Information Sciences 405 (2017) 175–189

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Link sign prediction by Variational Bayesian Probabilistic Matrix Factorization with Student-t Prior☆

Yisen Wang[a], Fangbing Liu[a], Shu-Tao Xia[a,b], Jia Wu[c,*]

[a] Department of Computer Science and Technology, Tsinghua University, China
[b] Graduate School at Shenzhen, Tsinghua University, China
[c] Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia

ARTICLE INFO

Article history:
Received 7 April 2016
Revised 9 April 2017
Accepted 11 April 2017
Available online 12 April 2017

Keywords:
Signed networks
Matrix factorization
Student-t distribution

ABSTRACT

In signed social networks, link sign prediction refers to using the observed link signs to infer the signs of the remaining links, which is important for mining and analyzing the evolution of social networks. The widely used matrix factorization-based approach – Bayesian Probabilistic Matrix Factorization (BMF), assumes that the noise between the real and predicted entry is Gaussian noise, and the prior of latent features is multivariate Gaussian distribution. However, Gaussian noise model is sensitive to outliers and is not robust. Gaussian prior model neglects the differences between latent features, that is, it does not distinguish between important and non-important features. Thus, Gaussian assumption based models perform poorly on real-world (sparse) datasets. To address these issues, a novel Variational Bayesian Probabilistic Matrix Factorization with Student-t prior model (TBMF) is proposed in this paper. A univariate Student-t distribution is used to fit the prediction noise, and a multivariate Student-t distribution is adopted for the prior of latent features. Due to the high kurtosis of Student-t distribution, TBMF can select informative latent features automatically, characterize long-tail cases and obtain reasonable representations on many real-world datasets. Experimental results show that TBMF improves the prediction performance significantly compared with the state-of-the-art algorithms, especially when the observed links are few.

1. Introduction

Signed social networks have increasingly become an actively researched field in social networks [6,44–46,48,49]. Each link of the signed network has a sign, expressing the attitude from the generator to the receiver, which can be positive (stating trust or approval) or negative (representing distrust or disapproval). Link sign prediction is defined as using the observed link signs to predict the unknown signs of the remaining links [15]. For an online retailer, *e.g.*, Amazon, the link sign can represent whether the user likes the item or not, so link sign prediction is significant for recommending items to users. It is the same as the video website, *e.g.*, Netflix and Hulu. The relations between users can also be more informative

☆ The first two authors contributed equally to this study. This research is supported in part by the National Natural Science Foundation of China (61371078), and the R&D Program of Shenzhen (Nos. JCYJ20140509172959977, JSGG20150512162853495, ZDSYS20140509172959989).

* Corresponding author.
E-mail addresses: wangys14@mails.tsinghua.edu.cn (Y. Wang), lfb13@mails.tsinghua.edu.cn (F. Liu), xiast@sz.tsinghua.edu.cn (S.-T. Xia), jia.wu@uts.edu.au (J. Wu).

http://dx.doi.org/10.1016/j.ins.2017.04.014
0020-0255/

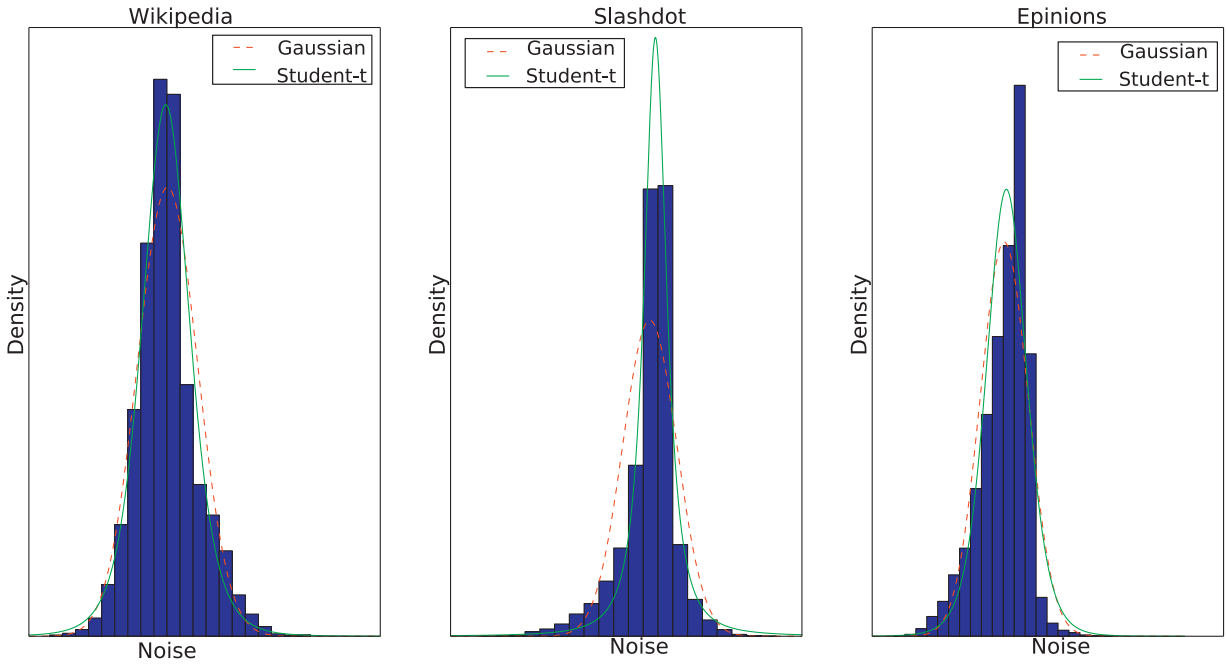

Fig. 1. Noise fitting results of different distributions. Compared to Gaussian distribution, Student-t distribution has higher kurtosis and owns better representation ability for noise. The figure is best viewed in color.

by allowing users to label each other as friends or enemies, compared to traditional social networks which only contain friend relations. Generally, there are two kinds of methods for solving link sign prediction, *i.e.*, the node similarity-based method [13,36] and the matrix factorization-based method [1,12]. The former not only relies on the completeness of the network structure, but also depends on high level features derived from social psychology theory [7,15,16,40]. The matrix factorization-based method overcomes the incompleteness of the network, and eliminates the need to extract structure features from limited observed links [4,26], which will be the focus of this work.

A signed social network is modeled in a matrix format $S \in \Re^{N \times M}$ with N users and M items. Each entry s_{ij} is defined as the attitude from the generator i to the receiver j, where $s_{ij} = 1$ for positive attitude, $s_{ij} = -1$ for negative attitude and $s_{ij} = 0$ for unknown cases. That is to say, s_{ij} is nonzero if the link sign is observed and zero otherwise. In matrix factorization models, the sign matrix $S \in \Re^{N \times M}$ can be factorized into the user latent feature matrix $U \in \Re^{K \times N}$ and the item latent feature matrix $V \in \Re^{K \times M}$, where $K \ll N, M$. The goal is to find appropriate U and V to minimize the discrepancy between S and $U^T V$, therefore, the noise is defined as the absolute value of $s_{ij} - u_i^T v_j$. A number of algorithms have been proposed to solve this problem, such as Variational Bayesian Matrix Factorization (VBMF) [17], Probabilistic Matrix Factorization (PMF) [30], Bayesian Probabilistic Matrix Factorization (BMF) [29] and Robust Bayesian Matrix Factorization (RBMF) [14]. We find that they all assume that the noise is Gaussian noise, while in real-world datasets, Gaussian distribution is not robust/reasonable and cannot characterize long-tail cases, which can be illustrated by the following experiments. On three widely used datasets (*i.e., Wikipedia, Slashdot* and *Epinions*) in the signed social network literature [16], we plot the histogram of the noise $|s_{ij} - u_i^T v_j|$ fitted by different distributions. Fig. 1 shows that the solid curves have higher peaks and heavier tails than the dashed curves, meaning that the solid curves can better fit the noise on real-world datasets. In other words, Student-t distribution is more suitable and more robust in fitting the noise than Gaussian distribution [37].

The fraction of the observed links with regard to the above three datasets is less than 1% (refer to Section 6.1), which is very sparse. As for the prior assumption of latent features, VBMF, PMF and BMF all adopt the multivariate Gaussian prior, which does not distinguish the importance of the latent features, making them perform poorly on the sparse datasets. Moreover, a study of the posteriors obtained by BMF on standard movie recommendation benchmarks also shows that the tails are significantly heavier than Gaussian distribution [14]. Thus, heavy-tailed distribution should be considered for the prior of latent features. To this end, RBMF has tried to assume that the prior of latent features is multivariate Student-t distribution, which obtains significant performance improvement. However, RBMF fixes the mean of Student-t distribution as 0, and still assumes that the noise is Gaussian noise.

To address the above issues, we propose a Variational Bayesian Probabilistic Matrix Factorization with Student-t prior model (TBMF), using Student-t distribution to describe both the noise and the prior of latent features. A univariate Student-t distribution is used to fit the prediction noise, and a multivariate Student-t distribution is adopted for the prior of latent features. A variational Bayesian inference is employed to estimate the parameters and find the solution that maximizes the posterior probability. In summary, our contributions are as follows:

- TBMF overcomes the shortcomings of the state-of-the-art algorithms. It is more robust to outliers and can deal with long-tail cases. TBMF also performs well on sparse datasets due to its heavy-tailed characteristic.
- Student-t distribution is used to characterize the noise model and the prior model of latent features.
- A variational inference algorithm is derived to estimate the parameters and solve our Student-t distribution-based model.
- Experiments on three signed network datasets, *i.e., Wikipedia, Slashdot* and *Epinions*, show that TBMF achieves better prediction performance than the state-of-the-art algorithms.

The rest of the paper is organized as follows: Section 2 presents related work on the link sign prediction problem. Section 3 describes the preliminaries of BMF and RBMF. Section 4 states our proposed TBMF. Section 5 demonstrates the details of the variational Bayesian inference procedure to approximate the posterior probability. Section 6 refers to the TBMF's experiments and analysis. Section 7 concludes the work.

2. Related work

The signed network is generally modeled in a matrix format, thus the link sign prediction problem can be regarded as a matrix completion problem [20], which can be solved by the matrix factorization method.

Lim and Teh [17] proposed Variational Bayesian-based Matrix Factorization (VBMF) for the movie recommendation. Nakajima et al. [22,23] analyzed VBMF theoretically. Probabilistic Matrix Factorization (PMF) [30] was also proposed to solve the movie recommendation problem. It assumes that the noise is Gaussian distribution and that the prior of latent features is multivariate Gaussian distribution. The gradient descent algorithm is used to find the local optimal solution, which requires careful regularization parameter tuning, though PMF achieves accurate results on the Netflix dataset. Bayesian Probabilistic Matrix Factorization (BMF) [29] addresses the parameter tuning issue via the Markov Chain Monte Carlo method (MCMC), and a number of variants of BMF by fusing the side information of social networks have been proposed [18,25]. Shan and Banerjee [34] extended PMF and BMF, and proposed a series of general PMF (GPMF) methods. However, these models have the same Gaussian assumption as PMF concerning the noise and the prior of latent features. Schmidt et al. [31,32] explored Gaussian noise and the exponential prior of latent features. Cemgil [5] assumed Poisson noise and the gamma prior of latent features, but the Poisson assumption is suitable only for integer data. Student-t distribution was introduced into representations of latent features in [14], and was called Robust Bayesian Matrix Factorization (RBMF). RBMF assumes that the prior of latent features is multivariate Student-t distribution, but the noise is still Gaussian noise. Actually, replacing Gaussian distribution by heavy-tailed distribution such as Student-t has been adopted in many research fields, *e.g.*, principle component analysis [50], logistic regression [9], approximate inference [8] and clustering [2]. Besides, many approximation methods have been proposed for the intractable posterior probability, such as the Markov Chain Monte Carlo (MCMC) method [11] and the variational Bayesian method [19,24,38].

The main challenge is how to incorporate Student-t distribution to characterize both the noise and the prior of latent features, which is the aim of this work. We employ a univariate Student-t distribution for the prediction noise, and a multivariate Student-t distribution for the prior of latent features. In order to find the solution that maximizes the posterior probability, we derive a variational Bayesian inference method to solve our model.

3. Preliminaries

3.1. BMF

Bayesian Probabilistic Matrix Factorization (BMF) is proposed in [29], and is a full Bayesian treatment of PMF. The observed sign matrix $S \in \Re^{N \times M}$ with N users and M items is factorized into two low dimensional matrices U and V:

$$S = U^T V, \tag{1}$$

where $U \in \Re^{K \times N}$ and $V \in \Re^{K \times M}$ are the K dimensional user and item latent feature matrices respectively, and each column represents the latent feature vector.

The noise in BMF is assumed as Gaussian noise, and the conditional probability of the observed sign matrix can be written as:

$$p(S|U, V, \tau) = \prod_{i=1}^{N} \prod_{j=1}^{M} [\mathcal{N}(s_{ij}|u_i^T v_j, \tau^{-1})]^{I_{ij}}. \tag{2}$$

$\mathcal{N}(x|\mu, \tau^{-1})$ is the Gaussian distribution with mean μ and precision τ. I is an indicator function. $I_{ij} = 1$ if the entry s_{ij} is observed and 0 otherwise.

The priors for the representations of latent features U and V are multivariate Gaussian distributions:

$$p(U|\mu_u, \tau_u^{-1}) = \prod_{i=1}^{N} \mathcal{N}(u_i|\mu_u, \tau_u^{-1}), \tag{3}$$

$$p(V|\mu_v, \tau_v^{-1}) = \prod_{j=1}^{M} \mathcal{N}(v_j|\mu_v, \tau_v^{-1}). \tag{4}$$

Gaussian–Wishart priors are given to the hyperparameters $\Theta_u = \{\mu_u, \tau_u\}$ and $\Theta_v = \{\mu_v, \tau_v\}$:

$$p(\Theta_u|\Theta_0) = P(\mu_u|\tau_u)P(\tau_u) = \mathcal{N}(\mu_u|\mu_0, (\beta_0\tau_u)^{-1})\mathcal{W}(\tau_u|W_0, \nu_0), \tag{5}$$

$$p(\Theta_v|\Theta_0) = P(\mu_v|\tau_v)P(\tau_v) = \mathcal{N}(\mu_v|\mu_0, (\beta_0\tau_v)^{-1})\mathcal{W}(\tau_v|W_0, \nu_0), \tag{6}$$

where $\Theta_0 = \{W_0, \nu_0\}$ is the top parameter. $\mathcal{W}(x|W, \nu)$ is the Wishart distribution with the scale matrix $W \in \Re^{K \times K}$ and the degrees of freedom ν.

The goal is to find matrices U and V that maximize the posterior probability $p(U, V|S, \Theta)$ to approximate the observed matrix S, which can be expressed in Bayes' rule:

$$p(U, V|S, \Theta) = \frac{p(S|U, V, \Theta)p(U|\Theta)p(V|\Theta)}{p(S)}, \tag{7}$$

where $\Theta = \{\tau, \Theta_u, \Theta_v, \Theta_0\}$ is the set of all parameters. The missing entry s_{ij}^* from user i to item j is predicted by

$$s_{ij}^* = u_i^T v_j. \tag{8}$$

In BMF, the adopted multivariate Gaussian prior does not discriminate the importance of the latent features. However, the sign matrix is usually sparse (only a few link signs are observed), and only part of the latent features are informative. Thus, we should establish a model which can select informative latent features firstly.

3.2. RBMF

Robust Bayesian Matrix Factorization (RBMF) uses the multivariate Student-t prior for user and item latent features [14], but its noise model is the same as BMF:

$$p(S|U, V, \tau) = \prod_{i=1}^{N}\prod_{j=1}^{M}[\mathcal{N}(s_{ij}|u_i^T v_j, \tau_{ij}^{-1})]^{I_{ij}}. \tag{9}$$

The multivariate Student-t priors for the latent feature matrices U and V are modeled as:

$$p(U|\alpha, \tau_u^{-1}) = \prod_{i=1}^{N}\int \mathcal{N}(u_i|0, (\alpha_i\tau_u)^{-1})p(\alpha_i)d\alpha_i, \tag{10}$$

$$p(V|\beta, \tau_v^{-1}) = \prod_{j=1}^{M}\int \mathcal{N}(v_j|0, (\beta_j\tau_v)^{-1})p(\beta_j)d\beta_j, \tag{11}$$

where $\alpha = \{\alpha_1, \ldots, \alpha_N\}$ is the set of user scaling factors, $\beta = \{\beta_1, \ldots, \beta_M\}$ is the set of item scaling factors, and the scaling factors α, β are given the Gamma prior:

$$p(\alpha|\Theta_0) = \prod_{i=1}^{N} Gamma\left(\alpha_i|\frac{a_0}{2}, \frac{b_0}{2}\right), \tag{12}$$

$$p(\beta|\Theta_0) = \prod_{j=1}^{M} Gamma\left(\beta_j|\frac{c_0}{2}, \frac{d_0}{2}\right), \tag{13}$$

where $\Theta_0 = \{a_0, b_0, c_0, d_0\}$ is the parameters of Gamma distribution. a_0, b_0 are shared by all users and c_0, d_0 are shared by all items. The Probability Density Function (PDF) of a variable $x \sim Gamma(x|a, b)$ obeying Gamma distribution [10] is

$$Gamma(x|a, b) = \frac{b^a}{\Gamma(a)}x^{a-1}\exp(-bx), \tag{14}$$

where a is the shape parameter and b is the rate parameter.

Both BMF and RBMF assume that the noise between the real entry s_{ij} and the predicted entry $s_{ij}^* = u_i^T v_j$ obeys Gaussian distribution. This assumption implies that the users and items share the same constant noise level, which is too strong and is not reasonable in most cases. The pre-processing normalisation method [25] may help to reduce the difference of the noise level between users and items. However, this method is associated with the dataset and may lead to severe estimation error when the dataset is very sparse or imbalanced. Moreover, as shown in Fig. 1, the noise in real-world datasets usually presents the long-tail characteristic. Because of the improper choice of the noise model and the prior model of latent features, BMF and RBMF perform poorly on many datasets. Therefore, we propose TBMF using Student-t distribution to characterize both the noise and the representation prior of latent features. The main differences between our proposed TBMF and PMF, BMF, RBMF are shown in Fig. 2.

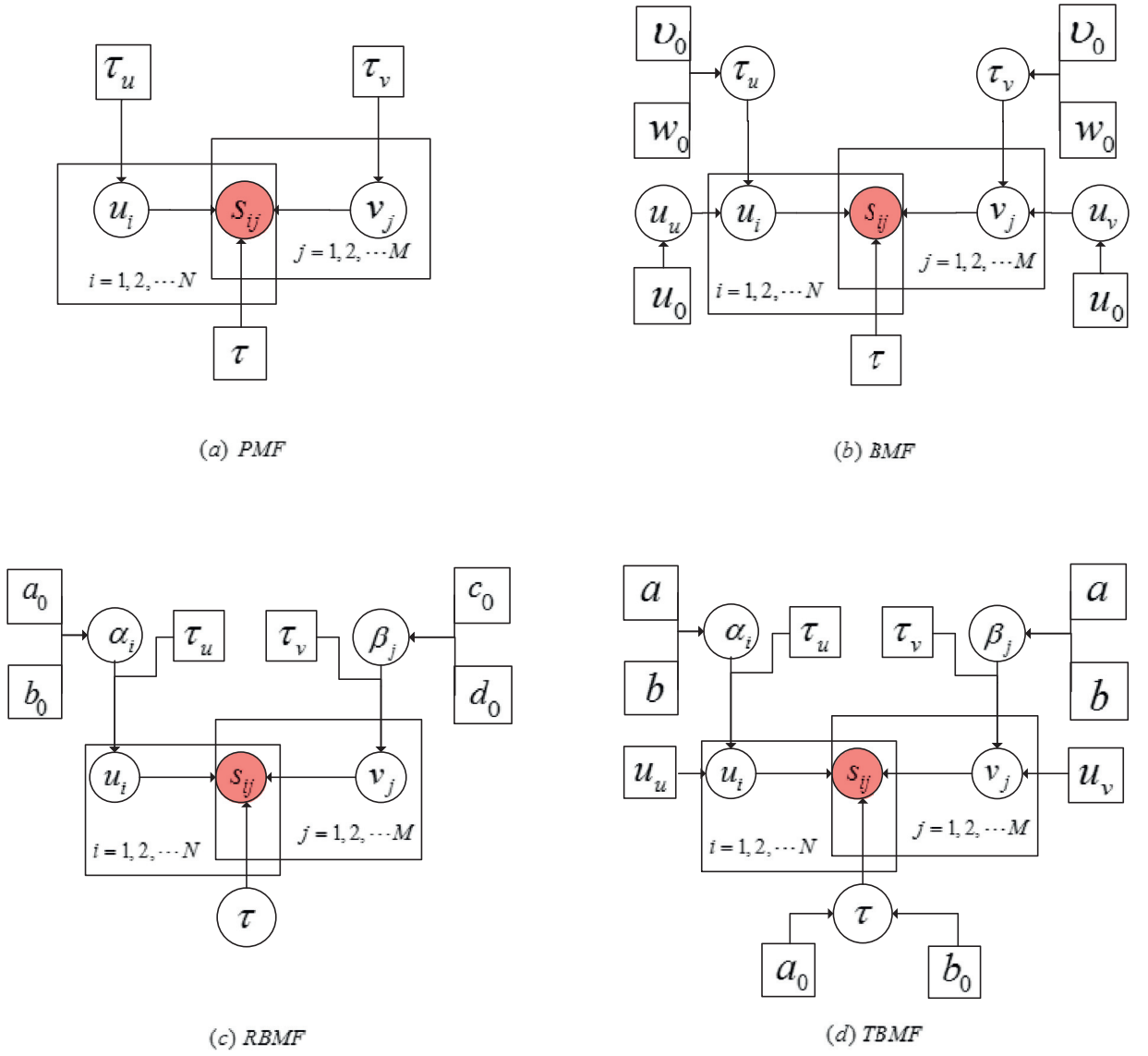

Fig. 2. Graphical description of various models. The non-transparent circle is the variable we want to infer. The transparent circles represent the latent variables and the transparent rectangles are the top parameters. In sub-figure (a), s_{ij} is the variable we want to infer, and the dependencies between variables can be seen from a hierarchical view. In the first layer, s_{ij} depends on u_i, v_j, τ. In the second layer, u_i, v_j obey multivariate Gaussian distribution with the same mean zero and precision τ_i, τ_j respectively. And τ, τ_u, τ_v are fixed top parameters. A similar explanation of sub-figure (b) is that u_i, v_j have different means μ_u, μ_v. Both mean values are given a prior with parameter μ_0. The precisions τ_u, τ_v are given priors with parameters w_0, ν_0, where μ_0, w_0, ν_0 obey the Gaussian–Wishart distribution. Sub-figures (c)(d) are more complex models. Scaling parameters α_i, β_j are applied to the precisions τ_u, τ_v, and α_i, β_j are the assumed Gamma distribution with parameters. The differences lie in u_i, v_j, τ, where sub-figure (d) assumes different mean priors μ_u, μ_v rather than the fixed zero mean in sub-figure (c), and τ is also given a Gamma prior with parameters a_0, b_0 in sub-figure (d).

4. The proposed TBMF

To better fit the signed data noise and obtain better representations of user and item latent matrices, a Bayesian probabilistic matrix factorization with Student-t prior model (TBMF) is proposed. It consists of two parts: the noise model part and the prior model part. The noise model is built via the univariate Student-t distribution, and the multivariate Student-t prior is placed on the latent features of users and items.

4.1. Noise model

The Probability Density Function (PDF) of a variable obeying Student-t distribution $x \sim t(x|\mu, \tau, \nu)$ can be written as:

$$t(x|\mu,\tau,\nu)=\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}\left(\frac{\tau}{\pi\nu}\right)^{1/2}\left[1+\frac{\tau(x-\mu)^2}{\nu}\right]^{-(\nu+1)/2}, \tag{15}$$

where μ is the mean, τ is the precision and ν is the degrees of freedom.

When $\nu\to\infty$, it is easy to find that

$$\left[1+\frac{\tau(x-\mu)^2}{\nu}\right]^{-(\nu+1)/2}\to\exp\left[-\frac{\tau(x-\mu)^2}{2}\right]. \tag{16}$$

Besides, according to Stirling's approximation [27], we have:

$$\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)}\left(\frac{\tau}{\pi\nu}\right)^{1/2}\to\left(\frac{\tau}{\pi}\right)^{\frac{1}{2}}\ \text{as}\ \nu\to\infty. \tag{17}$$

Thus, Eq. (15) can be rewritten as:

$$t(x|\mu,\tau,\nu)=\left(\frac{\tau}{\pi}\right)^{\frac{1}{2}}\exp\left[-\frac{\tau(x-\mu)^2}{2}\right]\ \text{as}\ \nu\to\infty, \tag{18}$$

which is exactly the Gaussian distribution with mean μ and precision τ.

Student-t distribution is therefore a generalization of Gaussian distribution with an additional parameter ν, and when $\nu\to\infty$, it degenerates to Gaussian distribution. The Student-t noise model is thus suitable for both cases in which the noise is exactly Student-t distribution and Gaussian distribution without known spectrum [28]. Moreover, Student-t distribution has the heavy-tailed characteristic and is robust to outliers.

The Student-t distribution of a variable x can be derived as follows: Given $x\sim\mathcal{N}(x|\mu,\tau^{-1})$ with mean μ and precision τ, and τ follows a *Gamma* prior $\tau\sim\Gamma(\tau|a_0,b_0)$ with shape a_0 and rate b_0, we obtain

$$\begin{aligned}
p(x|\mu,a_0,b_0)&=\int\mathcal{N}(x|\mu,\tau^{-1})Gamma(\tau|a_0,b_0)d\tau\\
&=\int\left(\frac{\tau}{2\pi}\right)^{1/2}\exp\left[-\frac{\tau(x-\mu)^2}{2}\right]\frac{b_0^{a_0}}{\Gamma(a_0)}\tau^{a_0-1}\exp(-b_0\tau)d\tau\\
&=\frac{b_0^{a_0}\Gamma(a_0+1/2)}{(2\pi)^{1/2}\Gamma(a_0)[(x-\mu)^2/2+b_0]^{(a_0+1/2)}}\\
&=\frac{\Gamma((2a_0+1)/2)}{\Gamma(2a_0/2)}\left[\frac{a_0/b_0}{2a_0\pi}\right]^{1/2}\left[1+\frac{a_0/b_0(x-\mu)^2}{2a_0}\right]^{-(2a_0+1)/2}\\
&=t(x|\mu,a_0/b_0,2a_0).
\end{aligned} \tag{19}$$

We can see that the variable x obeys a Student-t distribution with mean μ, precision $\tau=a_0/b_0$ and degrees of freedom $\nu=2a_0$. Thus, the Student-t noise model can be written as

$$p(S|U,V,\tau)=\prod_i\prod_j[t(s_{ij}|u_i^Tv_j,a_0/b_0,2a_0)]^{I_{ij}}. \tag{20}$$

4.2. Prior model

The multivariate Student-t prior has the ability to achieve sparse representation of latent features; that is, it can discern the importance of the latent features and select the most important ones.

From the hierarchy view, given each user distribution $u_i\sim\mathcal{N}(u_i|\mu_u,(\alpha_i\tau_u)^{-1})$ and the scaling factor prior $\alpha_i\sim Gamma(\alpha_i|a,b)$, we derive:

$$\begin{aligned}
p(u_i)&=\int\mathcal{N}(u_i|\mu_u,(\alpha_i\tau_u)^{-1})Gamma(\alpha_i|a,b)d\alpha_i\\
&=\frac{|\tau_u|^{1/2}}{(2\pi)^{K/2}}\frac{b^a}{\Gamma(a)}\int\alpha_i^{a+K/2-1}\exp\left[-\frac{(u_i-\mu_u)^T\tau_u(u_i-\mu_u)}{2}\right]d\alpha_i\\
&=\frac{|\tau_u|^{1/2}}{(2\pi)^{K/2}}\frac{b^a}{\Gamma(a)}\frac{\Gamma(K/2+a)}{[b+\frac{1}{2}(u_i-\mu_u)^T\tau_u(u_i-\mu_u)]^{K/2+a}}\\
&=t(u_i|\mu_u,a\tau_u/b,2a),
\end{aligned} \tag{21}$$

which shows that u_i is a multivariate Student-t distribution with mean μ_u, precision matrix $a\tau_u/b$, and degrees of freedom $2a$. Similarly, the multivariate Student-t distribution of item $v_j\sim t(v_j|\mu_v,a\tau_v/b,2a)$ is obtained following each item distribution $v_j\sim\mathcal{N}(v_j|\mu_v,(\beta_j\tau_v)^{-1})$ and the item scaling factor $\beta_j\sim Gamma(\beta_j|a,b)$.

The probabilities of users and items are thus modeled as:

$$p(U) = \prod_{i=1}^{N} t(u_i|\mu_u, a_i\tau_{ui}/b_i, 2a_i), \tag{22}$$

$$p(V) = \prod_{j=1}^{M} t(v_j|\mu_v, a_j\tau_{vj}/b_j, 2a_j). \tag{23}$$

To simplify the problem, the precision matrix τ_{ui} is limited to being diagonal and is initialized to be the same for all users $\tau_{ui}^{-1} = diag\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2\}$, which indicates that the latent features behind the user u_i are independent of each other. Similarly, $\tau_{vj}^{-1} = diag\{\rho_1^2, \rho_2^2, \ldots, \rho_K^2\}$.

The best U and V characterizing sign matrix S are the matrices that maximize the posterior probability $p(U, V|S)$:

$$\begin{aligned} U, V &= arg\max_{U,V} p(U, V|S) \\ &= arg\max_{U,V} \frac{p(S|U, V)p(U, V)}{p(S)}. \end{aligned} \tag{24}$$

5. Variational inference

Calculating posterior probability $p(U, V|S)$ is difficult because $p(S)$ cannot be computed easily, so we employ a variational Bayesian method [38] to approximate the posterior.

5.1. Variational expectation maximization approach

The core idea of variational Bayesian inference is to use the probability $q(U)$ and $q(V)$ to approximate the posterior probability $p(U|S)$ and $p(V|S)$, respectively. Following the graphical description of TBMF in Fig. 2, we use $Z = \{U, V, \alpha, \beta, \tau\}$ to denote the set of latent variables and $\Theta = \{\mu_u, \mu_v, \tau_u, \tau_v, a, b, a_0, b_0\}$ to denote the set of fixed parameters.

The log model evidence $p(S)$ can be expressed as:

$$\begin{aligned} \ln p(S; \Theta) &= KL[q(Z)||p(Z|S; \Theta)] + F[q(Z), \Theta], \\ KL[q(Z)||p(Z|S; \Theta)] &= -\int q(Z) \ln [p(Z|S; \Theta)/q(Z)]dZ, \\ F[q(Z), \Theta] &= \int q(Z) \ln [p(Z, S; \Theta)/q(Z)]dZ, \end{aligned} \tag{25}$$

where $KL[q(Z)||p(Z|S; \Theta)]$ is the Kullback Leibler(KL)-divergence [33] between $q(Z)$ and $p(Z|S; \Theta)$, and $F[q(Z), \Theta]$ is the variational free energy.

KL-divergence is nonnegative and gets zero when $q(Z) = p(Z|S; \Theta)$. As the log-likelihood $\ln p(S; \Theta) \geq F[q(Z), \Theta]$ holds, the variational free energy $F[q(Z), \Theta]$ provides a lower bound of the log-likelihood $\ln p(S; \Theta)$. That is to say, KL-divergence will be close to zero by maximizing the variational free energy $F[q(Z), \Theta]$, thus the optimization problem can be described as

$$Z, \Theta = \arg\max_{Z,\Theta} F[q(Z), \Theta] \tag{26}$$

As there are latent variables, we cannot obtain the optimization solution described in Eq. (26), but the variational Expectation Maximization (EM) algorithm [3] can be used to solve it by iteratively maximizing $F[q(Z), \Theta]$. In the E-step, the parameter set Θ^{old} is fixed, $F[q(Z), \Theta^{old}]$ is maximized with respect to $q(Z)$, where $q(Z) = p(Z|S; \Theta^{old})$. In M-step, $q(Z)$ is fixed, $F[q(Z), \Theta]$ is maximized with respect to Θ to gain new parameter values Θ^{new}.

We substitute $q(Z) = p(Z|S; \Theta^{old})$ into Eq. (25) and get

$$\begin{aligned} F[q(Z), \Theta] &= \int p(Z|X; \Theta^{old}) \ln p(Z, S; \Theta)dZ - \int p(Z|S; \Theta^{old}) \ln p(Z|S; \Theta^{old})dZ \\ &= Q(\Theta, \Theta^{old}) + C, \end{aligned} \tag{27}$$

where C is a constant and

$$Q(\Theta, \Theta^{old}) = \int p(Z|S; \Theta^{old}) \ln p(Z, S; \Theta)dZ = E_{p(Z|S;\Theta^{old})} \ln p(Z, S; \Theta), \tag{28}$$

which is the expectation of the log-likelihood of both the observations and the latent variables. This should be maximized in the M-step to obtain new parameter values Θ^{new}.

To sum up, the variational EM algorithm has two iterative steps,

- E-step: Compute $p(Z|S; \Theta^{old})$;

- M-step: Obtain new parameter set $\Theta^{new} = arg\max_{\Theta} Q(\Theta, \Theta^{old})$;

In this paper, we suppose that the parameter set Θ is fixed and the variational EM algorithm only contains the E-step. The optimization problem described in Eq. (24) is then equivalent to the optimization problem

$$U, V = arg\max_{U,V} F[q(U, V)]. \tag{29}$$

According to the mean-field assumption [39], the variational distribution $q(Z)$ can be factorized over the latent variables denoted as $q(Z) = \prod_i q_i(Z_i)$. Thus, $q(Z) = q(U)q(V)q(\tau)q(\alpha)q(\beta)$. Then Eq. (25) can be rewritten as

$$\begin{aligned} F[q(Z)] &= \int q(Z) ln \frac{p(Z,S)}{q(Z)} dZ \\ &= \int \prod_i q_i(Z_i) \times [\ln p(S,Z) - \sum_i \ln q_i(Z_i)] dZ \\ &= \int q_j(Z_j) \prod_{\setminus j} q_i(Z_i) (\ln p(S,Z) - \ln q_j(Z_j)) dZ - \int q_j(Z_j) \prod_{\setminus j} q_i(Z_i) \sum_{\setminus j} \ln q_i(Z_i) dZ \\ &= \int q_j(Z_j) \left(\int \prod_{\setminus j} q_i(Z_i) \ln p(Z,S) d\Theta_{\setminus j} - \ln q_j(Z_j) \right) d\Theta_j - \int q_j(Z_j) \int \prod_{\setminus j} q_i(Z_i) \ln \prod_{\setminus j} q_i(Z_i) d\Theta_{\setminus j} d\Theta_j \\ &= -KL[q_j(Z_j) || \exp(E_{q_{\setminus j}(Z_j)} \ln p(S,Z))] + c, \end{aligned} \tag{30}$$

where $q_{\setminus j}(Z_j)$ is the joint distribution of all parameters except Z_j.

Supposing the distribution $q_{\setminus j}(Z_j)$ is fixed, we obtain the approximated posterior probability $q_j(Z_j)$ that maximizes free energy $F[q(Z)]$:

$$\begin{aligned} q_j^*(Z_j) &= arg\max_{q_j(Z_j)} F[q(Z)] \\ &= arg\max_{q_j(Z_j)} \frac{1}{N_o} \exp[E_{q_{\setminus j}(Z_j)} \ln p(S,Z)], \end{aligned} \tag{31}$$

where N_o is a normalizing constant.

The log form of the best PDF of the jth latent variable that maximizes the variational free energy is thus

$$\ln q_j^*(Z_j) = E_{q_{\setminus j}(Z_j)} \ln p(S,Z) - \ln N_o. \tag{32}$$

In other words, the best $q_j^*(Z_j)$ can be gained by calculating the expectation of $\ln p(S, Z)$ when all the other latent variables are fixed.

In summary, we fix the latent variables in our variational EM algorithm except for the jth latent variable and calculate the expectation of the log-likelihood $\ln p(S, Z)$ as the log form of the PDF of the jth latent variable. We do this iteratively until the variational free energy $F[q(Z)]$ no longer changes.

5.2. Inference procedure

The mathematical description of our problem can be written as

$$p(S, U, V, \tau, \alpha, \beta) = p(S|U, V, \tau) p(U|\alpha) p(V|\beta) p(\tau) p(\alpha) p(\beta), \tag{33}$$

where the probabilities of the individual variables are as listed below:

$$p(S|U, V, \tau) = \prod_{i=1}^{N} \prod_{j=1}^{M} \mathcal{N}(s_{ij}|u_i^T v_j, \tau^{-1}), \tag{34}$$

$$p(U|\alpha) = \prod_{i=1}^{N} \mathcal{N}(u_i|\mu_u, (\alpha_i \tau_{ui})^{-1}), \tag{35}$$

$$p(V|\beta) = \prod_{j=1}^{M} \mathcal{N}(v_j|\mu_v, (\beta_j \tau_{vj})^{-1}), \tag{36}$$

$$p(\tau) = Gamma(\tau|a_0, b_0), \tag{37}$$

$$p(\alpha) = \prod_{i=1}^{N} Gamma(\alpha_i|a, b), \tag{38}$$

$$p(\beta) = \prod_{j=1}^{M} Gamma(\beta_j|a, b). \tag{39}$$

In our model, the latent variables $Z = \{U, V, \tau, \alpha, \beta\}$ can be derived using the variational EM algorithm. We also consider a structural variational approximation [43]. The approximated posteriors $q(U|\alpha)$, $q(V|\beta)$ of the users and items are in the same format as their priors; that is to say, they can be factorized into a product of Gaussian distribution with scaling factors

$$q(U|\alpha) = \prod_{i=1}^{N} \mathcal{N}(u_i|\hat{\mu}_{ui}, (\alpha_i \hat{\tau}_{ui})^{-1}), \tag{40}$$

$$q(V|\beta) = \prod_{j=1}^{M} \mathcal{N}(v_j|\hat{\mu}_{vj}, (\beta_j \hat{\tau}_{vj})^{-1}). \tag{41}$$

It is not difficult to obtain the form of the defined parameters in Eq. (40)

$$\begin{aligned} \hat{\tau}_{ui} &= \overline{\tau} \sum_{j \in Ne_i} (\overline{v_j v_j}^T + \hat{\tau}_{vj}) + \overline{\alpha}_i \tau_{ui}, \\ \hat{\mu}_{ui} &= \hat{\tau}_{ui}^{-1} \left(\overline{\tau} \sum_{j \in Ne_i} s_{ij} \overline{v}_j + \overline{\alpha}_i \hat{\tau}_{ui} \mu_{ui} \right), \end{aligned} \tag{42}$$

where Ne_i denotes the collection of items whose link signs from user i are observed, namely $s_{ij} \neq 0$. $\overline{\tau}, \overline{\alpha_i}, \overline{v_j}$ are the mean of τ, α_i, v_j.

Similarly, the derivation of parameters in Eq. (41) is

$$\begin{aligned} \hat{\tau}_{vj} &= \overline{\tau} \sum_{i \in Ne_j} (\overline{u_i u_i}^T + \hat{\tau}_{ui}^{-1}) + \overline{\beta}_j \tau_{vj}, \\ \hat{\mu}_{vj} &= \left(\overline{\tau} \sum_{i \in Ne_j} s_{ij} \overline{u_i}^T + \overline{\beta}_j \mu_v^T \hat{\tau}_{vj} \right)^T \hat{\tau}_{vj}. \end{aligned} \tag{43}$$

The distribution of τ is Gamma distribution $\tau \sim Gamma(\tau|\hat{a}_0, \hat{b}_0)$ and the parameter derivation is

$$\begin{aligned} \hat{a}_0 &= a_0 + \frac{1}{2}, \\ \hat{b}_0 &= b_0 + \frac{\sum_{ij \in \{s_{ij} \neq 0\}} (s_{ij} - \overline{u_i^T v_j})^2}{2n_s}, \end{aligned} \tag{44}$$

where n_s is the number of links whose signs are observed.

Because α is independent across i, its structural posterior approximation is

$$q(\alpha) = \prod_{i=1} Gamma(\alpha_i|\hat{c}_i, \hat{d}_i), \tag{45}$$

and the defined parameters are

$$\begin{aligned} \hat{c}_i &= c_i + \frac{1}{2}, \\ \hat{d}_i &= d_i + \frac{1}{2} \left(\frac{1}{\hat{\sigma}_1^2}, \ldots, \frac{1}{\hat{\sigma}_K^2} \right) (\overline{u}_i \cdot \overline{u}_i + (\hat{\sigma}_1^2, \ldots, \hat{\sigma}_K^2)^T) - \frac{1}{2} \overline{\mu}_u \tau_u \mu_u, \end{aligned} \tag{46}$$

where $\hat{\sigma}_1^2, \ldots, \hat{\sigma}_K^2$ are the entries on the diagonal of the user covariance matrix.

Similarly,

$$q(\beta) = \prod_{j=1}^{M} Gamma(\beta_j|\hat{e}_j, \hat{f}_j), \tag{47}$$

and the parameter derivations are

$$\begin{aligned} \hat{e}_j &= e_j + \frac{1}{2}, \\ \hat{f}_j &= f_j + \frac{1}{2} \left(\frac{1}{\hat{\rho}_1^2}, \ldots, \frac{1}{\hat{\rho}_K^2} \right)^t (\overline{v}_j \cdot \overline{v}_j + (\hat{\rho}_1^2, \ldots, \hat{\rho}_K^2)^T) - \frac{1}{2} \overline{\mu}_v \tau_v \mu_v, \end{aligned} \tag{48}$$

where $\hat{\rho}_1^2, \ldots, \hat{\rho}_K^2$ are the entries on the diagonal of the user covariance matrix.

The complete inference algorithm is presented in Algorithm 1.

Algorithm 1 Inference algorithm.

Input: *Observed Sign Matrix S*
Output: *Optimization Matrix U, V*

```
Init fixed parameters Θ = μ_u, μ_v, τ_u, τ_v, a_0, b_0, a, b;
repeat
  for all i ∈ 1...N do
    Update parameters τ_ui, μ_ui in q(u_i) according to Eq. 42
  end for
  for all j ∈ 1...M do
    Update parameters τ_vj, μ_vj in q(v_j) according to Eq. 43
  end for
  Update parameters a, b in q(τ) according to Eq. 44
  for all i ∈ 1...N do
    Update parameters c_i, d_i in q(α_i) according to Eq. 46
  end for
  for all j ∈ 1...M do
    Update parameters e_j, f_j in q(β_j) according to Eq. 48
  end for
until Convergence
Return U, V
```

Table 1
Detailed dataset information.

	Wikipedia	Slashdot	Epinions
Node	7118	82,144	131,828
Link	103,747	549,202	841372
Sparsity	0.41%	0.02%	0.01%

6. Experiments

In this section, we conduct numerous experiments on signed network datasets from the Stanford large network dataset collection[1]. We compare TBMF with the baseline models PMF [30], BMF [29], RBMF [14] and Nonnegative Matrix Factorization (NMF) [35]. In addition, we explore the effects of the sampling rate and the number of latent features on the prediction performance.

6.1. Datasets

Wikipedia, Slashdot and *Epinions* are three widely used datasets for signed network problems. *Wikipedia* is a free encyclopedia whose dataset records the Wikipedia adminship process by which community users vote to elect site administrators. '+1' indicates that a candidate is supported, while '−1' indicates that the election of a candidate is opposed. There are about 2800 elections with around 100,000 votes, and approximately 7000 users participate in the elections. *Slashdot* is a technology news website which has multiple user communities and allows users to tag each other as friend or foe. The dataset contains links and corresponding friend or foe markers, and consists of about 82,000 users and 549,000 links. *Epinions* is a consumer review site on which users can label their trusted users. The resulting trust relationships are then combined with the review ratings. The dataset contains about 130,000 users and 840,000 relationships. Statistical information about the three datasets is shown in Table 1. The sparsity of the datasets can be computed as the fraction of the true links in the full connected nodes, *e.g.*, for *Wikipedia*, the sparsity is $\frac{103747}{C_{7118}^2} = 0.41\%$. Table 1 presents that the sparsities of all datasets are below 1%.

6.2. Experimental setup

The initial parameters of our method are as follows. For the multivariate Student-t prior model, the mean values of the users and items μ_u, μ_v are set to all-zero vectors. The diagonal elements of the covariance matrix τ_u^{-1} of the users $\sigma_1^2, \ldots, \sigma_K^2$ are all set as 1, while the diagonal elements in the covariance matrix τ_v^{-1} of the items $\rho_1^2, \ldots, \rho_K^2$ are set as $1/K$. For the Student-t noise model, the Gamma distribution parameters a_0, b_0 for the precision τ are set to be non-informatively small, *e.g.*, 10^{-6}. This is the same as the Gamma distribution parameters *a, b* for the scaling factor of users and items in the

[1] https://snap.stanford.edu/data/

Table 2
Variation in ACC of models by sampling rate.

Sampling rate	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	
Wikipedia											
PMF	0.5020	0.5771	0.7014	0.7295	0.7424	0.7465	0.7488	0.7564	0.7596	0.7570	9.5%↑
BMF	0.5728	0.6479	0.7722	0.8003	0.8132	0.8173	0.8196	0.8272	0.8304	0.8278	2.4%↑
RBMF	0.5819	0.6070	0.7657	0.7986	0.8086	0.8135	0.8217	0.8262	0.8270	0.8286	2.9%↑
NMF	0.5206	0.5957	0.7200	0.7481	0.7610	0.7651	0.7674	0.7750	0.7782	0.7756	7.7%↑
TBMF	0.6258	0.7086	0.7756	0.8132	0.8308	0.8380	0.8425	0.8458	0.8514	0.8452	–
Slashdot											
PMF	0.4754	0.6189	0.6572	0.6588	0.7881	0.7783	0.7790	0.7959	0.7899	0.7979	8.2%↑
BMF	0.5492	0.6327	0.7240	0.7759	0.7978	0.8080	0.8159	0.8238	0.8240	0.8271	3.8%↑
RBMF	0.5587	0.6383	0.7025	0.7894	0.8098	0.8183	0.8186	0.8219	0.8183	0.8187	3.6%↑
NMF	0.5160	0.5767	0.5723	0.7076	0.7780	0.7863	0.7922	0.8089	0.8090	0.8176	7.9%↑
TBMF	0.6614	0.7498	0.7774	0.8025	0.8131	0.8225	0.8270	0.8325	0.8379	0.8363	–
Epinions											
PMF	0.5932	0.6375	0.8132	0.8729	0.8396	0.8883	0.8855	0.9056	0.9214	0.9027	9.6%↑
BMF	0.5641	0.6700	0.8429	0.8697	0.8799	0.9059	0.9252	0.9284	0.9291	0.9318	7.7%↑
RBMF	0.7374	0.8338	0.8717	0.8881	0.8958	0.9001	0.9307	0.9320	0.9323	0.9326	3.7%↑
NMF	0.6413	0.6815	0.8032	0.8835	0.8732	0.8937	0.9037	0.9176	0.9240	0.9237	6.5%↑
TBMF	0.8836	0.9056	0.9173	0.9245	0.9283	0.9301	0.9326	0.9329	0.9332	0.9340	–

multivariate Student-t prior model. For fair comparison, we initialize the parameters in PMF, BMF and RBMF in the same way. As for NMF, because the entries of the matrix are required to be non-negative, we modify the input matrix, *i.e.*, '+1' to '5' and '−1' to '1'. The unknown cases are still '0'. If the output of NMF is equal to or larger than '3', we regard it as '+1', and if it is less than '3', we regard it as '−1'.

A special 10 times 10-fold cross-validation is carried out in the experiments. As the dataset is unbalanced and most of the link signs are positive, we first divide the dataset into positive and negative parts and then divide the two parts into ten folds [41]. The final single fold consists of one fold from the positive part and one fold from the negative part.

We compare the models in term of predictive performance based on the following two metrics [21,42]:

Accuracy (ACC): is the proportion of true prediction results in the total number of cases examined.

$$ACC = \frac{\sum_{ij\in\Omega_{test}} I(s_{ij}s^*_{ij})}{\sum_{ij\in\Omega_{test}} 1} \tag{49}$$

where Ω_{test} is the test set and $I(x)$ is a sign function which takes 1 if $x > 0$ and takes 0 for other cases.

Area Under the ROC Curve (AUC): is a metric for evaluating performance on imbalanced datasets. Firstly, Receiver Operating Characteristic (ROC) curve is a graphical plot which illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting the fraction of true positives out of the positives (TPR) *vs.* the fraction of false positives out of the negatives (FPR), at various threshold settings. Then, AUC computes the area under the ROC curve, which uses one number to summarize the curve information [47].

6.3. Results and analysis

We evaluate the performance of TBMF, PMF, BMF, RBMF and NMF on the three datasets under the *ACC* and *AUC* metrics. We also analyze the effects of the sampling rate and the number of latent features on prediction results.

6.3.1. Effects of sampling rate

The training set is randomly sampled according to a sampling rate that ranges from 0.1 to 1.0. A sampling rate of 1.0 means the entire training set is used. We use 20 latent features to represent users and items.

Tables 2 and 3 report the performance of PMF, BMF, RBMF, NMF and TBMF on the three datasets when the sampling rate varies from 0.1 to 1.0. The last column is the average improvement of TBMF compared to other algorithms. We can see that TBMF performs better than PMF, BMF, RBMF and NMF on almost all datasets under both *ACC* and *AUC* metrics. Taking Table 2 as an example, we can find that the average *ACC* improvement on the three datasets is 9.1%, 4.6%, 3.4% and 7.3% when compared respectively with PMF, BMF, RBMF and NMF.

For a better view, we also plot the prediction performance curves on the three datasets with varied sampling rates, as shown in Fig. 3. We can clearly see the trend whereby *ACC* and *AUC* increase with the sampling rate, which indicates that more training data usually lead to better prediction performance. The first row of Fig. 3, *i.e.*, sub-figures (a)–(c), gives the results under the *ACC* metric. We find that there is a big gap between TBMF and the baseline models when the sampling rate is 0.1. That is to say, TBMF performs well when the observed link signs are few. The second row shows the results under the *AUC* metric (sub-figures (d)–(f)). We can see that TBMF still outperforms others except for the *Slashdot* dataset at a low sampling rate. The reason is that *Slashdot*'s degree of imbalance is high, which affects its performance when the sampling rate is low, *i.e.*, 0.1 or 0.2. Note that when the sampling rate is equal to or larger than 0.5, TBMF still performs better

Table 3
Variation in AUC of models by sampling rate.

Sampling rate	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	
Wikipedia											
PMF	0.7539	0.7424	0.7851	0.8055	0.8383	0.8363	0.8360	0.8245	0.8372	0.8394	4.2%↑
BMF	0.8032	0.8284	0.8388	0.8517	0.8552	0.8622	0.8492	0.8539	0.8534	0.8576	0.6%↑
RBMF	0.7825	0.7836	0.8220	0.8473	0.8573	0.8498	0.8476	0.8453	0.8437	0.8396	2.0%↑
NMF	0.7478	0.7697	0.7840	0.7905	0.8422	0.8368	0.8271	0.8352	0.8393	0.8460	4.0%↑
TBMF	0.8043	0.8231	0.8443	0.8609	0.8631	0.8638	0.8604	0.8670	0.8631	0.8716	–
Slashdot											
PMF	0.8493	0.8357	0.8326	0.8212	0.8434	0.8573	0.8441	0.8405	0.8221	0.8355	0.8%↑
BMF	0.8402	0.8458	0.8316	0.8315	0.8338	0.8366	0.8413	0.8456	0.8416	0.8407	0.7%↑
RBMF	0.8404	0.8434	0.8407	0.8531	0.8642	0.8712	0.8459	0.8362	0.8177	0.7980	0.4%↑
NMF	0.8179	0.8337	0.8364	0.8405	0.8458	0.8466	0.8498	0.8483	0.8442	0.8358	0.6%↑
TBMF	0.7736	0.8121	0.8383	0.8535	0.8548	0.8605	0.8607	0.8656	0.8703	0.8686	–
Epinions											
PMF	0.8432	0.8645	0.8646	0.8691	0.8666	0.8658	0.8906	0.8935	0.8972	0.8970	5.3%↑
BMF	0.8669	0.8599	0.8402	0.8446	0.8559	0.8994	0.9060	0.9012	0.8983	0.9044	5.1%↑
RBMF	0.8990	0.9035	0.9137	0.9219	0.9309	0.9305	0.9315	0.9338	0.9347	0.9365	0.5%↑
NMF	0.8734	0.8784	0.8868	0.8833	0.8805	0.8769	0.8971	0.9054	0.9182	0.9107	3.7%↑
TBMF	0.9205	0.9201	0.9234	0.9303	0.9319	0.9318	0.9346	0.9229	0.9301	0.9370	–

Table 4
Variation in ACC of models by latent feature number.

Feature no.	2	4	6	8	10	15	20	30	40	50	
Wikipedia											
PMF	0.7324	0.7363	0.7361	0.7330	0.7420	0.7440	0.7454	0.7504	0.7471	0.7573	9.3%↑
BMF	0.8035	0.8115	0.8128	0.8119	0.8092	0.8154	0.8162	0.8132	0.8170	0.8173	2.8%↑
RBMF	0.8235	0.8273	0.8300	0.8312	0.8332	0.8319	0.8353	0.8315	0.8327	0.8335	0.6%↑
NMF	0.7557	0.7564	0.7463	0.7551	0.7630	0.7548	0.766	0.7668	0.7719	0.7666	7.5%↑
TBMF	0.8331	0.8336	0.8357	0.8326	0.8350	0.8362	0.8385	0.8370	0.8396	0.8403	–
Slashdot											
PMF	0.7824	0.7826	0.7886	0.7898	0.7870	0.7870	0.7911	0.7925	0.7927	0.7938	2.9%↑
BMF	0.7995	0.7958	0.7965	0.7952	0.7978	0.7931	0.8002	0.7964	0.7977	0.8031	2.0%↑
RBMF	0.8008	0.8048	0.8067	0.8083	0.8094	0.8070	0.8108	0.8044	0.8067	0.8085	1.1%↑
NMF	0.7841	0.7817	0.7808	0.7816	0.7829	0.7832	0.7852	0.7870	0.7882	0.7804	3.4%↑
TBMF	0.8161	0.8150	0.8157	0.8190	0.8151	0.8167	0.8188	0.8203	0.8213	0.8219	–
Epinions											
PMF	0.8456	0.8616	0.8657	0.8671	0.8744	0.8749	0.8800	0.8857	0.8812	0.8807	5.6%↑
BMF	0.8430	0.8488	0.8456	0.8509	0.8509	0.8751	0.8742	0.8748	0.8799	0.8753	6.6%↑
RBMF	0.8303	0.8497	0.8634	0.8709	0.8763	0.8876	0.8930	0.8932	0.8929	0.8953	5.2%↑
NMF	0.8577	0.8530	0.8559	0.8700	0.8728	0.8644	0.8732	0.8830	0.8810	0.8895	5.8%↑
TBMF	0.9275	0.9272	0.9276	0.9267	0.9285	0.9268	0.9282	0.9285	0.9287	0.9290	–

on the *Slashdot* dataset. In summary, TBMF owns better generalization ability and achieves significantly better prediction performance than the baseline models on almost all datasets.

6.3.2. Effects of latent feature number

We also assess the effect of the latent feature number K on model performance. We change K from 2 to 50 and fix the sampling rate at 0.5.

Tables 4 and 5 present the performance of the models on the datasets under the *ACC* and *AUC* metrics when the latent feature number K varies from 2 to 50. The last column is the average improvement of TBMF compared to other algorithms. We find that TBMF performs better than the baseline models, especially when K is small. Taking Table 4 as an example, we can find that the average *ACC* improvement on the three datasets is 5.9%, 3.8%, 2.3% and 5.5% when compared respectively with PMF, BMF, RBMF and NMF. The reason is that TBMF can distinguish the most informative latent features and obtain a better representation of the users and items. Note that TBMF performs a little bit worse than RBMF under the *AUC* metric on the *Slashdot* dataset (0.3%↓). This is caused by our fixed sampling rate of 0.5 and the high degree of imbalance in the *Slashdot* dataset. We believe that if we increase the sampling rate, TBMF will perform better than RBMF, as indicated in Fig. 3(e).

The performance changing trend of the models with the varied latent feature number K is revealed in Fig. 4. We can see that the performance increases gradually and approaches to a stable value as the number of latent features K increases. When K is very small, the model cannot make full use of the latent features to approximate the sign matrix. When K becomes large, some necessary latent features are included, and if we keep increasing K, there will be redundant latent features, which will increase the computational cost of the matrix factorization, only providing slight performance improvement.

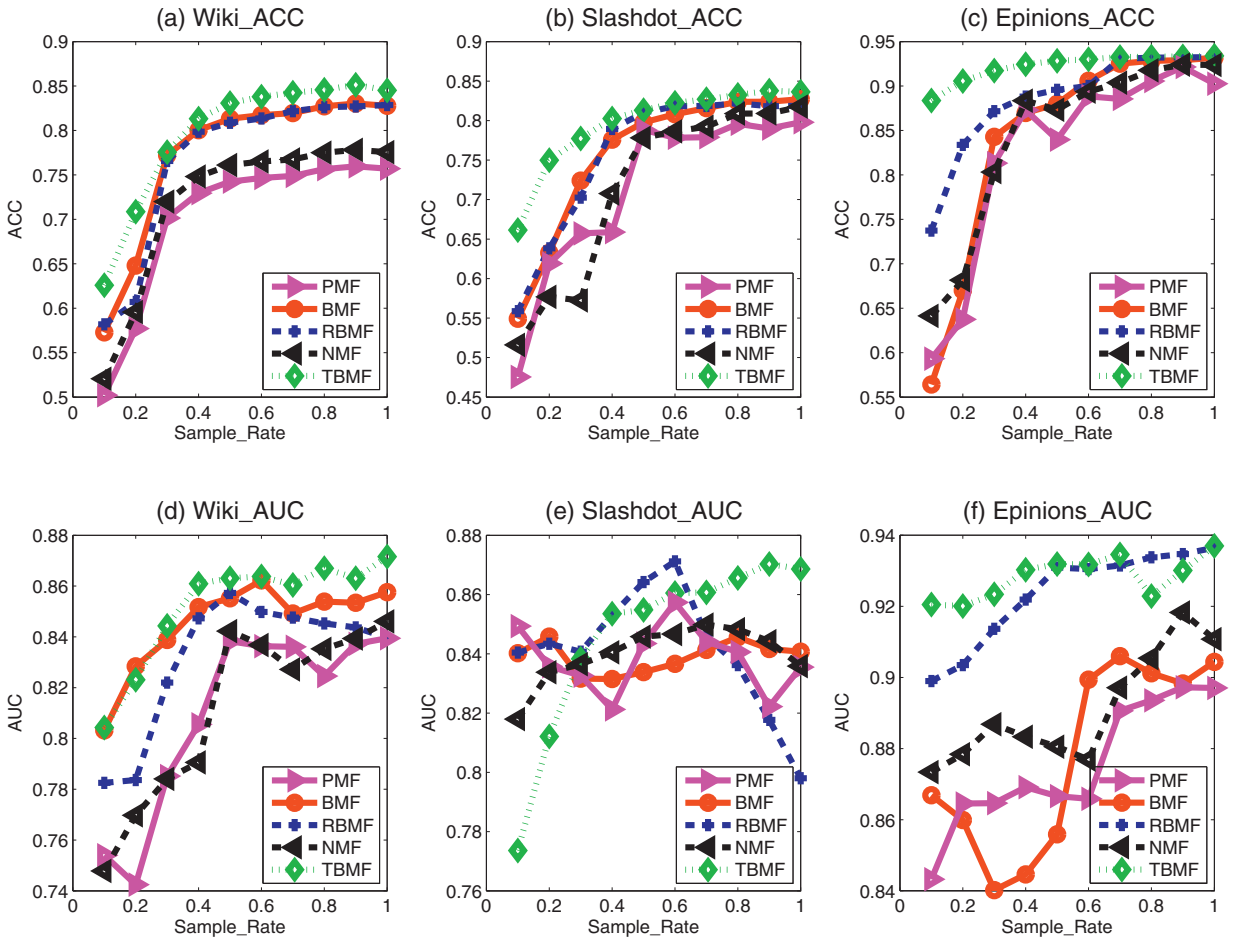

Fig. 3. Effects of sampling rate on prediction performance.

Table 5
Variation in AUC of models by latent feature number.

Feature no.	2	4	6	8	10	15	20	30	40	50	
Wikipedia											
PMF	0.8242	0.8238	0.8361	0.8311	0.8316	0.8362	0.8383	0.8446	0.8438	0.8436	3.5%↑
BMF	0.8583	0.8681	0.8575	0.8558	0.8565	0.8580	0.8534	0.8526	0.8536	0.8539	1.4%↑
RBMF	0.8399	0.8324	0.8386	0.8535	0.8480	0.8544	0.8565	0.8527	0.8572	0.8619	2.1%↑
NMF	0.8355	0.8405	0.8333	0.8359	0.8297	0.8359	0.8422	0.8436	0.8435	0.8424	3.3%↑
TBMF	0.8625	0.8797	0.8706	0.8733	0.8781	0.8680	0.8668	0.8671	0.8709	0.8717	–
Slashdot											
PMF	0.8335	0.8385	0.8408	0.8378	0.8320	0.8361	0.8404	0.8435	0.8438	0.8440	1.6%↑
BMF	0.8408	0.8336	0.8394	0.8362	0.8338	0.8396	0.8329	0.8360	0.8346	0.8357	1.9%↑
RBMF	0.8464	0.8547	0.8615	0.8629	0.8623	0.8663	0.8621	0.8616	0.8613	0.8622	0.3%↓
NMF	0.8239	0.8324	0.8325	0.8372	0.8383	0.8400	0.8458	0.8463	0.8459	0.8480	1.6%↑
TBMF	0.8553	0.8584	0.8612	0.8566	0.8518	0.8561	0.8506	0.8522	0.8531	0.8555	–
Epinions											
PMF	0.8583	0.8660	0.8627	0.8562	0.8608	0.8602	0.8666	0.8669	0.8675	0.8673	6.8%↑
BMF	0.8430	0.8416	0.8495	0.8500	0.8527	0.8575	0.8570	0.8587	0.8590	0.8600	7.8%↑
RBMF	0.9166	0.9228	0.9253	0.9280	0.9273	0.9293	0.9301	0.9296	0.9304	0.9312	0.4%↑
NMF	0.8641	0.8698	0.8772	0.8795	0.8670	0.8657	0.8805	0.8813	0.8818	0.8806	5.6%↑
TBMF	0.9315	0.9287	0.9315	0.9305	0.9337	0.9282	0.9312	0.9322	0.9324	0.9330	–

7. Conclusion

In this paper, we propose a novel Bayesian Matrix Factorization with Student-t prior model (TBMF) to address the link sign prediction problem. TBMF assumes that the sign data noise obeys the univariate Student-t distribution, which makes the model robust to outliers and can deal with the long-tail link signs. Besides, TBMF also performs well when the observed

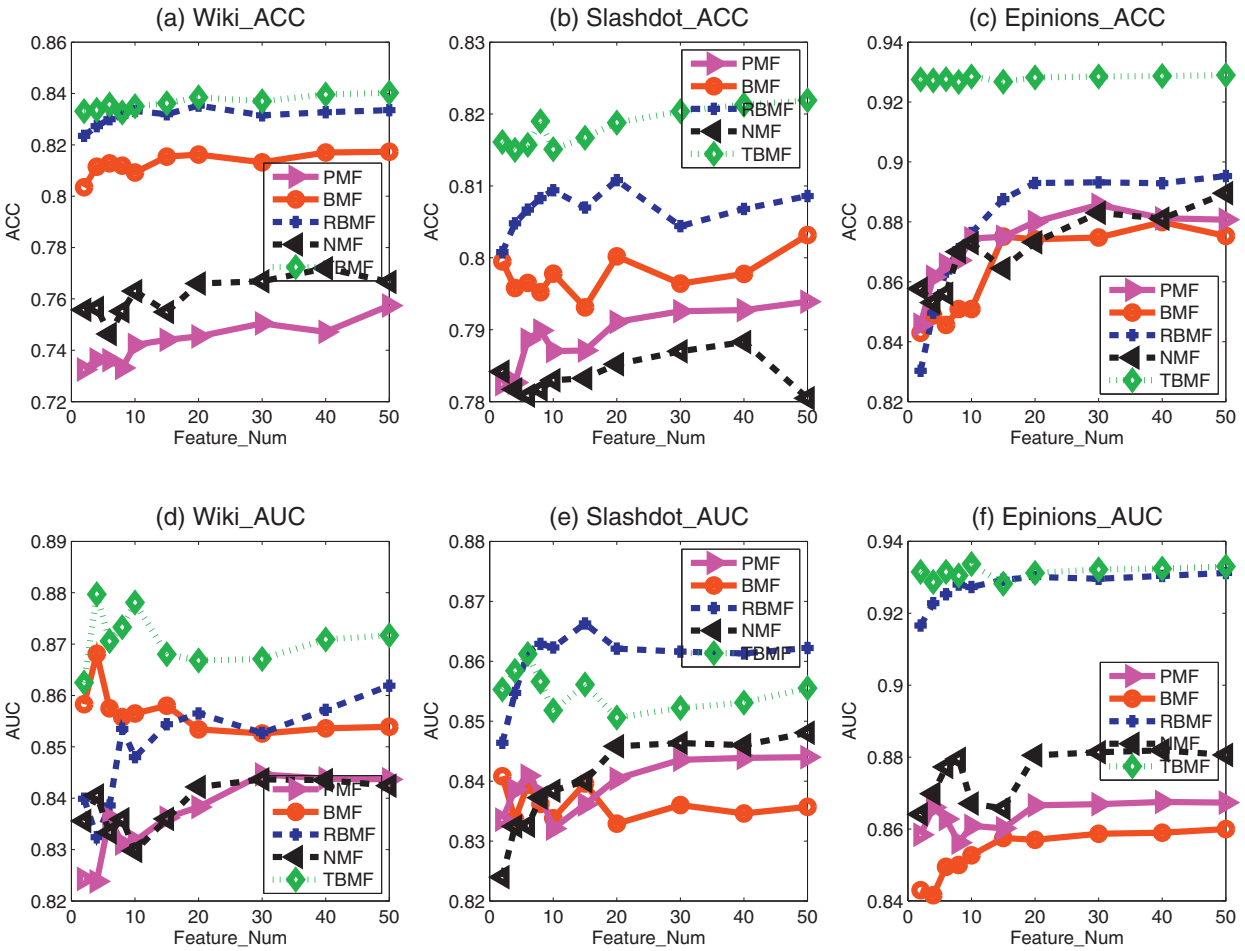

Fig. 4. Effects of latent feature number on prediction performance.

link signs are few; that is, it can adapt to the cold start problem. Furthermore, users and items are adopted the multivariate Student-t prior instead of the commonly used multivariate Gaussian prior for the representation of latent features. Thus, TBMF can obtain the sparse representation and discriminate between the informative and non-informative latent features. More importantly, a small number of latent features are sufficient for TBMF to achieve good prediction performance compared to other baseline models, which has great benefit in reducing computational complexity.

References

[1] P. Agrawal, V. Garg, R. Narayanam, Link label prediction in signed social networks., in: Proceedings of the 23th International Joint Conference on Artificial Intelligence, 2013, pp. 2591–2597.
[2] C. Archambeau, M. Verleysen, Robust Bayesian clustering, Neural Netw. 20 (1) (2007) 129–138.
[3] H. Attias, A variational Baysian framework for graphical models, Advances in Neural Information Processing Systems, 2000, pp. 209–215.
[4] A. Buchanan, A. Fitzgibbon, Damped newton algorithms for matrix factorization with missing data, in: IEEE Conference on Computer Vision and Pattern Recognition, 2, 2005, pp. 316–322.
[5] A.T. Cemgil, Bayesian inference for nonnegative matrix factorisation models, Comput. Intell. Neurosci. 2009 (2009).
[6] B. Chen, L. Chen, B. Li, A fast algorithm for predicting links to nodes of interest, Inf. Sci. 329 (2016) 552–567.
[7] K. Chiang, N. Natarajan, A. Tewari, I. Dhillon, Exploiting longer cycles for link prediction in signed networks, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, 2011, pp. 1157–1162.
[8] N. Ding, Y. Qi, S. Vishwanathan, t-divergence based approximate inference, in: Advances in Neural Information Processing Systems, 2011, pp. 1494–1502.
[9] N. Ding, S. Vishwanathan, t-logistic regression, in: Advances in Neural Information Processing Systems, 2010, pp. 514–522.
[10] C. Forbes, M. Evans, N. Hastings, B. Peacock, Statistical Distributions, John Wiley & Sons, 2011.
[11] W. Gilks, S. Richardson, D. Spiegelhalter, Markov Chain Monte Carlo in Practice, CRC Press, 1995.
[12] C. Hsieh, K. Chiang, I. Dhillon, Low rank modeling of signed networks, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, pp. 507–515.
[13] J. Kunegis, A. Lommatzsch, C. Bauckhage, The slashdot zoo: mining a social network with negative edges, in: Proceedings of the 18th International Conference on World Wide Web, 2009, pp. 741–750.
[14] B. Lakshminarayanan, G. Bouchard, C. Archambeau, Robust bayesian matrix factorisation., in: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, 2011, pp. 425–433.

[15] J. Leskovec, D. Huttenlocher, J. Kleinberg, Predicting positive and negative links in online social networks, in: Proceedings of the 19th International Conference on World Wide Web, 2010, pp. 641–650.
[16] J. Leskovec, D. Huttenlocher, J. Kleinberg, Signed networks in social media, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2010, pp. 1361–1370.
[17] Y.J. Lim, Y.W. Teh, Variational bayesian approach to movie rating prediction, in: Proceedings of KDD Cup and Workshop, vol. 7, 2007, pp. 15–21.
[18] J. Liu, C. Wu, W. Liu, Bayesian probabilistic matrix factorization with social relations and item contents for recommendation, Decis. Support. Syst. 55 (3) (2013) 838–850.
[19] Z. Ma, A.E. Teschendorff, A. Leijon, Y. Qiao, H. Zhang, J. Guo, Variational Bayesian matrix factorization for bounded support data, IEEE Trans. Pattern Anal. Mach. Intell. 37 (4) (2015) 876–889.
[20] A.K. Menon, C. Elkan, Link prediction via matrix factorization, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2011, pp. 437–452.
[21] C.E. Metz, Basic principles of roc analysis., Semin. Nucl. Med. 8 (4) (1978) 283–298.
[22] S. Nakajima, M. Sugiyama, Theoretical analysis of Bayesian matrix factorization, J Mach. Learn. Res. 12 (September) (2011) 2583–2648.
[23] S. Nakajima, M. Sugiyama, S.D. Babacan, R. Tomioka, Global analytic solution of fully-observed variational Bayesian matrix factorization, J Mach. Learn. Res. 14 (January) (2013) 1–37.
[24] J. Paisley, D. Blei, M.I. Jordan, Bayesian nonnegative matrix factorization with stochastic variational inference, Handbook of Mixed Membership Models and Their Applications, Chapman and Hall/CRC, 2014.
[25] I. Porteous, A. Asuncion, M. Welling, Bayesian matrix factorization with side information and Dirichlet process mixtures, in: Proceedings of the 24th AAAI Conference on Artificial Intelligence, 2010, pp. 563–568.
[26] J. Rennie, N. Srebro, Fast maximum margin matrix factorization for collaborative prediction, in: Proceedings of the 22nd International Conference on Machine Learning, 2005, pp. 713–719.
[27] D. Romik, Stirling'S approximation for n!: the ultimate short proof? Am. Math. Mon. 107 (6) (2000) 556.
[28] C. Röver, Student-t based filter for robust signal detection, Phys. Rev. D 84 (12) (2011) 122004.
[29] R. Salakhutdinov, A. Mnih, Bayesian probabilistic matrix factorization using Markov Chain Monte Carlo, in: Proceedings of the 25th International Conference on Machine Learning, 2008, pp. 880–887.
[30] R. Salakhutdinov, A. Mnih, Probabilistic matrix factorization, in: Advances in Neural Information Processing Systems, 2008, pp. 1257–1264.
[31] M.N. Schmidt, H. Laurberg, Nonnegative matrix factorization with Gaussian process priors, Comput. Intell. Neurosci. 2008 (2008) 3.
[32] M.N. Schmidt, O. Winther, L.K. Hansen, Bayesian non-negative matrix factorization, in: International Conference on Independent Component Analysis and Signal Separation, 2009, pp. 540–547.
[33] M.W. Seeger, D.P. Wipf, Variational Bayesian inference techniques, IEEE Signal Process. Mag. 27 (6) (2010) 81–91.
[34] H. Shan, A. Banerjee, Generalized probabilistic matrix factorizations for collaborative filtering, in: 2010 IEEE International Conference on Data Mining, 2010, pp. 1025–1030.
[35] D. Song, D.A. Meyer, M.R. Min, Fast nonnegative matrix factorization with rank-one ADMM, NIPS 2014 Workshop on Optimization for Machine Learning (OPT2014), 2014.
[36] P. Symeonidis, E. Tiakas, Y. Manolopoulos, Transitive node similarity for link prediction in social networks with positive and negative links, in: Proceedings of the 4th ACM Conference on Recommender Systems, 2010, pp. 183–190.
[37] Q. Tang, Y. Wang, S.-T. Xia, Student-t process regression with dependent student-t noise, in: Proceedings of the 22nd European Conference on Artificial Intelligence, 2016, pp. 82–89, doi:10.3233/978-1-61499-672-9-82.
[38] D.G. Tzikas, A.C. Likas, N.P. Galatsanos, The variational approximation for Bayesian inference, IEEE Sig. Process. Mag. 25 (6) (2008) 131–146.
[39] M.J. Wainwright, M.I. Jordan, Graphical models, exponential families, and variational inference, Found. Trends Mach. Learn. 1 (1–2) (2008) 1–305.
[40] Y. Wang, S. Romano, V. Nguyen, J. Bailey, X. Ma, S.-T. Xia, Unbiased multivariate correlation analysis, in: Proceedings of the 31th AAAI Conference on Artificial Intelligence, 2017, pp. 2754–2760.
[41] Y. Wang, S.-T. Xia, A novel feature subspace selection method in random forests for high dimensional data, in: 2016 IEEE International Joint Conference on Neural Networks, 2016, pp. 4383–4389.
[42] Y. Wang, S.-T. Xia, J. Wu, A less-greedy two-term tsallis entropy information metric approach for decision tree classification, Knowl. Based Syst. 120 (2017) 34–42.
[43] W. Wiegerinck, Variational approximations between mean field theory and the junction tree algorithm, in: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, 2000, pp. 626–633.
[44] J. Wu, S. Pan, X. Zhu, Z. Cai, Boosting for multi-graph classification, IEEE Trans. Cybern. 45 (3) (2015) 416–429, doi:10.1109/TCYB.2014.2327111.
[45] J. Wu, S. Pan, X. Zhu, C. Zhang, X. Wu, Positive and unlabeled multi-graph learning, IEEE Trans. Cybern. 47 (4) (2017) 818–829.
[46] J. Wu, S. Pan, X. Zhu, C. Zhang, P.S. Yu, Multiple structure-view learning for graph classification, IEEE Trans. Neural Netw. Learn. Syst. PP (99) (2017) 1–15.
[47] J. Wu, S. Pan, X. Zhu, P. Zhang, C. Zhang, Sode: self-adaptive one-dependence estimators for classification, Pattern Recognit. 51 (2016) 358–377. http://dx.doi.org/10.1016/j.patcog.2015.08.023.
[48] A.K. Young, P. Rasik, A trust prediction framework in rating-based experience sharing social networks without a web of trust, Inf. Sci. 191 (2012) 128–145.
[49] F. Zhou, J.R. Jiao, B. Lei, A linear threshold-hurdle model for product adoption prediction incorporating social network effects, Inf. Sci. 307 (2015) 95–109.
[50] J. Zhu, Z. Ge, Z. Song, HMM-driven robust probabilistic principal component analyzer for dynamic process fault classification, IEEE Trans. Ind. Electron. 62 (6) (2015) 3814–3821.