over the posterior distribution  $p(\mathbf{z}|\mathbf{X})$  [54]. In the light of Equation (23), minimizing  $KL(q||p)$  is equivalent to maximizing  $L(q)$ . What’s more, the lower bound can be further derived as below:

$$\begin{aligned}L(q) &= \int q(\mathbf{z}) \ln p(\mathbf{X}|\mathbf{z})d\mathbf{z} + \int q(\mathbf{z}) \ln p(\mathbf{z})d\mathbf{z} - \int q(\mathbf{z}) \ln q(\mathbf{z})d\mathbf{z} \\ &= \int q(\mathbf{z}) \ln p(\mathbf{X}|\mathbf{z})d\mathbf{z} + \int q(\mathbf{z}) \ln \frac{p(\mathbf{z})}{q(\mathbf{z})}d\mathbf{z} \\ &= \mathbb{E}_{q(\mathbf{z})} \ln p(\mathbf{X}|\mathbf{z}) - KL(q||p).\end{aligned}\tag{25}$$

The above Equation means that  $\ln p(\mathbf{X})$  can also be regarded as the expectation of the log likelihood function  $\ln p(\mathbf{X}|\mathbf{z})$  with respect to  $\mathbf{z}$ .

To reduce the difficulty of approximating the trial distribution  $q(\mathbf{z})$ , we partition all latent variables into  $M$  disjoint block variables  $\{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_M\}$  and assume that  $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_M$  are mutually independent. Based on the above assumption, we have

$$q(\mathbf{z}) = q(\mathbf{z}_1)q(\mathbf{z}_2) \cdots q(\mathbf{z}_M).\tag{26}$$

Afterwards, we utilize the mean field theory to obtain the approximation of  $q(\mathbf{z})$ . Concretely speaking, we update each  $q(\mathbf{z}_i)$  in turn. Let  $q(\mathbf{z}_i)$  be unknown and  $q(\mathbf{z}_j)$  ( $j \neq i$ ) be fixed. Under this circumstance, we can get the optimal  $q(\mathbf{z}_i)$  by solving the following variational maximization problem:

$$q^*(\mathbf{z}_i) = \operatorname{argmax}_{q(\mathbf{z}_i)} L(q(\mathbf{z}))\tag{27}$$

Plugging Equation (26) into the lower bound of  $\ln p(\mathbf{X})$ , we have

$$\begin{aligned}
 L(q) &= \int \prod_{j=1}^M q(\mathbf{z}_j) \ln p(\mathbf{X}, \mathbf{z}) \prod_{j=1}^M d\mathbf{z}_j - \int \prod_{j=1}^M q(\mathbf{z}_j) \sum_{j=1}^M \ln q(\mathbf{z}_j) \prod_{j=1}^M d\mathbf{z}_j \\
 &= \int q(\mathbf{z}_i) \left( \int \ln p(\mathbf{X}, \mathbf{z}) \prod_{j \neq i} q(\mathbf{z}_j) \prod_{j \neq i} d\mathbf{z}_j \right) d\mathbf{z}_i - \int q(\mathbf{z}_i) \ln q(\mathbf{z}_i) d\mathbf{z}_i + \text{const} \\
 &= \int q(\mathbf{z}_i) \ln \tilde{p}(\mathbf{X}, \mathbf{z}_i) d\mathbf{z}_i - \int q(\mathbf{z}_i) \ln q(\mathbf{z}_i) d\mathbf{z}_i + \text{const} \\
 &= \int q(\mathbf{z}_i) \ln \frac{\tilde{p}(\mathbf{X}, \mathbf{z}_i)}{q(\mathbf{z}_i)} d\mathbf{z}_i + \text{const}
 \end{aligned} \tag{28}$$

where “const” is a term independent of  $\mathbf{z}_i$  and  $\tilde{p}(\mathbf{X}, \mathbf{z}_i)$  is a new probability distribution whose log is defined by

$$\begin{aligned}
 \ln \tilde{p}(\mathbf{X}, \mathbf{z}_i) &= \int \ln p(\mathbf{X}, \mathbf{z}) \prod_{j \neq i} q(\mathbf{z}_j) \prod_{j \neq i} d\mathbf{z}_j + \text{const} \\
 &= \mathbb{E}_{\mathbf{z} \setminus \mathbf{z}_i} [p(\mathbf{X}, \mathbf{z})] + \text{const}.
 \end{aligned} \tag{29}$$

Therefore,  $\tilde{p}(\mathbf{X}, \mathbf{z}_i)$  is the optimal solution of problem (27) obtained by minimizing the KL divergence between  $q(\mathbf{z}_i)$  and  $\tilde{p}(\mathbf{X}, \mathbf{z}_i)$ .

An advantage of the VB methods is that they are immune to over fitting. But they also have some shortcomings. For example, the probability distributions derived via VB have always less probability mass in the wings than the true solution and this systematic bias may break applications. A VB method approximates a full posterior distribution by maximizing the corresponding lower bound on the marginal likelihood and it can only handle a smaller dataset. In contrast, the stochastic variational inference optimizes a subset at each iteration. This batch inference algorithm is scalable and outperforms traditional variational inference methods [90,91].

### 3.3. Comparisons between Gibbs Sampling and Variational Bayesian Inference

Evaluating the posterior distribution  $p(\mathbf{z}|\mathbf{X})$  is a central task in probabilistic models of low-rank matrix factorizations. In many practical applications, it is often infeasible to compute the posterior distribution or the expectations with respect to this distribution. Gibbs sampling and VB inference are two dominant methods for approximating the posterior distribution. The former is a stochastic approximation and the latter is a deterministic technique.

The fatal shortcoming of Expectation Maximization (EM) algorithm is that the posterior distribution of the latent variables should be given in advance. Both Gibbs sampling and VB inference can ameliorate the shortcoming of EM algorithm. Gibbs sampling is easy to implement due to the fact it adopts a Monte Carlo procedure. Another advantage of Gibbs sampling is that it can generate exact results [54]. But this method is often suitable for small-scale problems because it costs a large amount of computation. Compared with Gibbs sampling, VB inference does not generate exact results and has small computation complexity. Therefore, their strengths and weaknesses are complementary rather than competitive.

## 4. Probabilistic Models of Principal Component Analysis

Principal Component Analysis (PCA) is a special type of low-rank matrix factorizations. This section first introduces the classical PCA and then reviews its probabilistic models.

### 4.1. Principal Component Analysis

The classical PCA converts a set of samples with possibly correlated variables into another set of samples with linearly uncorrelated variables via an orthogonal transformation [1]. Based on this, PCA is an effective technique widely used in performing dimensionality reduction and extracting features.

Let  $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$  be a collection of  $N$  samples with  $d$  dimensions and  $\mathbf{I}_r$  ( $r < d$ ) an  $r$ -by- $r$  identity matrix. Given a projection transformation matrix  $\mathbf{W} \in \mathbb{R}^{d \times r}$ , the PCA model can be expressed as

$$\mathbf{x}_i = \mathbf{W}\mathbf{z}_i + \bar{\mathbf{x}} + \boldsymbol{\varepsilon}_i, \quad i = 1, 2, \dots, N \quad (30)$$

where the mean vector  $\bar{\mathbf{x}} = \sum_{i=1}^N \mathbf{x}_i / N$ ,  $\mathbf{W}$  satisfies  $\mathbf{W}^T \mathbf{W} = \mathbf{I}_r$ ,  $\mathbf{z}_i$  is a representation coefficient vector with  $r$  dimensions and  $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \dots, \boldsymbol{\varepsilon}_N \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$  are independent and identically distributed noise vectors. Denote  $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$ ,  $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N)$ ,  $\bar{\mathbf{X}} = (\bar{\mathbf{x}}, \bar{\mathbf{x}}, \dots, \bar{\mathbf{x}}) \in \mathbb{R}^{d \times N}$  and  $\mathbf{E} = (\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \dots, \boldsymbol{\varepsilon}_N)$ . Then PCA can be rewritten as the following matrix factorization formulation:

$$\mathbf{X} = \mathbf{W}\mathbf{Z} + \bar{\mathbf{X}} + \mathbf{E}. \quad (31)$$

According to the maximum likelihood estimation, the optimal  $\mathbf{W}$  and  $\mathbf{Z}$  can be obtained by solving the minimization problem:

$$\min_{\mathbf{W}, \mathbf{Z}} \|\mathbf{X} - \bar{\mathbf{X}} - \mathbf{W}\mathbf{Z}\|_F^2, \quad \text{s.t. } \mathbf{W}^T \mathbf{W} = \mathbf{I}_r \quad (32)$$

where  $\|\cdot\|_F$  is the Frobenius norm of a matrix. Although this optimization problem is not convex, we can obtain the optimal transform matrix  $\mathbf{W}^*$  by stacking  $r$  singular vectors corresponding to the first  $r$  largest singular values of the sample covariance matrix  $\mathbf{S} = \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T / (N - 1)$ . Let  $\mathbf{z}_i^*$  be the optimal low-dimensional representation of  $\mathbf{x}_i$ . Then it holds that  $\mathbf{z}_i^* = \mathbf{W}^{*T}(\mathbf{x}_i - \bar{\mathbf{x}})$ .

The PCA technique only supposes that the dataset is contaminated by isotropic Gaussian noise. The advantage of PCA is that it is very simple and effective in achieving the point estimations of  $\mathbf{W}$  and  $\mathbf{Z}$ . But we can not obtain their probability distributions. In fact, the probability distributions of parameters are more useful and valuable than the point estimations in exploiting the intrinsic essence and investigating the procedure of data generation. Probabilistic models of PCA are just a class of methods for inferring the probability distributions of parameters.

#### 4.2. Probabilistic Principal Component Analysis

In Equation (30), the low-dimensional representation  $\mathbf{z}_i$  is an unknown and deterministic parameter vector. In contrast, the original probabilistic PCA [10] regards  $\mathbf{z}_i$  as a stochastic vector. This probability model provides a general form of decomposing a  $d$ -dimensional input sample  $\mathbf{x}$ :

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\varepsilon} \quad (33)$$

where the latent random vector  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_r)$ , the noise  $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$  and  $\boldsymbol{\mu}$  is the mean vector. In the following, a Maximum Likelihood (ML) method is proposed to obtain the point estimations of  $\mathbf{W}$ ,  $\boldsymbol{\mu}$  and  $\sigma^2$ .

It is obvious that  $p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2 \mathbf{I}_d)$ . Hence we have

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \sigma^2 \mathbf{I}_d + \mathbf{W}\mathbf{W}^T). \quad (34)$$

Given the observed sample set  $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ , the log-likelihood function is

$$L(\mathbf{W}, \boldsymbol{\mu}, \sigma^2 | X) = \sum_{i=1}^N \ln p(\mathbf{x}_i) = \sum_{i=1}^N \ln \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}, \sigma^2 \mathbf{I}_d + \mathbf{W}\mathbf{W}^T). \quad (35)$$

The optimal  $\mathbf{W}, \boldsymbol{\mu}, \sigma^2$  can be obtained by maximizing  $L(\mathbf{W}, \boldsymbol{\mu}, \sigma^2 | X)$ . By letting the partial derivative of  $L(\mathbf{W}, \boldsymbol{\mu}, \sigma^2 | X)$  with respect to  $\boldsymbol{\mu}$  be the zero vector, we can easily get the maximum likelihood estimation of  $\boldsymbol{\mu}$ :  $\boldsymbol{\mu}_{\text{ML}} = \bar{\mathbf{x}}$ . Hence, the optimal  $\mathbf{W}$  and  $\sigma^2$  can be achieved by the stationary points of  $L(\mathbf{W}, \boldsymbol{\mu}_{\text{ML}}, \sigma^2 | X)$  with respect to  $\mathbf{W}$  and  $\sigma^2$ .

The aforementioned method is slightly complex when computing  $\mathbf{W}$  and  $\sigma^2$ . For this purpose, an Expectation Maximization (EM) algorithm was also presented to solve probabilistic PCA. If  $\mathbf{W}$ ,  $\boldsymbol{\mu}$  and  $\sigma^2$  are given, then the joint probability distribution of  $\mathbf{z}$  and  $\mathbf{x}$  can be derived as follows:

$$\begin{aligned} p(\mathbf{z}, \mathbf{x} | \mathbf{W}, \boldsymbol{\mu}, \sigma^2) &= p(\mathbf{z} | \mathbf{W}, \boldsymbol{\mu}, \sigma^2) p(\mathbf{x} | \mathbf{z}, \mathbf{W}, \boldsymbol{\mu}, \sigma^2) \\ &= \mathcal{N}(\mathbf{z} | \mathbf{0}, \mathbf{I}_r) \mathcal{N}(\mathbf{x} | \mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2 \mathbf{I}_d). \end{aligned} \tag{36}$$

The posterior distribution of  $\mathbf{z}$  for given  $\mathbf{x}$  is

$$p(\mathbf{z} | \mathbf{x}, \mathbf{W}, \boldsymbol{\mu}, \sigma^2) = \frac{p(\mathbf{z}, \mathbf{x} | \mathbf{W}, \boldsymbol{\mu}, \sigma^2)}{p(\mathbf{x} | \mathbf{W}, \boldsymbol{\mu}, \sigma^2)} \propto \mathcal{N}(\mathbf{z} | \mathbf{0}, \mathbf{I}_r) \mathcal{N}(\mathbf{x} | \mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2 \mathbf{I}_d). \tag{37}$$

Because

$$\begin{aligned} &\mathcal{N}(\mathbf{z} | \mathbf{0}, \mathbf{I}_r) \mathcal{N}(\mathbf{x} | \mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2 \mathbf{I}_d) \\ &\propto \exp\left(-\frac{1}{2} \mathbf{z}^T \mathbf{z}\right) \exp\left(-\frac{1}{2\sigma^2} (\mathbf{x} - \mathbf{W}\mathbf{z} - \boldsymbol{\mu})^T (\mathbf{x} - \mathbf{W}\mathbf{z} - \boldsymbol{\mu})\right) \\ &\propto \exp\left(-\frac{1}{2} \left[\mathbf{z} - (\sigma^2 \mathbf{I} + \mathbf{W}^T \mathbf{W})^{-1} \mathbf{W}^T (\mathbf{x} - \boldsymbol{\mu})\right]^T \frac{\sigma^2 \mathbf{I} + \mathbf{W}^T \mathbf{W}}{\sigma^2} \left[\mathbf{z} - (\sigma^2 \mathbf{I} + \mathbf{W}^T \mathbf{W})^{-1} \mathbf{W}^T (\mathbf{x} - \boldsymbol{\mu})\right]\right) \end{aligned} \tag{38}$$

we have

$$p(\mathbf{z} | \mathbf{x}, \mathbf{W}, \boldsymbol{\mu}, \sigma^2) \sim \mathcal{N}(\langle \mathbf{z} \rangle, \boldsymbol{\Sigma}_z) \tag{39}$$

where  $\boldsymbol{\Sigma}_z = \text{Cov}(\mathbf{z}, \mathbf{z}) = \sigma^2 (\sigma^2 \mathbf{I} + \mathbf{W}^T \mathbf{W})^{-1}$  is the covariance matrix of  $\mathbf{z}$  and  $\langle \mathbf{z} \rangle = \boldsymbol{\Sigma}_z \mathbf{W}^T (\mathbf{x} - \boldsymbol{\mu}) / \sigma^2$  is the mean of  $\mathbf{z}$ . Based on the above analysis, we give the complete-data log-likelihood function:

$$\begin{aligned} L_C(\mathbf{W}, \boldsymbol{\mu}, \sigma^2 | X) &= \sum_{i=1}^N \ln p(\mathbf{x}_i, \mathbf{z}_i | \mathbf{W}, \boldsymbol{\mu}, \sigma^2) \\ &= \sum_{i=1}^N \ln \mathcal{N}(\mathbf{z}_i | \mathbf{0}, \mathbf{I}_r) + \ln \mathcal{N}(\mathbf{x}_i | \mathbf{W}\mathbf{z}_i + \boldsymbol{\mu}, \sigma^2 \mathbf{I}) \\ &= -\frac{dN}{2} \ln \sigma^2 - \frac{1}{2} \sum_{i=1}^N \left( \mathbf{z}_i^T \mathbf{z}_i + (\mathbf{x}_i - \mathbf{W}\mathbf{z}_i - \boldsymbol{\mu})^T (\mathbf{x}_i - \mathbf{W}\mathbf{z}_i - \boldsymbol{\mu}) / \sigma^2 \right) + \text{const} \\ &= -\frac{dN}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^N \left( \mathbf{z}_i^T (\sigma^2 \mathbf{I} + \mathbf{W}^T \mathbf{W}) \mathbf{z}_i + (\mathbf{x}_i - \boldsymbol{\mu})^T (\mathbf{x}_i - \boldsymbol{\mu}) + 2(\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{W}\mathbf{z}_i \right) + \text{const} \end{aligned} \tag{40}$$

where “const” is a term independent of  $\mathbf{W}$ ,  $\boldsymbol{\mu}$  and  $\sigma^2$ .

In the expectation step, we take expectation on  $L_C(\mathbf{W}, \boldsymbol{\mu}, \sigma^2 | X)$  with respect to  $\mathbf{z}$ :

$$\begin{aligned} &\mathbb{E}_z [L_C(\mathbf{W}, \boldsymbol{\mu}, \sigma^2 | X)] \\ &= -\frac{dN}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^N \left( \text{tr}((\sigma^2 \mathbf{I} + \mathbf{W}^T \mathbf{W}) \mathbb{E}[\mathbf{z}_i \mathbf{z}_i^T]) + (\mathbf{x}_i - \boldsymbol{\mu})^T (\mathbf{x}_i - \boldsymbol{\mu}) - 2(\mathbf{x}_i - \boldsymbol{\mu})^T \mathbf{W} \mathbb{E}[\mathbf{z}_i] \right) \\ &\quad + \text{const}. \end{aligned} \tag{41}$$

According to Equation (39), we have  $\mathbb{E}[\mathbf{z}_i] = \boldsymbol{\Sigma}_z \mathbf{W}^T (\mathbf{x}_i - \boldsymbol{\mu}) / \sigma^2$  and  $\mathbb{E}[\mathbf{z}_i \mathbf{z}_i^T] = \boldsymbol{\Sigma}_z + \mathbb{E}[\mathbf{z}_i] \mathbb{E}[\mathbf{z}_i]^T$ .

In the maximization step, we first obtain the maximum likelihood estimation of  $\boldsymbol{\mu}$ :  $\boldsymbol{\mu}_{ML} = \bar{\mathbf{x}}$  by setting the partial derivative of  $\mathbb{E}_z [L_C(\mathbf{W}, \boldsymbol{\mu}, \sigma^2 | X)]$  with respect to  $\boldsymbol{\mu}$  be a zero vector. Similarly, we can also obtain the optimal estimations of  $\mathbf{W}$  and  $\sigma^2$  by solving the equation set:

$$\begin{cases} \frac{\partial}{\partial \mathbf{W}} \mathbb{E}_z [L_C(\mathbf{W}, \boldsymbol{\mu}_{ML}, \sigma^2 | X)] = \mathbf{0}, \\ \frac{\partial}{\partial \sigma^2} \mathbb{E}_z [L_C(\mathbf{W}, \boldsymbol{\mu}_{ML}, \sigma^2 | X)] = 0. \end{cases} \tag{42}$$

Moreover, a mixture model for probabilistic PCA was proposed in [92]. Khan et al. replaced the Gaussian noise with Gaussian process and incorporated the information of preference pairs into the collaborative filtering [93].

#### 4.3. Bayesian Principal Component Analysis

In probabilistic PCA, both  $\mathbf{z}$  and  $\boldsymbol{\varepsilon}$  obey Gaussian distributions, and both  $\mathbf{W}$  and  $\boldsymbol{\mu}$  are non-stochastic parameters. Now, we further treat  $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_r$  and  $\boldsymbol{\mu}$  as independent random variables, that is,  $\mathbf{w}_i \sim \mathcal{N}(\mathbf{0}, \alpha_i^{-1} \mathbf{I}_d)$  and  $\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}, \beta^{-1} \mathbf{I}_d)$ . At this moment, the corresponding probabilistic model of PCA is called Bayesian PCA [11]. Set  $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_r)^T$  and  $\tau = \sigma^{-2}$ . Then we give the prior distributions of  $\boldsymbol{\alpha}$  and  $\tau$  as follows:

$$p(\boldsymbol{\alpha}) = \prod_{i=1}^r \text{Gam}(\alpha_i | a_0, b_0), p(\tau) = \text{Gam}(\tau | c_0, d_0) \tag{43}$$

where  $a_0, b_0, c_0, d_0$  are four given hyperparameters.

The true joint probability distribution of complete data is given by

$$\begin{aligned} p(\mathbf{X}, \mathbf{W}, \boldsymbol{\alpha}, \mathbf{Z}, \boldsymbol{\mu}, \tau) &= p(\mathbf{W}, \boldsymbol{\alpha}, \mathbf{Z}, \boldsymbol{\mu}, \tau) p(\mathbf{X} | \mathbf{W}, \boldsymbol{\alpha}, \mathbf{Z}, \boldsymbol{\mu}, \tau) \\ &= p(\boldsymbol{\alpha}) p(\mathbf{W} | \boldsymbol{\alpha}) p(\mathbf{Z}) p(\boldsymbol{\mu}) p(\tau) p(\mathbf{X} | \mathbf{W}, \mathbf{Z}, \boldsymbol{\mu}). \end{aligned} \tag{44}$$

We suppose that the trial distribution of  $p(\mathbf{X}, \mathbf{W}, \boldsymbol{\alpha}, \mathbf{Z}, \boldsymbol{\mu}, \tau)$  has the following form:

$$q(\mathbf{W}, \boldsymbol{\alpha}, \mathbf{Z}, \boldsymbol{\mu}, \tau) = q(\mathbf{W}) q(\boldsymbol{\alpha}) q(\mathbf{Z}) q(\boldsymbol{\mu}) q(\tau). \tag{45}$$

By making use of VB inference, we can obtain the trial probability distributions of  $\mathbf{W}, \boldsymbol{\alpha}, \mathbf{Z}, \boldsymbol{\mu}$  and  $\tau$  respectively. In detailed implementation, the hyperparameters  $a_0, b_0, c_0, d_0, \beta$  can be set to be small positive numbers to obtain the broad priors. Unlike other approximation methods, the proposed method maximizes a rigorous lower bound.

#### 4.4. Robust L1 Principal Component Analysis

The probabilistic model of robust L1 PCA [12] regards both  $\boldsymbol{\mu}$  and  $\mathbf{W}$  as deterministic parameters and  $\boldsymbol{\mu}$  is set to a zero vector without loss of the generality. We still suppose that  $\mathbf{z}$  obeys a spherical multivariate Gaussian distribution, that is,  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_r)$ . To improve the robustness, the noise  $\boldsymbol{\varepsilon}_i$  is assumed to follow a multivariate Laplace distribution:

$$p(\boldsymbol{\varepsilon}_i | \sigma) = \left(\frac{1}{2\sigma}\right)^{d/2} \exp\left(-\frac{1}{\sigma} \|\boldsymbol{\varepsilon}_i\|_1\right) \tag{46}$$

where  $\|\cdot\|_1$  is the L1 norm of a vector. Due to the fact that the Laplace distribution has heavy tail, the proposed model in [12] is more robust against data outliers.

We use the hierarchical model to deal with the Laplace distribution. The probability density distribution of a univariate Laplace distribution  $\text{Lap}(0, \sigma)$  can be rewritten as

$$p(\boldsymbol{\varepsilon} | \sigma) = \int \sqrt{\frac{\beta}{2\pi\sigma^2}} \exp\left(-\frac{\beta}{2\sigma^2} \boldsymbol{\varepsilon}^2\right) \frac{1}{2} \beta^{-2} \exp\left(-\frac{1}{2\beta}\right) d\beta \tag{47}$$

Hence, we can set  $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, (\sigma^2/\beta_i)\mathbf{I}_d)$  and  $\beta_i \sim \text{IGam}(1, 1/2)$ . Let  $\rho = 1/\sigma^2$  and give its prior distribution:

$$p(\rho) = \text{Gam}(\rho | a, b) \tag{48}$$

where  $a$  and  $b$  are two given hyperparameters.

We take  $\mathbf{Z}$ ,  $\beta$  and  $\rho$  as latent variables and  $\mathbf{W}$  as the hyperparameters. For fixed  $\mathbf{W}$ , the true joint probability distribution of all latent variables is

$$p(\mathbf{Z}, \beta, \rho, \mathbf{X}|\mathbf{W}) = p(\rho)p(\beta)p(\mathbf{Z})p(\mathbf{X}|\mathbf{W}, \mathbf{Z}). \quad (49)$$

The trial joint distribution of  $\mathbf{Z}$ ,  $\beta$  and  $\rho$  is chosen as  $q(\mathbf{Z}, \beta, \rho) = q(\mathbf{Z})q(\beta)q(\rho)$ . By applying the VB inference, the probability distributions of  $\mathbf{Z}$ ,  $\beta$  and  $\rho$  are approximated respectively. What's more,  $\mathbf{W}$  is updated by minimizing the robust reconstruction error of all samples.

#### 4.5. Bayesian Robust Factor Analysis

In previous probabilistic models of PCA, the noise  $\varepsilon$  obeys the same Gaussian distributions. However, different features maybe have different noise levels in practical applications. Now, we set  $\varepsilon \sim \mathcal{N}(\mathbf{0}, \text{diag}(\boldsymbol{\tau}))$ , where  $\boldsymbol{\tau} = (\tau_1, \tau_2, \dots, \tau_d)^T$  and  $\text{diag}(\boldsymbol{\tau})$  is a diagonal matrix. Probabilistic Factor Analysis (or PCA) [13] further assumes that  $W_{ij} \sim \mathcal{N}(0, \tau_i^{-1}\alpha_j^{-1})$ ,  $\mu_i \sim \mathcal{N}(0, \tau_i^{-1}\beta^{-1})$ . In other words, different  $W_{ij}$  or  $\mu_i$  have different variances and the variances of  $W_{ij}$  and  $\mu_i$  also have a coupling relationship. Given  $\mathbf{W}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\tau}$ , the conditional probability distribution of  $\mathbf{x}$  is written as

$$p(\mathbf{x}|\mathbf{W}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\tau}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \text{diag}(\boldsymbol{\tau})). \quad (50)$$

Meanwhile, the prior distributions for  $\tau_i, \alpha_j$  and  $\beta$  are given as follows:

$$\tau_i \sim \text{Gam}(a_0, b_0), \alpha_j \sim \text{Gam}(c_0, d_0), \beta \sim \text{Gam}(e_0, f_0) \quad (51)$$

where  $\{a_0, b_0, c_0, d_0, e_0, f_0\}$  is the set of hyperparameters.

The robust version of Bayesian factor analysis [13] considers the Student's  $t$ -distributions instead of Gaussian noise corruptions due to the fact that the heavy tail of Student's  $t$ -distributions makes it more robust to outliers or large sparse noise. Let  $\varepsilon \sim \text{St}(0, \boldsymbol{\tau}^{-1}, \mathbf{v})$ , where  $\boldsymbol{\tau}^{-1} = (1/\tau_1, 1/\tau_2, \dots, 1/\tau_N)^T$  and  $\mathbf{v} = (v_1, v_2, \dots, v_N)^T$ . In this case, the probability distribution of  $\mathbf{x}$  for given  $\mathbf{W}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{v}$  is

$$p(\mathbf{x}|\mathbf{W}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{v}) = \text{St}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \boldsymbol{\tau}^{-1}, \mathbf{v}). \quad (52)$$

We can represent hierarchically the above Student's  $t$ -distributions. Firstly, the conditional probability distribution of  $\mathbf{x}_k$  can be expressed as:

$$p(\mathbf{x}_k|\mathbf{W}, \mathbf{z}_k, \boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{u}_k) = \mathcal{N}(\mathbf{x}_k|\mathbf{W}\mathbf{z}_k + \boldsymbol{\mu}, \text{diag}(\boldsymbol{\tau}^{-1} * \mathbf{u}_k^{-1})) \quad (53)$$

where  $\mathbf{u}_k$  is a  $d$ -dimensional column vector, " $*$ " is the Hadamard product (also known as entrywise product). Then we give the prior of  $\mathbf{u}_k$  for fixed  $d$ -dimensional hyper parameter vector  $\mathbf{v}$  as below:

$$p(\mathbf{u}_k|\mathbf{v}) = \text{Gam}(\mathbf{u}_k|\mathbf{v}/2, \mathbf{v}/2). \quad (54)$$

Another Bayesian robust factor analysis is on the basis of the Laplace distribution. At this moment, we assume that  $\varepsilon \sim \text{Lap}(0, \boldsymbol{\tau}^{-1/2})$ , where  $\boldsymbol{\tau}^{-1/2} = (\tau_1^{-1/2}, \tau_2^{-1/2}, \dots, \tau_N^{-1/2})^T$ . In this case, we have

$$p(\mathbf{x}|\mathbf{W}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\tau}) = \text{Lap}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \boldsymbol{\tau}^{-1/2}) \quad (55)$$

The Laplace distribution generally leads to adverse effects on inferring the probability distributions of other random variables. Here, we still employ the hierarchical method, that is,

$$p(\mathbf{x}|\mathbf{W}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{u}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \text{diag}(\boldsymbol{\tau}^{-1} * \mathbf{u}^{-1})), p(\mathbf{u}) = \text{IGam}(\mathbf{u}|1, 1/2). \quad (56)$$

Under this circumstance, we have

$$\begin{aligned}
 p(\mathbf{x}|\mathbf{W}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\tau}) &= \int p(\mathbf{x}, \mathbf{u}|\mathbf{W}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\tau}) d\mathbf{u} = \int p(\mathbf{x}|\mathbf{W}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\tau}, \mathbf{u}) p(\mathbf{u}) d\mathbf{u} \\
 &= \int \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \text{diag}(\boldsymbol{\tau}^{-1} * \mathbf{u}^{-1})) \text{IGam}(\mathbf{u}|1, 1/2) d\mathbf{u} \\
 &\propto \int \prod_{i=1}^N (\tau_i u_i)^{1/2} \exp\left(-\frac{1}{2} \tau_i u_i (\mathbf{x} - \mathbf{W}\mathbf{z} - \boldsymbol{\mu})_i^2\right) u_i^{-2} \exp\left(-\frac{1}{2u_i}\right) du_i \\
 &\propto \prod_{i=1}^N \int u_i^{-3/2} \exp\left(-\frac{1}{2u_i}\right) \exp\left(-\frac{1}{2} \tau_i (\mathbf{x} - \mathbf{W}\mathbf{z} - \boldsymbol{\mu})_i^2 u_i\right) du_i \\
 &\propto \prod_{i=1}^N \mathcal{L}\left\{u_i^{-3/2} \exp\left(-\frac{1}{2u_i}\right)\right\} \left(\frac{1}{2} \tau_i (\mathbf{x} - \mathbf{W}\mathbf{z} - \boldsymbol{\mu})_i^2\right)
 \end{aligned} \tag{57}$$

where  $(\mathbf{x} - \mathbf{W}\mathbf{z} - \boldsymbol{\mu})_i$  is the  $i$ -th element of vector  $\mathbf{x} - \mathbf{W}\mathbf{z} - \boldsymbol{\mu}$ . Because

$$\mathcal{L}^{-1}\left\{\frac{\sqrt{2}}{2} \exp\left(-\sqrt{\tau_i (\mathbf{x} - \mathbf{W}\mathbf{z} - \boldsymbol{\mu})_i^2}\right)\right\}(t) = \frac{1}{2\sqrt{\pi}} t^{-3/2} \exp\left(-\frac{1}{2t}\right) \tag{58}$$

we have  $p(\mathbf{x}|\mathbf{W}, \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\tau}) \propto \prod_{i=1}^N \exp(-|\sqrt{\tau_i}(\mathbf{x} - \mathbf{W}\mathbf{z} - \boldsymbol{\mu})_i|)$ . Hence, Equation (56) holds.

For the aforementioned two probabilistic models of robust factor analysis, the VB inference was proposed to approximate the posterior distributions [13]. In practice, the probability distribution of the noise should be chosen based on the application. The probabilistic models of PCA are compared in Table 2.

**Table 2.** Probabilistic models of Principal Component Analysis (PCA) with the factorization  $\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}$ .

Probabilistic Model	Deterministic Variables	Random Variables	Prior Distributions	Solving Strategy
Probabilistic PCA [10]	$\mathbf{W}, \boldsymbol{\mu}$	$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d),$ $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_r).$	-	ML EM
Bayesian PCA [11]	-	$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d),$ $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_r),$ $\mathbf{w}_i \sim \mathcal{N}(\mathbf{0}, \alpha_i^{-1} \mathbf{I}_d),$ $\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}, \beta^{-1} \mathbf{I}).$	$p(\boldsymbol{\alpha}) = \prod_{i=1}^r \text{Gam}(\alpha_i   a_0, b_0),$ $p(\boldsymbol{\tau}) = \text{Gam}(\boldsymbol{\tau}   c_0, d_0),$ $\boldsymbol{\tau} = \sigma^2.$	VB
Robust L1 PCA [12]	$\mathbf{W}, \boldsymbol{\mu}$ ( $\boldsymbol{\mu} = \mathbf{0}$ )	$\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, (\sigma^2 / \beta_i) \mathbf{I}_d),$ $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d).$	$\beta_i \sim \text{IGam}(1, 1/2),$ $p(\rho) = \text{Gam}(\rho   a, b),$ $\rho = 1/\sigma^2.$	VB
Probabilistic factor analysis [13]	-	$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d),$ $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_r),$ $w_{ij} \sim \mathcal{N}(0, \tau_i^{-1} \alpha_j^{-1}),$ $\mu_i \sim \mathcal{N}(0, \tau_i^{-1} \beta^{-1}).$	$\tau_i \sim \text{Gam}(a_0, b_0),$ $\alpha_j \sim \text{Gam}(c_0, d_0),$ $\beta \sim \text{Gam}(e_0, f_0).$	VB
Bayesian robust PCA I [13]	-	$\boldsymbol{\varepsilon}_k \sim \mathcal{N}(\mathbf{0}, \text{diag}(\boldsymbol{\tau}^{-1} * \mathbf{u}_k^{-1})),$ $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_r),$ $W_{ij} \sim \mathcal{N}(0, \tau_i^{-1} \alpha_j^{-1}),$ $\mu_i \sim \mathcal{N}(0, \tau_i^{-1} \beta^{-1}).$	$p(\mathbf{u}_k   \mathbf{v}) = \text{Gam}(\mathbf{u}_k   \mathbf{v}/2, \mathbf{v}/2),$ $\tau_i \sim \text{Gam}(a_0, b_0),$ $\alpha_j \sim \text{Gam}(c_0, d_0),$ $\beta \sim \text{Gam}(e_0, f_0).$	VB
Bayesian robust PCA II [13]	-	$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \text{diag}(\boldsymbol{\tau}^{-1} * \mathbf{u}^{-1})),$ $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_r),$ $w_{ij} \sim \mathcal{N}(0, \tau_i^{-1} \alpha_j^{-1}),$ $\mu_i \sim \mathcal{N}(0, \tau_i^{-1} \beta^{-1}).$	$p(\mathbf{u}) = \text{IGam}(\mathbf{u}   1, 1/2),$ $\tau_i \sim \text{Gam}(a_0, b_0),$ $\alpha_j \sim \text{Gam}(c_0, d_0),$ $\beta \sim \text{Gam}(e_0, f_0).$	VB

### 5. Probabilistic Models of Matrix Factorizations

Matrix factorizations are a type of methods to approximating the data matrix by the product of two low-rank matrices. They can be regarded as a special case of PCA without considering the mean. This section will discuss the existing probabilistic models of matrix factorizations.



### 5.1. Matrix Factorizations

For given data matrix  $\mathbf{X}$ , its low-rank factorization model is written as

$$\mathbf{X} = \mathbf{WZ} + \mathbf{E} \quad (59)$$

where  $\mathbf{W} \in \mathbb{R}^{d \times r}$ ,  $\mathbf{Z} \in \mathbb{R}^{r \times N}$ ,  $\mathbf{E}$  is the noise matrix and  $r < d$ . We assume that  $E_{ij} \sim \mathcal{N}(0, \sigma^2)$  are independent and identically distributed. We can get the optimal  $\mathbf{W}$  and  $\mathbf{Z}$  according to the maximum likelihood estimation. More specifically, we need to solve the following minimization problem:

$$\min_{\mathbf{W}, \mathbf{Z}} \|\mathbf{X} - \mathbf{WZ}\|_F. \quad (60)$$

The closed-form solution to problem (60) can be obtained by the Singular Value Decomposition (SVD).

To enhance the robustness to outliers or large sparse noise, we now assume that  $E_{ij} \sim \text{Lap}(0, \sigma)$ . For the moment, we solve the following optimization problem:

$$\min_{\mathbf{W}, \mathbf{Z}} \|\mathbf{X} - \mathbf{WZ}\|_1 \quad (61)$$

where  $\|\cdot\|_1$  is the L1-norm of a matrix (i.e., the sum of the absolute value of all elements). This problem is also called L1-norm PCA and the corresponding optimization methods were proposed in [3,4]. Srebro and Jaakkola considered the weighted low-rank approximations problems and provided an EM algorithm [94].

### 5.2. Probabilistic Matrix Factorization

We still consider Gaussian noise corruptions, that is,  $E_{ij} \sim \mathcal{N}(0, \sigma^2)$ . Furthermore, the zero-mean spherical Gaussian priors are respectively imposed on each row of  $\mathbf{W}$  and each column of  $\mathbf{Z}$ :

$$p(\mathbf{W}|\sigma_{\mathbf{W}}^2) = \prod_{i=1}^d \mathcal{N}(\mathbf{w}_i | \mathbf{0}, \sigma_{\mathbf{W}}^2 \mathbf{I}_r), p(\mathbf{Z}|\sigma_{\mathbf{Z}}^2) = \prod_{j=1}^N \mathcal{N}(\mathbf{z}_j | \mathbf{0}, \sigma_{\mathbf{Z}}^2 \mathbf{I}_r). \quad (62)$$

Probabilistic matrix factorization (PMF) [15] regards both  $\sigma_{\mathbf{W}}^2$  and  $\sigma_{\mathbf{Z}}^2$  as two deterministic parameters. The point estimations of  $\mathbf{W}, \mathbf{Z}$  can be obtained by maximizing the posterior distribution with the following form:

$$\begin{aligned} p(\mathbf{W}, \mathbf{Z} | \mathbf{X}, \sigma^2, \sigma_{\mathbf{W}}^2, \sigma_{\mathbf{Z}}^2) &\propto p(\mathbf{W}, \mathbf{Z}, \mathbf{X} | \sigma^2, \sigma_{\mathbf{W}}^2, \sigma_{\mathbf{Z}}^2) \\ &\propto p(\mathbf{W} | \sigma_{\mathbf{W}}^2) p(\mathbf{Z} | \sigma_{\mathbf{Z}}^2) p(\mathbf{X} | \mathbf{W}, \mathbf{Z}, \sigma^2). \end{aligned} \quad (63)$$

If the priors are respectively placed on  $\sigma_{\mathbf{W}}^2$  and  $\sigma_{\mathbf{Z}}^2$ , we can obtain the generalized model of probabilistic matrix factorization [15]. In this case, the likelihood function is derived as

$$\begin{aligned} p(\mathbf{W}, \mathbf{Z}, \sigma^2, \sigma_{\mathbf{W}}^2, \sigma_{\mathbf{Z}}^2 | \mathbf{X}) &\propto p(\mathbf{W}, \mathbf{Z}, \sigma^2, \sigma_{\mathbf{W}}^2, \sigma_{\mathbf{Z}}^2, \mathbf{X}) \\ &\propto p(\mathbf{X} | \mathbf{W}, \mathbf{Z}, \sigma^2) p(\mathbf{W} | \sigma_{\mathbf{W}}^2) p(\mathbf{Z} | \sigma_{\mathbf{Z}}^2) p(\sigma_{\mathbf{W}}^2) p(\sigma_{\mathbf{Z}}^2). \end{aligned} \quad (64)$$

By maximizing the above posterior distribution, we can obtain the point estimations of parameters  $\{\mathbf{W}, \mathbf{Z}\}$  and hyperparameters  $\{\sigma_{\mathbf{W}}^2, \sigma_{\mathbf{Z}}^2\}$ . Furthermore, two derivatives of PMF are also presented, i.e., PMF with a adaptive prior and PMF with constraining user-specific feature vectors.

In [89], a fully-observed variational Bayesian matrix factorization, an extension of PMF, was discussed. Meanwhile, it is shown that this new probabilistic matrix factorization can weaken the decomposability assumption and obtain the exact global analytic solution for rectangular cases.

### 5.3. Variational Bayesian Approach to Probabilistic Matrix Factorization

In PMF,  $W_{ik}$  are independent and identically distributed and so are  $Z_{kj}$ . Variational Bayesian PMF [14] assumes the entries from different columns of  $\mathbf{W}$  or  $\mathbf{Z}^T$  have different variances, that is,  $W_{ik} \sim \mathcal{N}(0, \sigma_k^2)$ ,  $Z_{kj} \sim \mathcal{N}(0, \rho_k^2)$ ,  $k = 1, 2, \dots, r$ . For given hyperparameters  $\{\sigma^2, \sigma_{\mathbf{W}}^2, \rho_{\mathbf{Z}}^2\}$ , we get the joint probability distribution:

$$p(\mathbf{X}, \mathbf{W}, \mathbf{Z} | \sigma^2, \sigma_{\mathbf{W}}^2, \rho_{\mathbf{Z}}^2) = p(\mathbf{X} | \mathbf{W}, \mathbf{Z}, \sigma^2) p(\mathbf{W} | \sigma_{\mathbf{W}}^2) p(\mathbf{Z} | \rho_{\mathbf{Z}}^2) \tag{65}$$

where  $\sigma_{\mathbf{W}}^2 = (\sigma_1^2, \sigma_2^2, \dots, \sigma_r^2)^T$  and  $\rho_{\mathbf{Z}}^2 = (\rho_1^2, \rho_2^2, \dots, \rho_r^2)^T$ .

We assume that the trial joint distribution of  $\mathbf{W}$  and  $\mathbf{Z}$  is decomposable, that is,  $q(\mathbf{W}, \mathbf{Z}) = q(\mathbf{W})q(\mathbf{Z})$ . Using VB method, we can update alternatively  $q(\mathbf{W})$  and  $q(\mathbf{Z})$ . The variances  $\{\sigma^2, \sigma_{\mathbf{W}}^2, \rho_{\mathbf{Z}}^2\}$  can be determined by maximizing the following lower bound:

$$L(q(\mathbf{W}, \mathbf{Z})) = \mathbb{E}_{q(\mathbf{W}), q(\mathbf{Z})} \left[ \ln p(\mathbf{X}, \mathbf{W}, \mathbf{Z} | \sigma^2, \sigma_{\mathbf{W}}^2, \rho_{\mathbf{Z}}^2) - \ln q(\mathbf{W})q(\mathbf{Z}) \right]. \tag{66}$$

The experimental results in Netflix Prize competition show that the Variational Bayesian approach has superiority over MAP and greedy residual fitting.

### 5.4. Bayesian Probabilistic Matrix Factorizations Using Markov Chain Monte Carlo

Variational Bayesian approach to PMF only discusses the case that  $W_{ik}$  (or  $Z_{kj}$ ) are independent and identically distributed and their means are zeros. However, Bayesian PMF [16] assumes that  $\mathbf{w}_i$ . (or  $\mathbf{z}_j$ ) are independent and identically distributed and their mean vectors are not zero vectors. Concretely speaking, we stipulate that

$$p(\mathbf{W} | \mu_{\mathbf{W}}, \Lambda_{\mathbf{W}}) = \prod_{i=1}^d \mathcal{N}(\mathbf{w}_i | \mu_{\mathbf{W}}, \Lambda_{\mathbf{W}}^{-1}), p(\mathbf{Z} | \mu_{\mathbf{Z}}, \Lambda_{\mathbf{Z}}) = \prod_{j=1}^N \mathcal{N}(\mathbf{z}_j | \mu_{\mathbf{Z}}, \Lambda_{\mathbf{Z}}^{-1}). \tag{67}$$

Let  $\Theta_{\mathbf{W}} = \{\mu_{\mathbf{W}}, \Lambda_{\mathbf{W}}\}$ ,  $\Theta_{\mathbf{Z}} = \{\mu_{\mathbf{Z}}, \Lambda_{\mathbf{Z}}\}$ . We further suppose the prior distributions of  $\Theta_{\mathbf{W}}$  and  $\Theta_{\mathbf{Z}}$  are Gaussian–Wishart priors:

$$\begin{cases} p(\Theta_{\mathbf{W}} | \Theta_0) = p(\mu_{\mathbf{W}} | \Lambda_{\mathbf{W}}) p(\Lambda_{\mathbf{W}}) = \mathcal{N}(\mu_{\mathbf{W}} | \mu_0, (\beta_0 \Lambda_{\mathbf{W}})^{-1}) \mathcal{W}(\Lambda_{\mathbf{W}} | \mathbf{W}_0, v_0) \\ p(\Theta_{\mathbf{Z}} | \Theta_0) = p(\mu_{\mathbf{Z}} | \Lambda_{\mathbf{Z}}) p(\Lambda_{\mathbf{Z}}) = \mathcal{N}(\mu_{\mathbf{Z}} | \mu_0, (\beta_0 \Lambda_{\mathbf{Z}})^{-1}) \mathcal{W}(\Lambda_{\mathbf{Z}} | \mathbf{W}_0, v_0) \end{cases} \tag{68}$$

where  $\Theta_0 = \{\mu_0, v_0, \mathbf{W}_0\}$ ,  $\beta_0$  is a hyper parameter.

We can initialize the parameters  $\Theta_0$  as follows:  $\mu_0 = \mathbf{0}, v_0 = d, \mathbf{W}_0 = \mathbf{I}_d$ . In theory, the predictive probability distribution of  $X_{ij}^*$  can be obtained by marginalizing over model parameters and hyperparameters:

$$\begin{aligned} p(X_{ij}^* | \mathbf{X}, \Theta_0) &= \int \int \int \int p(X_{ij}^* | \mathbf{W}, \mathbf{Z}, \Theta_{\mathbf{W}}, \Theta_{\mathbf{Z}} | \mathbf{X}, \Theta_0) d\mathbf{W} d\mathbf{Z} d\Theta_{\mathbf{W}} d\Theta_{\mathbf{Z}} \\ &= \int \int \int \int p(X_{ij}^* | \mathbf{w}_i, \mathbf{z}_j) p(\mathbf{W}, \mathbf{Z} | \mathbf{X}, \Theta_{\mathbf{W}}, \Theta_{\mathbf{Z}}) p(\Theta_{\mathbf{W}} | \Theta_0) p(\Theta_{\mathbf{Z}} | \Theta_0) d\mathbf{W} d\mathbf{Z} d\Theta_{\mathbf{W}} d\Theta_{\mathbf{Z}}. \end{aligned} \tag{69}$$

However, the above integral is analytically intractable due to the fact that it is very difficult to determine the posterior distribution. Based on this, Gibbs sampling, one of the simplest Markov chain Monte Carlo, was proposed to approximate the predictive distribution  $p(X_{ij}^* | \mathbf{X}, \Theta_0)$ . It is noted that MCMC methods for large-scale problems require especial care for efficient proposals and may be very slow if the sample correlation is too long.

### 5.5. Sparse Bayesian Matrix Completion

We consider the case that some elements of data matrix  $\mathbf{X}$  are missing and the observed index set is denoted by  $\Omega$ . Matrix completion assumes that  $\mathbf{X}$  is approximately low-rank and its goal is to recover all missing elements from observed elements.

For noise matrix  $\mathbf{E}$ , we assume that  $E_{ij} \sim \mathcal{N}(0, \beta^{-1})$ . In sparse Bayesian matrix completion [19], the Gaussian distributions are imposed on two low-rank matrices, that is,

$$p(\mathbf{W}|\boldsymbol{\gamma}) = \prod_{i=1}^r \mathcal{N}(\mathbf{w}_{\cdot i}|\gamma_i^{-1}\mathbf{I}_d), p(\mathbf{Z}|\boldsymbol{\gamma}) = \prod_{i=1}^r \mathcal{N}(\mathbf{z}_{i\cdot}|\gamma_i^{-1}\mathbf{I}_r). \tag{70}$$

Moreover, the priors of  $\gamma_i$  are given by  $\gamma_i \sim \text{Gam}(a, b)$  and the prior of  $\beta$  is assigned the noninformative Jeffrey’s prior:  $p(\beta) \propto \beta^{-1}$ . It is obvious that

$$p(\mathbf{X}|\mathbf{W}, \mathbf{Z}, \beta) = \prod_{(i,j) \in \Omega} \mathcal{N}(X_{ij}|\mathbf{w}_{i\cdot}\mathbf{z}_{\cdot j}, \beta^{-1}). \tag{71}$$

Then the joint probability distribution is

$$p(\mathbf{X}, \mathbf{W}, \mathbf{Z}, \boldsymbol{\gamma}, \beta) = p(\mathbf{X}|\mathbf{W}, \mathbf{Z}, \beta)p(\mathbf{W}|\boldsymbol{\gamma})p(\mathbf{Z}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})p(\beta). \tag{72}$$

Let  $q(\mathbf{W}, \mathbf{Z}, \boldsymbol{\gamma}, \beta)$  be the approximated distribution of  $p(\mathbf{W}, \mathbf{Z}, \boldsymbol{\gamma}, \beta)$ . The approximated procedure can be achieved by VB inference. It is demonstrated that the proposed method achieves a better prediction error in recommendation systems.

### 5.6. Robust Bayesian Matrix Factorization

In previous probabilistic models of matrix factorizations, there is no relationship among the variances of  $W_{ik}, Z_{kj}, E_{ij}$ . Now, we reconsider the probability distributions of  $W_{ik}, Z_{kj}, E_{ij}$ . The noise  $E_{ij}$  is chosen as the heteroscedastic Gaussian scale mixture distribution:

$$E_{ij} \sim \mathcal{N}(0, (\tau\alpha_i\beta_j)^{-1}) \tag{73}$$

and the probability distributions of  $\mathbf{w}_{i\cdot}$  and  $\mathbf{z}_{\cdot j}$  are given by:

$$p(\mathbf{w}_{i\cdot}|\alpha_i) = \mathcal{N}(\mathbf{w}_{i\cdot}|\mathbf{0}, (\alpha_i\boldsymbol{\Lambda}_{\mathbf{W}})^{-1}), p(\mathbf{z}_{\cdot j}|\beta_j) = \mathcal{N}(\mathbf{z}_{\cdot j}|\mathbf{0}, (\beta_j\boldsymbol{\Lambda}_{\mathbf{Z}})^{-1}). \tag{74}$$

We also impose Gamma distribution priors on  $\alpha_i$  and  $\beta_j$ :

$$p(\alpha_i) = \text{Gam}(\alpha_i|a_0/2, b_0/2), p(\beta_j) = \text{Gam}(\beta_j|c_0/2, d_0/2) \tag{75}$$

where  $\{a_0, b_0, c_0, d_0\}$  is a given set of hyper-parameters. To reduce this problem’s complexity, we restrict  $\boldsymbol{\Lambda}_{\mathbf{W}}$  and  $\boldsymbol{\Lambda}_{\mathbf{Z}}$  to be diagonal matrices, that is,

$$\boldsymbol{\Lambda}_{\mathbf{W}}^{-1} = \text{diag}(\sigma_1^2, \dots, \sigma_r^2), \boldsymbol{\Lambda}_{\mathbf{Z}}^{-1} = \text{diag}(\rho_1^2, \dots, \rho_r^2). \tag{76}$$

Let  $\theta = \{\tau, \boldsymbol{\Lambda}_{\mathbf{W}}, \boldsymbol{\Lambda}_{\mathbf{Z}}, a_0, b_0, c_0, d_0\}$ . For the fixed parameters  $\theta$ , the joint probability distribution is

$$p(\mathbf{X}, \mathbf{W}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}|\theta) = p(\mathbf{X}|\mathbf{W}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \theta)p(\mathbf{W}|\boldsymbol{\alpha})p(\mathbf{Z}|\boldsymbol{\beta})p(\boldsymbol{\alpha}|\theta)p(\boldsymbol{\beta}|\theta). \tag{77}$$

We consider two types of approximated posteriors of  $p(\mathbf{W}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}|\mathbf{X}, \theta)$ , one is

$$q(\mathbf{W}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = q(\mathbf{W})q(\mathbf{Z})q(\boldsymbol{\alpha})q(\boldsymbol{\beta}) \tag{78}$$

and another has the form

$$q(\mathbf{W}, \mathbf{Z}, \alpha, \beta) = q(\mathbf{W}, \alpha)q(\mathbf{Z}, \beta). \quad (79)$$

For the above two cases, we can obtain respectively the trial probability distribution  $q(\mathbf{W}, \mathbf{Z}, \alpha, \beta)$  by VB method [17].

In addition, a structured variational approximation was also proposed in [17]. We assume that the variational posteriors of  $\mathbf{W}$  and  $\mathbf{Z}$  obey Gaussian distributions:

$$q(\mathbf{W}|\alpha) = \prod_{i=1}^d \mathcal{N}(\mathbf{w}_i | \bar{\mathbf{w}}_i, \alpha_i \bar{\mathbf{S}}_i), q(\mathbf{Z}|\beta) = \prod_{j=1}^N \mathcal{N}(\mathbf{z}_j | \bar{\mathbf{z}}_j, \beta_j \bar{\mathbf{R}}_j). \quad (80)$$

The free energy function is defined by the negative lower bound:

$$-L(q) = -\int \int \int q(\mathbf{W}, \mathbf{Z}, \alpha, \beta) \ln \frac{p(\mathbf{X}, \mathbf{W}, \mathbf{Z}, \alpha, \beta | \theta)}{q(\mathbf{W}, \mathbf{Z}, \alpha, \beta)} d\mathbf{W} d\mathbf{Z} d\alpha d\beta \quad (81)$$

By directly minimizing the free energy function with respect to  $\bar{\mathbf{w}}_i, \bar{\mathbf{S}}_i, \bar{\mathbf{z}}_j, \bar{\mathbf{R}}_j$ , we can obtain the optimal  $\bar{\mathbf{w}}_i, \bar{\mathbf{S}}_i, \bar{\mathbf{z}}_j, \bar{\mathbf{R}}_j$ . The variational posteriors of scale variables  $\alpha$  and  $\beta$  can also be recognized as the generalized inverse Gaussians.

The parameters  $\theta$  can be estimated by type II maximum likelihood or empirical Bayes. In other words, we update the parameters by minimizing directly the negative lower bound. Robust Bayesian matrix factorization shows that the heavy-tailed distributions are useful to incorporate robustness information to the probabilistic models.

### 5.7. Probabilistic Robust Matrix Factorization

The model of probabilistic robust matrix factorization [18] considers the sparse noise corruptions. In this model, the Gaussian distributions are also imposed on  $W_{ik}$  and  $Z_{kj}$ :

$$p(W_{ik} | \lambda_{\mathbf{W}}) = \mathcal{N}(W_{ik} | 0, \lambda_{\mathbf{W}}^{-1}), p(Z_{kj} | \lambda_{\mathbf{Z}}) = \mathcal{N}(Z_{kj} | 0, \lambda_{\mathbf{Z}}^{-1}) \quad (82)$$

and the Laplace noise is placed on  $E_{ij}$ , that is,  $E_{ij} \sim \text{Lap}(0, \lambda)$ .

From Bayes' rule, we have

$$p(\mathbf{W}, \mathbf{Z} | \mathbf{X}, \lambda, \lambda_{\mathbf{W}}, \lambda_{\mathbf{Z}}) \propto p(\mathbf{X} | \mathbf{W}, \mathbf{Z}, \lambda) p(\mathbf{W} | \lambda_{\mathbf{W}}) p(\mathbf{Z} | \lambda_{\mathbf{Z}}). \quad (83)$$

We construct the hierarchical model for the Laplace distribution:

$$p(X_{ij} | \mathbf{W}, \mathbf{Z}, \mathbf{T}) = \mathcal{N}(X_{ij} | \mathbf{w}_i \cdot \mathbf{z}_j, T_{ij}), p(T_{ij} | \lambda) = \text{Exp}(T_{ij} | \lambda / 2). \quad (84)$$

We regard  $\mathbf{T}$  as a latent variable matrix and denote  $\theta = \{\mathbf{W}, \mathbf{Z}\}, \hat{\theta} = \{\hat{\mathbf{W}}, \hat{\mathbf{Z}}\}$ , where  $\hat{\theta}$  is the current estimation of  $\theta$ . An EM algorithm was proposed to inferring  $\mathbf{W}$  and  $\mathbf{Z}$  [18]. To this end, we construct the so-called Q-function:

$$Q(\mathbf{Z} | \hat{\theta}) = \mathbb{E}_{\mathbf{T}} [\log p(\mathbf{Z} | \hat{\mathbf{W}}, \mathbf{X}, \mathbf{T}) | \mathbf{X}, \hat{\theta}]. \quad (85)$$

The posterior of complete-data is

$$\begin{aligned} p(\mathbf{Z} | \hat{\mathbf{W}}, \mathbf{X}, \mathbf{T}) &= \frac{p(\mathbf{Z}, \mathbf{X} | \hat{\mathbf{W}}, \mathbf{T})}{p(\mathbf{X} | \hat{\mathbf{W}}, \mathbf{T})} \propto p(\mathbf{Z}, \mathbf{X} | \hat{\mathbf{W}}, \mathbf{T}) \\ &\propto p(\mathbf{X} | \mathbf{Z}, \hat{\mathbf{W}}, \mathbf{T}) p(\mathbf{Z}) \end{aligned} \quad (86)$$

Hence, its log is

$$\log p(\mathbf{Z}|\hat{\mathbf{W}}, \mathbf{X}, \mathbf{T}) = -\frac{1}{2} \sum_{i=1}^d \sum_{j=1}^N T_{ij}^{-1} (X_{ij} - \hat{\mathbf{w}}_{i,\mathbf{z},j})^2 - \frac{\lambda_{\mathbf{Z}}}{2} \sum_j \mathbf{z}_j^T \mathbf{z}_j + \text{const} \tag{87}$$

where “const” is a term independent of  $\mathbf{Z}$ .

To compute the expectations of  $\mathbf{T}$ , we derive the conditional probability distribution of  $\mathbf{T}$  as follows:

$$p(\mathbf{T}|\mathbf{X}, \hat{\mathbf{W}}, \hat{\mathbf{Z}}) = \frac{p(\mathbf{T}, \mathbf{X}|\hat{\mathbf{W}}, \hat{\mathbf{Z}})}{p(\mathbf{X}|\hat{\mathbf{W}}, \hat{\mathbf{Z}})} \propto p(\mathbf{T}, \mathbf{X}|\hat{\mathbf{W}}, \hat{\mathbf{Z}}) \propto p(\mathbf{T})p(\mathbf{X}|\mathbf{T}, \hat{\mathbf{W}}, \hat{\mathbf{Z}}). \tag{88}$$

Hence,  $T_{ij}^{-1}$  follows an inverse Gaussian distribution and its posterior expectation is given by

$$\mathbb{E} [T_{ij}^{-1}|\mathbf{X}, \hat{\mathbf{W}}, \hat{\mathbf{Z}}] = \frac{\sqrt{\lambda}}{|X_{ij} - (\hat{\mathbf{W}}\hat{\mathbf{Z}})_{ij}|}. \tag{89}$$

Thus, we get  $Q(\mathbf{Z}|\hat{\theta})$ .

To obtain the update of  $\mathbf{Z}$ , we maximize the function  $Q(\mathbf{Z}|\hat{\theta})$  with respect to  $\mathbf{z}_j$ . By setting  $\frac{\partial}{\partial \mathbf{z}_j} Q(\mathbf{Z}|\hat{\theta}) = \mathbf{0}$ , we can get the closed-form solution of  $\mathbf{z}_j$ . The update rule for  $\mathbf{W}$  is similar to that of  $\mathbf{Z}$ . The proposed probabilistic model is robust again outliers and missing data and equivalent to robust PCA under mild conditions [18].

### 5.8. Bayesian Robust Matrix Factorization

Another robust probabilistic model of matrix factorizations is Bayesian robust matrix factorization [20]. Gaussian distributions are still imposed on  $\mathbf{W}$  and  $\mathbf{Z}$ , that is,

$$\mathbf{w}_i \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{W}}, \boldsymbol{\Lambda}_{\mathbf{W}}^{-1}), \mathbf{z}_j \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{Z}}, \boldsymbol{\Lambda}_{\mathbf{Z}}^{-1}), i = 1, 2, \dots, d, j = 1, 2, \dots, N. \tag{90}$$

We further assume both  $(\boldsymbol{\mu}_{\mathbf{W}}, \boldsymbol{\Lambda}_{\mathbf{W}})$  and  $(\boldsymbol{\mu}_{\mathbf{Z}}, \boldsymbol{\Lambda}_{\mathbf{Z}})$  follow Gaussian–Wishart distributions:

$$\boldsymbol{\Lambda}_{\mathbf{W}} \sim \mathcal{W}(\mathbf{W}_0, v_0), \boldsymbol{\mu}_{\mathbf{W}} \sim \mathcal{N}(\boldsymbol{\mu}_0, (\beta_0 \boldsymbol{\Lambda}_{\mathbf{W}})^{-1}), \boldsymbol{\Lambda}_{\mathbf{Z}} \sim \mathcal{W}(\mathbf{W}_0, v_0), \boldsymbol{\mu}_{\mathbf{Z}} \sim \mathcal{N}(\boldsymbol{\mu}_0, (\beta_0 \boldsymbol{\Lambda}_{\mathbf{Z}})^{-1}) \tag{91}$$

where  $\mathbf{W}_0, v_0, \boldsymbol{\mu}_0, \beta_0$  are the hyperparameters.

To enhance the robustness, we suppose the noise is the mixture of a Laplace distribution and a GIG distribution. Concretely speaking, the noise term  $E_{ij} \sim \text{Lap}(0, \eta_{ij})$  and the prior of  $\eta_{ij}$  is given by GIG( $p, a, b$ ), where  $p, a, b$  are three hyperparameters. Hence,  $E_{ij} \sim \text{Lap}(0, \eta_{ij})$  can be represented by two steps:  $E_{ij} \sim \mathcal{N}(0, T_{ij}), T_{ij} \sim \text{Exp}(\eta_{ij}/2)$ . According to the above probability distributions, we can generate  $\boldsymbol{\Lambda}_{\mathbf{W}}, \boldsymbol{\mu}_{\mathbf{W}}, \mathbf{W}, \boldsymbol{\mu}_{\mathbf{Z}}, \boldsymbol{\Lambda}_{\mathbf{Z}}, \mathbf{Z}, \boldsymbol{\eta}, \mathbf{T}$  and  $\mathbf{X}$  in turn.

Gibbs sampling is proposed to infer the posterior distributions. For this purpose, we need to derive the posterior distributions of all random variables. Due to the fact that the derivation process of  $\{\boldsymbol{\mu}_{\mathbf{Z}}, \boldsymbol{\Lambda}_{\mathbf{Z}}, \mathbf{Z}\}$  is similar to that of  $\{\boldsymbol{\mu}_{\mathbf{W}}, \boldsymbol{\Lambda}_{\mathbf{W}}, \mathbf{W}\}$ , the following only considers approximating the posterior distributions of  $(\boldsymbol{\mu}_{\mathbf{W}}, \boldsymbol{\Lambda}_{\mathbf{W}}, \mathbf{W})$  for brevity. Firstly, the posterior distribution of  $(\boldsymbol{\mu}_{\mathbf{W}}, \boldsymbol{\Lambda}_{\mathbf{W}})$  is a Gaussian–Wishart distribution because that

$$\begin{aligned} p(\boldsymbol{\mu}_{\mathbf{W}}, \boldsymbol{\Lambda}_{\mathbf{W}}|\mathbf{W}, \mathbf{W}_0, \boldsymbol{\mu}_0, v_0, \beta_0) &\propto p(\boldsymbol{\mu}_{\mathbf{W}}, \boldsymbol{\Lambda}_{\mathbf{W}}, \mathbf{W}, \mathbf{W}_0, \boldsymbol{\mu}_0, v_0, \beta_0) \\ &\propto \prod_{i=1}^d p(\mathbf{w}_i|\boldsymbol{\mu}_{\mathbf{W}}, \boldsymbol{\Lambda}_{\mathbf{W}}) p(\boldsymbol{\Lambda}_{\mathbf{W}}|\mathbf{W}_0, v_0) p(\boldsymbol{\mu}_{\mathbf{W}}|\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_{\mathbf{W}}, \beta_0) \end{aligned} \tag{92}$$

Then, we compute the conditional probability distribution of  $\mathbf{w}_i$ :

$$\begin{aligned} p(\mathbf{w}_i | \mathbf{X}, \mathbf{Z}, \boldsymbol{\mu}_W, \boldsymbol{\Lambda}_W, \mathbf{T}) &\propto p(\mathbf{w}_i, \mathbf{X}, \mathbf{Z}, \boldsymbol{\mu}_W, \boldsymbol{\Lambda}_W, \mathbf{T}) \\ &\propto p(\mathbf{x}_i | \mathbf{w}_i, \mathbf{z}_j, \mathbf{t}_i) p(\mathbf{w}_i | \boldsymbol{\mu}_W, \boldsymbol{\Lambda}_W) \\ &\propto \prod_{j=1}^N \mathcal{N}(X_{ij} | \mathbf{w}_i, \mathbf{z}_j, T_{ij}) \mathcal{N}(\mathbf{w}_i | \boldsymbol{\mu}_W, \boldsymbol{\Lambda}_W). \end{aligned} \quad (93)$$

According to the above Equation,  $p(\mathbf{w}_i | \mathbf{X}, \mathbf{Z}, \boldsymbol{\mu}_W, \boldsymbol{\Lambda}_W, \mathbf{T})$  is also Gaussian. Next, for given  $\{X_{ij}, \mathbf{w}_i, \mathbf{z}_j, \eta_{ij}\}$ , the probability distribution of  $T_{ij}$  is derived as below:

$$\begin{aligned} p(T_{ij} | X_{ij}, \mathbf{w}_i, \mathbf{z}_j, \eta_{ij}) &\propto p(T_{ij}, X_{ij}, \mathbf{w}_i, \mathbf{z}_j, \eta_{ij}) \\ &\propto p(X_{ij} | \mathbf{w}_i, \mathbf{z}_j, T_{ij}) p(T_{ij} | \eta_{ij}). \end{aligned} \quad (94)$$

So,  $T_{ij} | X_{ij}, \mathbf{w}_i, \mathbf{z}_j, \eta_{ij} \sim \text{GIG}(1/2, \eta_{ij}, r_{ij}^2)$ , where  $r_{ij} = X_{ij} - \mathbf{w}_i \cdot \mathbf{z}_j$ .

Finally, the posterior of  $\eta_{ij}$  satisfies that

$$p(\eta_{ij} | T_{ij}, p, a, b) \propto p(\eta_{ij}, T_{ij}, p, a, b) \propto p(T_{ij} | \eta_{ij}) p(\eta_{ij} | p, a, b) \quad (95)$$

Hence, it holds  $\eta_{ij} | T_{ij}, p, a, b \sim \text{GIG}(p + 1, T_{ij} + a, b)$ . Bayesian robust matrix factorization incorporates spatial or temporal proximity in computer vision applications and batch algorithms are proposed to infer parameters.

### 5.9. Bayesian Model for L1-Norm Low-Rank Matrix Factorizations

For low-rank matrix factorizations, L1-norm minimization is more robust than L2-norm minimization in the presence of outliers or non-Gaussian noises. Based on this, we assume that the noise  $E_{ij}$  follows the Laplace distribution:  $E_{ij} \sim \text{Lap}(0, \sqrt{\lambda/2})$ . Since the Laplace noise is inconvenient for Bayesian inference, a hierarchical Bayesian model was formulated in [21]. Concretely speaking, a two-level hierarchical prior is imposed on the Laplace prior:

$$E_{ij} \sim \mathcal{N}(0, T_{ij}), T_{ij} \sim \text{Exp}(\lambda). \quad (96)$$

The generative models of  $\mathbf{W}$  and  $\mathbf{Z}$  are constructed as follows:

$$\mathbf{w}_i \sim \mathcal{N}(\mathbf{0}, \tau_W \mathbf{I}_r), \mathbf{z}_j \sim \mathcal{N}(\mathbf{0}, \tau_Z \mathbf{I}_r), i = 1, 2, \dots, d, j = 1, 2, \dots, N. \quad (97)$$

In addition, Gamma priors are placed on the precision parameters of the above Gaussian distributions:

$$\tau_W \sim \text{Gam}(a_0, b_0), \tau_Z \sim \text{Gam}(c_0, d_0). \quad (98)$$

The trial posterior distribution for  $\{\mathbf{W}, \mathbf{Z}, \tau_W, \tau_Z, \mathbf{T}\}$  is specified as:

$$q(\mathbf{W}, \mathbf{Z}, \tau_W, \tau_Z, \mathbf{T}) = \prod_{i=1}^d q(\mathbf{w}_i) \prod_{j=1}^N q(\mathbf{z}_j) q(\tau_W) q(\tau_Z). \quad (99)$$

And the joint probability distribution is expressed as

$$p(\mathbf{W}, \mathbf{Z}, \tau_W, \tau_Z, \mathbf{T}, \mathbf{X}) = p(\mathbf{X} | \mathbf{W}, \mathbf{Z}, \mathbf{T}) p(\mathbf{W} | \tau_W) p(\mathbf{Z} | \tau_Z) p(\tau_W) p(\tau_Z) p(\mathbf{T}). \quad (100)$$

VB inference was adopted to approximate the full posterior distribution [21]. Furthermore, varying precision parameters are also considered for different rows of  $\mathbf{W}$  or different columns of  $\mathbf{Z}$ . All parameters are automatically tuned to adapt to the data, and the proposed method is applied in computer vision problems to validate its efficiency and robustness.

In Table 3, we list all probabilistic models of matrix factorizations discussed in this section, and compare their probability distributions, priors and solving strategy.

**Table 3.** Probabilistic models of matrix factorizations with  $\mathbf{X} = \mathbf{WZ} + \mathbf{E}$ .

Probabilistic Model	Random Variables	Prior Distributions	Solving Strategy
PMF [15]	$E_{ij} \sim \mathcal{N}(0, \sigma^2),$ $P(\mathbf{W} \sigma_{\mathbf{W}}^2) = \prod_{i=1}^r \mathcal{N}(\mathbf{w}_i 0, \sigma_{\mathbf{W}}^2 \mathbf{I}_d),$ $P(\mathbf{Z} \sigma_{\mathbf{Z}}^2) = \prod_{j=1}^N \mathcal{N}(\mathbf{z}_j 0, \sigma_{\mathbf{Z}}^2 \mathbf{I}_r).$	-	MAP
Variational Bayesian PMF [14]	$E_{ij} \sim \mathcal{N}(0, \sigma^2),$ $W_{ik} \sim \mathcal{N}(0, \sigma_k^2),$ $Z_{kj} \sim \mathcal{N}(0, \rho_k^2).$	-	VB
Bayesian PMF [16]	$E_{ij} \sim \mathcal{N}(0, \beta^{-1}),$ $p(\mathbf{W} \mu_{\mathbf{W}}, \Lambda_{\mathbf{W}}) = \prod_{i=1}^d \mathcal{N}(\mathbf{w}_i \mu_{\mathbf{W}}, \Lambda_{\mathbf{W}}^{-1}),$ $p(\mathbf{Z} \mu_{\mathbf{Z}}, \Lambda_{\mathbf{Z}}) = \prod_{j=1}^N \mathcal{N}(\mathbf{z}_j \mu_{\mathbf{Z}}, \Lambda_{\mathbf{Z}}^{-1}).$	$p(\Theta_{\mathbf{W}} \Theta_0) = \mathcal{N}(\mu_{\mathbf{W}} \mu_0, (\beta_0 \Lambda_{\mathbf{W}})^{-1}) \mathcal{W}(\Lambda_{\mathbf{W}} \mathbf{W}_0, v_0),$ $p(\Theta_{\mathbf{Z}} \Theta_0) = \mathcal{N}(\mu_{\mathbf{Z}} \mu_0, (\beta_0 \Lambda_{\mathbf{Z}})^{-1}) \mathcal{W}(\Lambda_{\mathbf{Z}} \mathbf{W}_0, v_0).$	Gibbs sampling
Sparse Bayesian matrix completion [19]	$E_{ij} \sim \mathcal{N}(0, \beta^{-1}),$ $p(\mathbf{W} \gamma) = \prod_{i=1}^r \mathcal{N}(\mathbf{w}_i \gamma_i^{-1} \mathbf{I}_d),$ $p(\mathbf{Z} \gamma) = \prod_{i=1}^r \mathcal{N}(\mathbf{z}_i \gamma_i^{-1} \mathbf{I}_r).$	$\gamma_i \sim \text{Gam}(a, b),$ $p(\beta) \propto \beta^{-1}.$	VB
Robust Bayesian matrix factorization [17]	$E_{ij} \sim \mathcal{N}(0, (\tau \alpha_i \beta_j)^{-1}),$ $p(\mathbf{w}_i \alpha_i) = \mathcal{N}(\mathbf{w}_i \mathbf{0}, (\alpha_i \Lambda_{\mathbf{W}})^{-1}),$ $p(\mathbf{z}_j \beta_j) = \mathcal{N}(\mathbf{z}_j \mathbf{0}, (\beta_j \Lambda_{\mathbf{Z}})^{-1}).$	$p(\alpha_i) = \text{Gam}(\alpha_i a_0/2, b_0/2),$ $p(\beta_j) = \text{Gam}(\beta_j c_0/2, d_0/2).$	VB, type II ML, empirical Bayes
Probabilistic robust matrix factorization [18]	$E_{ij} \sim \mathcal{N}(X_{ij} 0, T_{ij}),$ $p(W_{ik} \lambda_{\mathbf{W}}) = \mathcal{N}(W_{ik} 0, \lambda_{\mathbf{W}}^{-1}),$ $p(Z_{kj} \lambda_{\mathbf{Z}}) = \mathcal{N}(Z_{kj} 0, \lambda_{\mathbf{Z}}^{-1}).$	$T_{ij} \sim \text{Exp}(\lambda/2).$	EM
Bayesian robust matrix factorization [20]	$E_{ij} \sim \mathcal{N}(0, T_{ij}),$ $\mathbf{w}_i \sim \mathcal{N}(\mu_{\mathbf{W}}, \Lambda_{\mathbf{W}}^{-1}),$ $\mathbf{z}_j \sim \mathcal{N}(\mu_{\mathbf{Z}}, \Lambda_{\mathbf{Z}}^{-1}).$	$T_{ij} \sim \text{Exp}(\eta_{ij}/2),$ $\Lambda_{\mathbf{W}} \sim \mathcal{W}(\mathbf{W}_0, v_0),$ $\mu_{\mathbf{W}} \sim \mathcal{N}(\mu_0, (\beta_0 \Lambda_{\mathbf{W}})^{-1}),$ $\Lambda_{\mathbf{Z}} \sim \mathcal{W}(\mathbf{W}_0, v_0),$ $\mu_{\mathbf{Z}} \sim \mathcal{N}(\mu_0, (\beta_0 \Lambda_{\mathbf{Z}})^{-1}).$	Gibbs sampling
Bayesian model for L1-norm low-rank matrix factorizations [21]	$E_{ij} \sim \mathcal{N}(0, T_{ij}),$ $\mathbf{w}_i \sim \mathcal{N}(\mathbf{0}, \tau_{\mathbf{W}} \mathbf{I}_r),$ $\mathbf{z}_j \sim \mathcal{N}(\mathbf{0}, \tau_{\mathbf{Z}} \mathbf{I}_r).$	$T_{ij} \sim \text{Exp}(\lambda),$ $\tau_{\mathbf{W}} \sim \text{Gam}(a_0, b_0),$ $\tau_{\mathbf{Z}} \sim \text{Gam}(c_0, d_0).$	VB

### 6. Probabilistic Models of Robust PCA

Compared with traditional PCA, robust PCA is more robust to outliers or large sparse noise. Stable robust PCA [95], a stable version of robust PCA, decomposes the data matrix into the sum of a low-rank matrix, a sparse noise matrix and a dense noise matrix. The low-rank matrix obtained by solving a relaxed principal component pursuit is simultaneously stable to small noise and robust to gross sparse errors. This section will review probabilistic models of robust PCA.

#### 6.1. Bayesian Robust PCA

In [22], a stable robust PCA is modeled as:

$$\mathbf{X} = \mathbf{W}(\mathbf{D}\mathbf{\Lambda})\mathbf{Z} + \mathbf{B} * \mathbf{S} + \mathbf{E} \tag{101}$$

where  $\mathbf{D} = \text{diag}(d_{11}, d_{22}, \dots, d_{rr}), \mathbf{\Lambda} = \text{diag}(\lambda_{11}, \lambda_{22}, \dots, \lambda_{rr}), \mathbf{B} \in \{0, 1\}^{d \times N}$ . The three terms  $\mathbf{W}(\mathbf{D}\mathbf{\Lambda})\mathbf{Z}, \mathbf{B} * \mathbf{S}$  and  $\mathbf{E}$  are the low-rank, the sparse noise and the dense noise terms respectively. If we do not consider the sparse noise term  $\mathbf{B} * \mathbf{S}$ , then Equation (101) is equivalent to Equation (59). If all columns of  $\mathbf{B} * \mathbf{S}$  are same, then the stable robust PCA becomes to be the PCA model (31).

The following considers the probability distributions of all matrices in the right of Equation (101). We assume that  $\mathbf{w}_{.i} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d/d)$ ,  $d_{ii} \sim \text{Bern}(p_i)$ ,  $\lambda_{ii} \sim \mathcal{N}(0, \tau^{-1})$ ,  $\mathbf{z}_{.j} \sim \mathcal{N}(0, \mathbf{I}_r/r)$ ,  $\mathbf{b}_{.j} \sim \prod_{k=1}^r \text{Bern}(\pi_k)$ ,  $\mathbf{s}_{.j} \sim \mathcal{N}(0, v^{-1}\mathbf{I}_d)$ ,  $E_{ij} \sim \mathcal{N}(0, \gamma^{-1})$ ,  $i = 1, 2, \dots, r, j = 1, 2, \dots, N$ . The priors of  $p_i$ ,  $\tau$ ,  $\pi_k$ ,  $v$  and  $\gamma$  are given by  $p_i \sim \text{Beta}(\alpha_0, \beta_0)$ ,  $\tau \sim \text{Gam}(a_0, b_0)$ ,  $\pi_k \sim \text{Beta}(\alpha_1, \beta_1)$ ,  $v \sim \text{Gam}(c_0, d_0)$  and  $\gamma \sim \text{Gam}(e_0, f_0)$  respectively.

Two methods were proposed in [22], that is, Gibbs sampling and VB inference. For the second method, the joint probability distribution is

$$\begin{aligned} p(\mathbf{X}, \mathbf{W}, \mathbf{Z}, \mathbf{D}, \mathbf{\Lambda}, \mathbf{B}, \mathbf{S}, \mathbf{p}, \tau, \boldsymbol{\pi}, v, \gamma) \\ = p(\mathbf{X}|\mathbf{W}, \mathbf{Z}, \mathbf{D}, \mathbf{\Lambda}, \mathbf{B}, \mathbf{S}, \boldsymbol{\pi}, v, \gamma) p(\mathbf{W}) p(\mathbf{Z}) p(\tau) p(\boldsymbol{\Lambda}|\tau) p(\mathbf{D}|\mathbf{p}) p(\mathbf{p}) \\ p(\boldsymbol{\pi}) p(\mathbf{B}|\boldsymbol{\pi}) p(v) p(\mathbf{S}|v) p(\gamma). \end{aligned} \tag{102}$$

VB inference is employed to approximating the posterior distribution  $p(\mathbf{W}, \mathbf{Z}, \mathbf{D}, \mathbf{\Lambda}, \mathbf{B}, \mathbf{S}, \boldsymbol{\pi}, v, \gamma|\mathbf{X})$ . Bayesian robust PCA is robust to different noise levels without changing model hyperparameters and exploits additional structure of the matrix in video applications.

### 6.2. Variational Bayes Approach to Robust PCA

In Equation (101), if both  $\mathbf{D}$  and  $\mathbf{\Lambda}$  are set to be identical matrices, then we have another form of stable robust PCA:

$$\mathbf{X} = \mathbf{WZ} + \mathbf{B} * \mathbf{S} + \mathbf{E}. \tag{103}$$

Essentially, both Equations (101) and (103) are equivalent. Formally, Equation (103) can also be transformed into another formula:

$$\mathbf{X} = \mathbf{WZ} + \mathbf{B} * \mathbf{S} + (1 - \mathbf{B}) * \mathbf{E}. \tag{104}$$

We assume that  $\mathbf{w}_{.i} \sim \mathcal{N}(\boldsymbol{\mu}_0, \sigma_0^2 \mathbf{I}_d)$ ,  $\mathbf{z}_{.j} \sim \mathcal{N}(\mathbf{v}_0, \sigma_0^2 \mathbf{I}_N)$ ,  $b_{ij} \sim \text{Bern}(b_0)$ . Let the means of  $S_{ij}$  and  $E_{ij}$  be zeros and their precision be  $\tau_S$  and  $\tau_E$  respectively. The priors of  $\tau_S$  and  $\tau_E$  are given by

$$\tau_S \sim \text{Gam}(\alpha_S, \beta_S), \tau_E \sim \text{Gam}(\alpha_E, \beta_E). \tag{105}$$

A naïve VB approach was proposed in [23]. The trial distribution is stipulated as

$$q(\mathbf{W}, \mathbf{Z}, \mathbf{B}, \tau_S, \tau_E) = q(\mathbf{W}) q(\mathbf{Z}) q(\mathbf{B}) q(\tau_S) q(\tau_E). \tag{106}$$

The likelihood function is constructed as

$$L = p(\mathbf{X}|\mathbf{W}, \mathbf{Z}, \mathbf{B}, \tau_S, \tau_E) = \prod_{i,j} \mathcal{N}(X_{ij} - \mathbf{w}_{.i} \cdot \mathbf{z}_{.j} | \tau_S)^{B_{ij}} \mathcal{N}(X_{ij} - \mathbf{w}_{.i} \cdot \mathbf{z}_{.j} | \tau_E)^{1-B_{ij}}. \tag{107}$$

The prior distribution is represented by

$$\pi(\mathbf{W}, \mathbf{Z}, \mathbf{B}, \tau_S, \tau_E) = p(\mathbf{W}|\boldsymbol{\mu}_0, \sigma_0^2 \mathbf{I}_d) p(\mathbf{Z}|\mathbf{v}_0, \sigma_0^2 \mathbf{I}_N) p(\mathbf{B}) p(\tau_S) p(\tau_E). \tag{108}$$

We first construct a function:  $G(q) = \mathbb{E}_q[\log L] - D_{KL}(q||\pi)$ . To simplify this problem, we let

$$q(\mathbf{w}_{.i}) \sim \mathcal{N}(\boldsymbol{\mu}_{wi}, \boldsymbol{\Sigma}_{wi}), q(\mathbf{z}_{.i}) \sim \mathcal{N}(\boldsymbol{\mu}_{zi}, \boldsymbol{\Sigma}_{zi}). \tag{109}$$

To find the updates for  $\mathbf{W}, \mathbf{Z}$ , we can maximize the function  $G$  with respect to  $\boldsymbol{\mu}_{wi}, \boldsymbol{\Sigma}_{wi}, \boldsymbol{\mu}_{zi}, \boldsymbol{\Sigma}_{zi}$  respectively. The main advantage of the proposed approach is that it can incorporate additional prior information and cope with missing data.



### 6.3. Sparse Bayesian Robust PCA

In Equation (101), we replace  $\mathbf{B} * \mathbf{S}$  by  $\mathbf{S}$  for the sake of simplicity. Thus, we have a model of sparse Bayesian robust PCA [19]. Assume that  $\mathbf{w}_{\cdot k} \sim \mathcal{N}(\mathbf{0}, \gamma_k^{-1} \mathbf{I}_d)$ ,  $\mathbf{z}_{k \cdot} \sim \mathcal{N}(\mathbf{0}, \gamma_k^{-1} \mathbf{I}_N)$ ,  $S_{ij} \sim \mathcal{N}(0, \alpha_{ij}^{-1})$  and  $E_{ij} \sim \mathcal{N}(0, \beta^{-1})$ , where  $k = 1, 2, \dots, r, i = 1, 2, \dots, d, j = 1, 2, \dots, N$ . The priors of  $\gamma_k$  are given by  $\gamma_k \sim \text{Gam}(a, b)$ . What's more, we assign Jeffrey's priors to  $\alpha_{ij}$  and  $\beta$ :

$$p(\beta) = \beta^{-1}, p(\alpha_{ij}) = \alpha_{ij}^{-1}, i = 1, 2, \dots, d, j = 1, 2, \dots, N. \tag{110}$$

The joint distribution is expressed as

$$p(\mathbf{X}, \mathbf{W}, \mathbf{Z}, \mathbf{S}, \gamma, \alpha, \beta) = p(\mathbf{X}|\mathbf{W}, \mathbf{Z}, \mathbf{S}, \beta)p(\mathbf{W}|\gamma)p(\mathbf{Z}|\gamma)p(\mathbf{S}|\alpha)p(\gamma)p(\alpha)p(\beta). \tag{111}$$

VB inference was used to approximate the posterior distributions of all variables matrices [19]. Experimental results in video background/foreground separation show that the proposed method is more effective than MAP and Gibbs sampling.

## 7. Probabilistic Models of Non-Negative Matrix Factorization

Non-negative matrix factorization (NMF) decomposes a non-negative data matrix into the product of two non-negative low-rank matrices. Mathematically, we can formulate NMF as follows:

$$\mathbf{X} = \mathbf{WZ} + \mathbf{E} \tag{112}$$

where  $X_{ij}, W_{ik}, Z_{kj}$  are non-negative. Multiplicative algorithms [96] are often used to obtain the point estimations of both  $\mathbf{W}$  and  $\mathbf{Z}$ . This section will introduce probabilistic models of NMF.

### 7.1. Probabilistic Non-Negative Matrix Factorization

Equation (112) can be rewritten as  $\mathbf{X} \approx \mathbf{WZ} = \sum_{k=1}^r \mathbf{w}_{\cdot k} \mathbf{z}_{k \cdot}$ . Let  $\theta_k = \{\mathbf{w}_{\cdot k}, \mathbf{z}_{k \cdot}\}$  and  $\theta = \{\theta_1, \theta_2, \dots, \theta_r\}$ . Probabilistic non-negative matrix factorization [24] introduces a generative model:

$$X_{ij} = \sum_{k=1}^r C_{k,ij}, C_{k,ij} \sim p(C_{k,ij}|\theta_k). \tag{113}$$

The probability distributions of  $C_{k,ij}$  can be assumed to follow Gaussian or Poisson distributions because they are closed under summation. This assumption means that we can get easily the probability distributions of  $X_{ij}$ . Four algorithms were proposed in [24], that is, multiplicative, EM, Gibbs sampling and VB algorithms.

### 7.2. Bayesian Inference for Nonnegative Matrix Factorization

For arbitrary  $k \in \{1, 2, \dots, r\}$ , Bayesian non-negative matrix factorization [25] introduces variables  $S_k = \{S_{ikj} | i = 1, 2, \dots, d, j = 1, 2, \dots, N\}$  as latent sources. The hierarchical model of  $X_{ij}$  is given by

$$S_{ikj} \sim \text{Poiss}(W_{ik}Z_{kj}), X_{ij} = \sum_{k=1}^r S_{ikj}. \tag{114}$$

In view of the fact that a Gamma distribution is the conjugate prior to Poisson distribution, the hierarchical priors of  $W_{ik}$  and  $Z_{kj}$  are proposed as follows

$$W_{ik} \sim \text{Gam}(a_{ik}^{\mathbf{W}}, b_{ik}^{\mathbf{W}}), Z_{kj} \sim \text{Gam}(a_{ik}^{\mathbf{Z}}, b_{ik}^{\mathbf{Z}}). \tag{115}$$

Let  $\Theta = \{a_{ik}^{\mathbf{W}}, b_{ik}^{\mathbf{W}}, a_{ik}^{\mathbf{Z}}, b_{ik}^{\mathbf{Z}}\}$  and  $S = \{S_1, S_2, \dots, S_r\}$ . For given  $\Theta$  and  $\mathbf{X}$ , the posterior distribution is expressed as  $p(\mathbf{W}, \mathbf{Z}, S | \mathbf{X}, \Theta)$ . We assume the trial distribution of  $p(\mathbf{W}, \mathbf{Z}, S | \mathbf{X}, \Theta)$  is factorable:

$q(\mathbf{W}, \mathbf{Z}, S) = q(\mathbf{W})q(\mathbf{Z})q(S)$ . Both VB inference and Gibbs sampling were proposed to infer the probability distributions of all variables [25].

The above Bayesian nonnegative matrix factorization is not a matrix factorization approach to latent Dirichlet allocation. To this end, another Bayesian extension of the nonnegative matrix factorization algorithm was proposed in [27]. What’s more, Paisley et al. also provided a correlated nonnegative matrix factorization based on the correlated topic model [27]. The stochastic variational inference algorithms were presented to solve the proposed two models.

### 7.3. Bayesian Nonparametric Matrix Factorization

Gamma process nonnegative matrix factorization (GaP-NMF) was developed in [26]. This Bayesian nonparametric matrix factorization considers the case that the number of sources  $r$  is unknown. Let non-negative hidden variable  $\theta_k$  be the overall gain of the  $k$ -th source and  $L$  a large number of sources. We assume that

$$W_{ik} \sim \text{Gam}(a, a), Z_{kj} \sim \text{Gam}(b, b), X_{ij} \sim \text{Exp}\left(\sum_{k=1}^r \theta_k W_{ik} Z_{kj}\right), \theta_k \sim \text{Gam}(\alpha/L, \alpha c). \tag{116}$$

The posterior distribution is expressed as  $p(\mathbf{W}, \mathbf{Z}, \boldsymbol{\theta} | \mathbf{X}, a, b, \alpha, c)$  for given hyperparameters  $a, b, \alpha, c$ . The trial distribution of  $\boldsymbol{\theta}, \mathbf{W}, \mathbf{Z}$  is assumed to be factorable, that is,  $q(\mathbf{W}, \mathbf{Z}, \boldsymbol{\theta}) = q(\mathbf{W})q(\mathbf{Z})q(\boldsymbol{\theta})$ . The flexible generalized inverse-Gaussian distributions are imposed on  $W_{ik}, Z_{kj}$  and  $\theta_k$  respectively,

$$q(W_{ik}) \sim \text{GIG}(\gamma_{ik}^{\mathbf{W}}, \rho_{ik}^{\mathbf{W}}, \tau_{ik}^{\mathbf{W}}), q(Z_{kj}) \sim \text{GIG}(\gamma_{kj}^{\mathbf{Z}}, \rho_{kj}^{\mathbf{Z}}, \tau_{kj}^{\mathbf{Z}}), q(\theta_k) \sim \text{GIG}(\gamma_k^{\boldsymbol{\theta}}, \rho_k^{\boldsymbol{\theta}}, \tau_k^{\boldsymbol{\theta}}). \tag{117}$$

The lower bound of the marginal likelihood is computed as

$$\begin{aligned} \log p(\mathbf{X}, a, b, \alpha, c) &\geq \mathbb{E}_q[\log p(\mathbf{X} | \mathbf{W}, \mathbf{Z}, \boldsymbol{\theta})] \\ &+ \mathbb{E}_q[\log p(\mathbf{W} | a)] - \mathbb{E}_q[\log q(\mathbf{W})] \\ &+ \mathbb{E}_q[\log p(\mathbf{Z} | b)] - \mathbb{E}_q[\log q(\mathbf{Z})] \\ &+ \mathbb{E}_q[\log p(\boldsymbol{\theta} | \alpha, c)] - \mathbb{E}_q[\log q(\boldsymbol{\theta})] \end{aligned} \tag{118}$$

By maximizing the lower bound of  $\log p(\mathbf{X}, a, b, \alpha, c)$ , we can yield an approximation distribution of  $q(\mathbf{W}, \mathbf{Z}, \boldsymbol{\theta})$ . GaP-NMF is applied in recorded music and the number of latent sources is discovered automatically.

### 7.4. Beta-Gamma Non-Negative Matrix Factorization

For the moment, we do not factorize the data matrix  $\mathbf{X}$  into the product of two low-rank matrices. Different from previous probabilistic models of NMF, we assume that  $X_{ij}$  is generated from a Beta distribution:  $\text{Beta}(A_{ij}, B_{ij})$ . For two matrices  $\mathbf{A}$  and  $\mathbf{B}$ , we jointly factorize them as

$$\mathbf{A} \approx \mathbf{C}\mathbf{H}, \mathbf{B} \approx \mathbf{D}\mathbf{H} \tag{119}$$

where  $\mathbf{C} \in \mathbb{R}_+^{d \times r}, \mathbf{D} \in \mathbb{R}_+^{d \times r}, \mathbf{H} \in \mathbb{R}_+^{r \times N}$ .

In Beta-Gamma non-negative matrix factorization [28], the generative model is given by

$$X_{ij} \sim \text{Beta}\left(\sum_{k=1}^r C_{ik} H_{kj}, \sum_{k=1}^r D_{ik} H_{kj}\right), C_{ik} \sim \text{Gam}(\mu_0, \alpha_0), D_{ik} \sim \text{Gam}(v_0, \beta_0), H_{kj} \sim \text{Gam}(\rho_0, \zeta_0) \tag{120}$$

Variational inference framework was adopted and a new lower-bound was proposed to approximate the objective function to derive an analytically tractable approximate solution of the posterior distribution. Beta-Gamma non-negative matrix factorization is used in source separation, collaborative filtering and cancer epigenomics analysis.

## 8. Other Probabilistic Models of Low-Rank Matrix/Tensor Factorizations

Besides the probabilistic models discussed in foregoing sections, there are many other types of probabilistic low-rank matrix factorization models. A successful application of probabilistic low-rank matrix factorization is the collaborative filtering in recommendation systems. In collaborative filtering, there are several other Poisson models in which the observations are usually modeled with a Poisson distribution, and these models mainly include [97–105]. As a matter of fact, the Poisson factorization roots in the nonnegative matrix factorization and takes advantage of the sparse essence of user behavior data and scales [103]. For some probabilistic models with respect to collaborative filtering, the Poisson distribution is changed into other probability distributions and this change deals with logistic function [106–108], Heaviside step function [107,109], Gaussian cumulative density function [110] and so on. In addition, side information on the a low-dimensional latent presentations is integrated into probabilistic low-rank matrix factorization models [111–113], and the case that the data is missing not at random is taken into consideration [109,114,115].

It is worthy to pay attention to other applications of probabilistic low-rank matrix factorization models. For instance, [116] developed a probabilistic model for low-rank subspace clustering. In [88], a sparse additive matrix factorization was proposed by a Bayesian regularization effect and the corresponding model was applied into a foreground/background video separation problem.

Recently, probabilistic low-rank matrix factorizations have been extended into the case of tensor decompositions (factorizations). Tucker decomposition and CP decomposition are two popular tensor decomposition approaches. The probabilistic Tucker decomposition models mainly include probabilistic Tucker decomposition [34], exponential family tensor factorization [35] and InfTucker model [36]. Probabilistic Tucker decomposition was closely related to probabilistic PCA. In [35], an integration method was proposed to model heterogeneously attributed data tensors. InfTucker, a tensor-variate latent nonparametric Bayesian model, conducted Tucker decomposition in an infinite feature space.

More probabilistic models of tensor factorizations focus on CP tensor decomposition model. For example, Ermis and Cemgil investigated variational inference for probabilistic latent tensor factorization [37]. Based on hierarchical dirichlet process, a Bayesian probabilistic model for unsupervised tensor factorization was proposed [38]. In [39], a novel probabilistic tensor factorization was proposed by extending probabilistic matrix factorization. A probabilistic latent tensor factorization was proposed in [40] to address the task of link pattern prediction. Based on the Polya-Gamma augmentation strategy and online expectation maximization algorithm, [41] proposed a scalable probabilistic tensor factorization framework. As the generalization of Poisson matrix factorization, Poisson tensor factorization was presented in [42]. In [43], a Bayesian tensor factorization models was proposed to infer the latent group structures from dynamic pairwise interaction patterns. In [44], a Bayesian non-negative tensor factorization model was presented for count-valued tensor data and scalable inference algorithms were developed. A scalable Bayesian framework for low-rank CP decomposition was presented and it can analyses both continuous and binary datasets [45]. A zero-truncated Poisson tensor factorization for binary tensors was proposed in [46]. A Bayesian robust tensor factorization [47] was proposed and it is the extension of probabilistic stable robust PCA. And in [48], the CP factorization was formulated by a hierarchical probabilistic model.

## 9. Conclusions and Future Work

In this paper, we have made a survey on probabilistic models of low-rank matrix factorizations and the related works. To classify the main probabilistic models, we divide low-rank matrix factorizations into several groups such as PCA, matrix factorizations, robust PCA, non-negative matrix factorization and so on. For each category, we list representative probabilistic models, describe the probability distributions of all random matrices or latent variables, present the corresponding inference methods and compare their similarity and difference. Besides, we further provide an overview of probabilistic models of low-rank tensor factorizations and discuss other probabilistic matrix factorizations models.

Although probabilistic low-rank matrix/tensor factorizations have made some progresses, we still face some challenges in theories and applications. Future research may concern the following aspects:

- Scalable algorithms to infer the probability distributions and parameters. Although both Gibbs sampling and variational Bayesian inference have their own advantages, they need large computation cost for real large-scale problems. A promising future direction is to design scalable algorithms.
- Constructing new probabilistic models of low-rank matrix factorizations. It is necessary to develop other probabilistic models according to the actual situation. For example, we can consider different types of sparse noise and different probability distributions (including the prior distributions) of low-rank components or latent variables.
- Probabilistic models of non-negative tensor factorizations. There is not much research on this type of probabilistic models. Compared with probabilistic models of tensor factorizations, the probabilistic non-negative tensor factorizations models are more complex and difficult in inferring the posterior distributions.
- Probabilistic TT format. In contrast to both CP and Tucker decompositions, the TT format provides stable representations and is formally free from the curse of dimensionality. Hence, probabilistic model of the TT format would be an interesting research issue.

**Acknowledgments:** This work is partially supported by the National Natural Science Foundation of China (No. 61403298, No. 11401457), China Postdoctoral Science Foundation (No. 2017M613087) and the Scientific Research Program Funded by the Shaanxi Provincial Education Department (No. 16JK1435).

**Author Contributions:** Jiarong Shi reviewed probabilistic models of low-rank matrix factorizations and wrote the manuscript. Xiuyun Zheng wrote the preliminaries. Wei Yang presented literature search and wrote other probabilistic models of low-rank matrix factorizations. All authors were involved in organizing and refining the manuscript. All authors have read and approved the final manuscript.

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

## References

1. Jolliffe, I. *Principal Component Analysis*; John Wiley & Sons, Ltd.: Mississauga, ON, Canada, 2002.
2. Golub, G.H.; Reinsch, C. Singular value decomposition and least squares solutions. *Numer. Math.* **1970**, *14*, 403–420. [[CrossRef](#)]
3. Ke, Q.; Kanade, T. Robust L1 norm factorizations in the presence of outliers and missing data by alternative convex programming. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE Computer Society: Washington, DC, USA, 2005.
4. Kwak, N. Principal component analysis based on L1-norm maximization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1672–1680. [[CrossRef](#)] [[PubMed](#)]
5. Nie, F.; Huang, H. Non-greedy L21-norm maximization for principal component analysis. *arXiv* **2016**, arXiv:1603.08293.
6. Candès, E.J.; Li, X.; Ma, Y.; Wright, J. Robust principal component analysis? *J. ACM* **2011**, *58*, 11. [[CrossRef](#)]
7. Lee, D.D.; Seung, H.S. Learning the parts of objects by non-negative matrix factorizations. *Nature* **1999**, *401*, 788–791. [[PubMed](#)]
8. Kolda, T.G.; Bader, B.W. Tensor decompositions and applications. *SIAM Rev.* **2009**, *51*, 455–500. [[CrossRef](#)]
9. Candès, E.J.; Recht, B. Exact matrix completion via convex optimization. *Found. Comput. Math.* **2009**, *9*, 717–772. [[CrossRef](#)]
10. Tipping, M.E.; Bishop, C.M. Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B* **1999**, *21*, 611–622. [[CrossRef](#)]
11. Bishop, C.M. Variational principal components. In *Proceedings of the Ninth International Conference on Artificial Neural Networks*, Edinburgh, UK, 7–10 September 1999.
12. Gao, J. Robust L1 principal component analysis and its Bayesian variational inference. *Neural Comput.* **2008**, *20*, 55–572. [[CrossRef](#)] [[PubMed](#)]

13. Luttinen, J.; Ilin, A.; Karhunen, J. Bayesian robust PCA of incomplete data. *Neural Process. Lett.* **2012**, *36*, 189–202. [[CrossRef](#)]
14. Lim, Y.J.; Teh, Y.W. Variational Bayesian approach to movie rating prediction. In Proceedings of the KDD Cup and Workshop, San Jose, CA, USA, 12 August 2007.
15. Salakhutdinov, R.; Mnih, A. Probabilistic matrix factorization. In Proceedings of the 20th Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 3–6 December 2007.
16. Salakhutdinov, R.; Mnih, A. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008.
17. Lakshminarayanan, B.; Bouchard, G.; Archambeau, C. Robust Bayesian matrix factorization. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Lauderdale, FL, USA, 11–13 April 2011.
18. Wang, N.; Yao, T.; Wang, J.; Yeung, D.-Y. A probabilistic approach to robust matrix factorization. In Proceedings of the 12th European Conference on Computer Vision (ECCV), Florence, Italy, 7–13 October 2012.
19. Babacan, S.D.; Luessi, M.; Molina, R.; Katsaggelos, K. Sparse Bayesian methods for low-rank matrix estimation. *IEEE Trans. Signal Process.* **2012**, *60*, 3964–3977. [[CrossRef](#)]
20. Wang, N.; Yeung, D.-Y. Bayesian robust matrix factorization for image and video processing. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013.
21. Zhao, Q.; Meng, D.; Xu, Z.; Zuo, W.; Yan, Y. L1-norm low-rank matrix factorization by variational Bayesian method. *IEEE Trans. Neural Netw. Learn. Syst.* **2015**, *26*, 825–839. [[CrossRef](#)] [[PubMed](#)]
22. Ding, X.; He, L.; Carin, L. Bayesian robust principal component analysis. *IEEE Trans. Image Process.* **2011**, *20*, 3419–3430. [[CrossRef](#)] [[PubMed](#)]
23. Aicher, C. A variational Bayes approach to robust principal component analysis. In *SFI REU 2013 Report*; University of Colorado Boulder: Boulder, CO, USA, 2013.
24. Févotte, C.; Cemgil, A.T. Nonnegative matrix factorizations as probabilistic inference in composite models. In Proceedings of the IEEE 17th European Conference on Signal Processing, Glasgow, UK, 24–28 August 2009.
25. Cemgil, A.T. Bayesian inference for nonnegative matrix factorization models. *Comput. Intell. Neurosci.* **2009**, *2009*, 785152. [[CrossRef](#)] [[PubMed](#)]
26. Hoffman, M.; Cook, P.R.; Blei, D.M. Bayesian nonparametric matrix factorization for recorded music. In Proceedings of the International Conference on Machine Learning, Washington, DC, USA, 12–14 December 2010.
27. Paisley, J.; Blei, D.; Jordan, M. Bayesian nonnegative matrix factorization with stochastic variational inference. In *Handbook of Mixed Membership Models and Their Applications*, Chapman and Hall; CRC: Boca Raton, FL, USA, 2014.
28. Ma, Z.; Teschendorff, A.E.; Leijon, A.; Qiao, Y.; Zhang, H.; Guo, J. Variational Bayesian matrix factorizations for bounded support data. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 876–889. [[CrossRef](#)] [[PubMed](#)]
29. Hackbusch, W.; Kühn, S. A new scheme for the tensor representation. *J. Fourier Anal. Appl.* **2009**, *15*, 706–722. [[CrossRef](#)]
30. Grasedyck, L.; Hackbusch, W. An introduction to hierarchical (H-) rank and TT-rank of tensors with examples. *Comput. Methods Appl. Math.* **2011**, *11*, 291–304. [[CrossRef](#)]
31. Oseledets, I.V.; Tyrtshnikov, E.E. Breaking the curse of dimensionality, or how to use SVD in many dimensions. *Soc. Ind. Appl. Math.* **2009**, *31*, 3744–3759. [[CrossRef](#)]
32. Oseledets, I.V. Tensor-train decomposition. *SIAM J. Sci. Comput.* **2011**, *33*, 2295–2317. [[CrossRef](#)]
33. Holtz, S.; Rohwedder, T.; Schneider, R. The alternating linear scheme for tensor optimization in the tensor train format. *SIAM J. Sci. Comput.* **2012**, *34*, A683–A713. [[CrossRef](#)]
34. Chu, W.; Ghahramani, Z. Probabilistic models for incomplete multi-dimensional arrays. In Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS), Clearwater, FL, USA, 16–18 April 2009.
35. Hayashi, K.; Takenouchi, T.; Shibata, T.; Kamiya, Y.; Kato, D.; Kunieda, K.; Yamada, K.; Ikeda, K. Exponential family tensor factorizations for missing-values prediction and anomaly detection. In Proceedings of the IEEE 10th International Conference on Data Mining, Sydney, Australia, 13–17 December 2010.
36. Xu, Z.; Yan, F.; Qi, Y. Bayesian nonparametric models for multiway data analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 475–487. [[CrossRef](#)] [[PubMed](#)]



37. Ermis, B.; Cemgil, A. A Bayesian tensor factorization model via variational inference for link prediction. *arXiv* **2014**, arXiv:1409.8276.
38. Porteous, I.; Bart, E.; Welling, M. Multi-HDP: A nonparametric Bayesian model for tensor factorization. In Proceedings of the National Conference on Artificial Intelligence, Chicago, IL, USA, 13–17 July 2008.
39. Xiong, L.; Chen, X.; Huang, T.; Schneiderand, J.G.; Carbonell, J.G. Temporal collaborative filtering with Bayesian probabilistic tensor factorization. In Proceedings of the SIAM Data Mining, Columbus, OH, Canada, 29 April–1 May 2010.
40. Gao, S.; Denoyer, L.; Gallinari, P.; Guo, J. Probabilistic latent tensor factorizations model for link pattern prediction in multi-relational networks. *J China Univ. Posts Telecommun.* **2012**, *19*, 172–181. [[CrossRef](#)]
41. Rai, P.; Wang, Y.; Guo, S.; Chen, G.; Dunson, D.; Carin, L. Scalable Bayesian low-rank decomposition of incomplete multiway tensors. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014.
42. Schein, A.; Paisley, J.; Blei, D.M.; Wallach, H. Inferring polyadic events with Poisson tensor factorization. In Proceedings of the NIPS 2014 Workshop on Networks: From Graphs to Rich Data, Montreal, QC, Canada, 13 December 2014.
43. Schein, A.; Paisley, J.; Blei, D.M. Bayesian Poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Sydney, Australia, 10–13 August 2015.
44. Hu, C.; Rai, P.; Chen, C.; Harding, M.; Carin, L. Scalable Bayesian non-negative tensor factorization for massive count data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer International Publishing AG: Cham, Switzerland, 2015.
45. Rai, P.; Hu, C.; Harding, M.; Carin, L. Scalable probabilistic tensor factorization for binary and count data. In Proceedings of the 24th International Conference on Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015.
46. Hu, C.; Rai, P.; Carin, L. Zero-truncated Poisson tensor factorization for massive binary tensors. *arXiv* **2015**, arXiv:1508.04210.
47. Zhao, Q.; Zhou, G.; Zhang, L.; Cichocki, A.; Amari, S.I. Bayesian robust tensor factorization for incomplete multiway data. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *27*, 736–748. [[CrossRef](#)] [[PubMed](#)]
48. Zhao, Q.; Zhang, L.; Cichocki, A. Bayesian CP factorization of incomplete tensors with automatic rank determination. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *37*, 1751–1763. [[CrossRef](#)] [[PubMed](#)]
49. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. Ser. B* **1977**, *39*, 1–38.
50. Jeff Wu, C.F. On the convergence properties of the EM algorithm. *Ann. Stat.* **1983**, *11*, 95–103.
51. Xu, L.; Jordan, M.I. On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Comput.* **1995**, *8*, 129–151. [[CrossRef](#)]
52. Hunter, D.R.; Lange, K. A tutorial on MM algorithms. *Am. Stat.* **2004**, *58*, 30–37. [[CrossRef](#)]
53. Gelman, A.; Carlin, J.; Stern, H.S.; Dunson, D.; Vehtari, A.; Rubin, D. *Bayesian Data Analysis*; CRC Press: Boca Raton, FL, USA, 2014.
54. Bishop, C. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.
55. Geman, S.; Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **1984**, *6*, 721–741. [[CrossRef](#)] [[PubMed](#)]
56. Casella, G.; George, E.I. Explaining the Gibbs sampler. *Am. Stat.* **1992**, *46*, 167–174.
57. Gilks, W.R.; Wild, P. Adaptive rejection sampling for Gibbs sampling. *J. R. Stat. Soc. Ser. C Appl. Stat.* **1992**, *41*, 337–348. [[CrossRef](#)]
58. Liu, J.S. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Am. Stat. Assoc.* **1994**, *89*, 958–966. [[CrossRef](#)]
59. Gilks, W.R.; Best, N.G.; Tan, K.K.C. Adaptive rejection metropolis sampling within Gibbs sampling. *J. R. Stat. Soc. Ser. C Appl. Stat.* **1995**, *44*, 455–472. [[CrossRef](#)]
60. Martino, L.; Míguez, J. A generalization of the adaptive rejection sampling algorithm. *Stat. Comput.* **2010**, *21*, 633–647. [[CrossRef](#)]
61. Martino, L.; Read, J.; Luengo, D. Independent doubly adaptive rejection metropolis sampling within Gibbs sampling. *IEEE Trans. Signal Process.* **2015**, *63*, 3123–3138. [[CrossRef](#)]
62. Bernardo, J.M.; Smith, A.F.M. *Bayesian Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2001.

63. Jordan, M.I.; Ghahramani, Z.; Jaakkola, T.; Saul, L.K. An introduction to variational methods for graphical models. *Mach. Learn.* **1999**, *37*, 183–233. [[CrossRef](#)]
64. Attias, H. A variational Bayesian framework for graphical models. *Adv. Neural Inf. Process. Syst.* **2000**, *12*, 209–215.
65. Beal, M.J. *Variational Algorithms for Approximate Bayesian Inference*; University of London: London, UK, 2003.
66. Smídl, V.; Quinn, A. *The Variational Bayes Method in Signal Processing*; Springer: New York, NY, USA, 2005.
67. Blei, D.; Jordan, M. Variational inference for Dirichlet process mixtures. *Bayesian Anal.* **2006**, *1*, 121–144. [[CrossRef](#)]
68. Beal, M.J.; Ghahramani, Z. The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures. In Proceedings of the IEEE International Conference on Acoustics Speech & Signal Processing, Toulouse, France, 14–19 May 2006.
69. Schölkopf, B.; Platt, J.; Hofmann, T. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2007.
70. Tzikas, D.G.; Likas, A.C.; Galatsanos, N.P. The variational approximation for Bayesian inference. *IEEE Signal Process. Mag.* **2008**, *25*, 131–146. [[CrossRef](#)]
71. Chen, Z.; Babacan, S.D.; Molina, R.; Katsaggelos, A.K. Variational Bayesian methods for multimedia problems. *IEEE Trans. Multimed.* **2014**, *16*, 1000–1017. [[CrossRef](#)]
72. Fink, D. A Compendium of Conjugate Priors. Available online: <https://www.johndcook.com/CompendiumOfConjugatePriors.pdf> (accessed on 17 August 2017).
73. Pearl, J. Evidential reasoning using stochastic simulation. *Artif. Intell.* **1987**, *32*, 245–257. [[CrossRef](#)]
74. Tierney, L. Markov chains for exploring posterior distributions. *Ann. Stat.* **1994**, *22*, 1701–1762. [[CrossRef](#)]
75. Besag, J.; Green, P.J.; Hidgon, D.; Mengersen, K. Bayesian computation and stochastic systems. *Stat. Sci.* **1995**, *10*, 3–66. [[CrossRef](#)]
76. Gilks, W.R.; Richardson, S.; Spiegelhalter, D.J. *Markov Chain Monte Carlo in Practice*; Chapman and Hall: Suffolk, UK, 1996.
77. Brooks, S.P. Markov chain Monte Carlo method and its application. *J. R. Stat. Soc. Ser. D Stat.* **1998**, *47*, 69–100. [[CrossRef](#)]
78. Beichl, I.; Sullivan, F. The Metropolis algorithm. *Comput. Sci. Eng.* **2000**, *2*, 65–69. [[CrossRef](#)]
79. Liu, J.S. *Monte Carlo Strategies in Scientific Computing*; Springer: Berlin, Germany, 2001.
80. Andrieu, C.; Freitas, N.D.; Doucet, A.; Jordan, M.I. An Introduction to MCMC for machine learning. *Mach. Learn.* **2003**, *50*, 5–43. [[CrossRef](#)]
81. Von der Linden, W.; Dose, V.; von Toussaint, U. *Bayesian Probability Theory: Applications in the Physical Sciences*; Cambridge University Press: Cambridge, UK, 2014.
82. Liu, J.S.; Liang, F.; Wong, W.H. The Multiple-try method and local optimization in Metropolis sampling. *J. Am. Stat. Assoc.* **2000**, *95*, 121–134. [[CrossRef](#)]
83. Martino, L.; Read, J. On the flexibility of the design of multiple try Metropolis schemes. *Comput. Stat.* **2013**, *28*, 2797–2823. [[CrossRef](#)]
84. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [[CrossRef](#)]
85. Kullback, S. *Information Theory and Statistics*; John Wiley & Sons: Hoboken, NJ, USA, 1959.
86. Johnson, D.H.; Sinanovic, S. Symmetrizing the Kullback–Leibler distance. *IEEE Trans. Inf. Theory* **2001**, *9*, 96–99.
87. Erven, T.V.; Harremo, P.H. Rényi divergence and Kullback–Leibler divergence. *IEEE Trans. Inf. Theory* **2013**, *60*, 3797–3820. [[CrossRef](#)]
88. Nakajima, S.; Sugiyama, M.; Babacan, S.D. Variational Bayesian sparse additive matrix factorization. *Mach. Learn.* **2013**, *92*, 319–347. [[CrossRef](#)]
89. Nakajima, S.; Sugiyama, M.; Babacan, S.D. Global analytic solution of fully-observed variational Bayesian matrix factorization. *J. Mach. Learn. Res.* **2013**, *14*, 1–37.
90. Paisley, J.; Blei, D.; Jordan, M. Variational Bayesian inference with stochastic search. *arXiv* **2012**, arXiv:1206.6430.
91. Hoffman, M.; Blei, D.; Wang, C.; Paisley, J. Stochastic variational inference. *J. Mach. Learn. Res.* **2013**, *14*, 1303–1347.
92. Tipping, M.E.; Bishop, C.M. Mixtures of probabilistic principal component analyzers. *Neural Comput.* **1999**, *11*, 443–482. [[CrossRef](#)] [[PubMed](#)]

93. Khan, M.E.; Young, J.K.; Matthias, S. Scalable collaborative Bayesian preference learning. In Proceedings of the 17th International Conference on Artificial Intelligence and Statistics, Reykjavik, Iceland, 22–24 April 2014.
94. Srebro, N.; Tommi, J. Weighted low-rank approximations. In Proceedings of the International Conference on Machine Learning, Washington, DC, USA, 21–24 August 2003.
95. Zhou, Z.; Li, X.; Wright, J.; Candès, E.J.; Ma, Y. Stable principal component pursuit. In Proceedings of the IEEE International Symposium on Information Theory, Austin, TX, USA, 13–18 June 2010.
96. Lee, D.D.; Seung, H.S. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2011.
97. Ma, H.; Liu, C.; King, I.; Lyu, M.R. Probabilistic factor models for web site recommendation. In Proceedings of the 34th international ACM SIGIR Conference on Research and Development in Information Retrieval, Beijing, China, 24–28 July 2011.
98. Seeger, M.; Bouchard, G. Fast variational Bayesian inference for non-conjugate matrix factorization models. *J. Mach. Learn. Res. Proc. Track* **2012**, *22*, 1012–1018.
99. Hoffman, M. Poisson-uniform nonnegative matrix factorization. In Proceedings of the Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012.
100. Gopalan, P.; Hofman, J.M.; Blei, D.M. Scalable recommendation with Poisson factorization. *arXiv* **2014**, arXiv:1311.1704.
101. Gopalan, P.; Ruiz, F.; Ranganath, R.; Blei, D. Bayesian nonparametric Poisson factorization for recommendation systems. In Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, Reykjavik, Iceland, 22–25 April 2014.
102. Gopalan, P.; Charlin, L.; Blei, D. Content-based recommendations with Poisson factorization. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2014.
103. Gopalan, P.; Ruiz, F.; Ranganath, R.; Blei, D. Bayesian nonparametric Poisson factorization for recommendation systems. *J. Mach. Learn. Res.* **2014**, *33*, 275–283.
104. Gopalan, P.; Hofman, J.M.; Blei, D. Scalable Recommendation with hierarchical Poisson factorization. In Proceedings of the Conference on Uncertainty in Artificial Intelligence, Amsterdam, The Netherlands, 6–10 July 2015.
105. Gopalan, P.; Hofman, J.; Blei, D. Scalable recommendation with Poisson factorization. In Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, Amsterdam, The Netherlands, 6–10 July 2015.
106. Ma, H.; Yang, H.; Lyu, M.R.; King, I. SoRec: Social recommendation using probabilistic matrix factorization. In Proceedings of the 17th ACM Conference on Information and Knowledge Management, Napa Valley, CA, USA, 26–30 October 2008.
107. Paquet, U.; Koenigstein, N. One-class collaborative filtering with random graphs. In Proceedings of the 22nd International Conference on World Wide Web, ACM, Rio de Janeiro, Brazil, 13–17 May 2013.
108. Liu, Y.; Wu, M.; Miao, C. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS Comput. Biol.* **2016**, *12*, e1004760. [[CrossRef](#)] [[PubMed](#)]
109. Hernández-Lobato, J.M.; Houlisby, N.; Ghahramani, Z. Probabilistic matrix factorization with non-random missing data. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014.
110. Koenigstein, N.; Nice, N.; Paquet, U.; Schleyen, N. The Xbox recommender system. In Proceedings of the Sixth ACM Conference on Recommender Systems, Dublin, Ireland, 9–13 September 2012.
111. Shan, H.; Banerjee, A. Generalized probabilistic matrix factorizations for collaborative filtering. In Proceedings of the Data Mining (ICDM), Sydney, Australia, 13–17 December 2010.
112. Zhou, T.; Shan, H.; Banerjee, A.; Sapiro, G. Kernelized probabilistic matrix factorization: Exploiting graphs and side information. In Proceedings of the 2012 SIAM International Conference on Data Mining, Anaheim, CA, USA, 26–28 April 2012.
113. Gonen, M.; Suleiman, K.; Samuel, K. Kernelized Bayesian matrix factorization. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013.
114. Marlin, B.M.; Zemel, R.S. Collaborative prediction and ranking with non-random missing data. In Proceedings of the Third ACM Conference on Recommender Systems, New York, NY, USA, 23–25 October 2009.



115. Bolgár, B.; Péter, A. Bayesian matrix factorization with non-random missing data using informative Gaussian process priors and soft evidences. In Proceedings of the Eighth International Conference on Probabilistic Graphical Models, Lugano, Switzerland, 6–9 September 2016.
116. Babacan, S.D.; Nakajima, S.; Do, M. Probabilistic low-rank subspace clustering. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2012.



© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).