Fast Max-Margin Matrix Factorization with Data Augmentation

Minjie Xu XUJ-10@MAILS.TSINGHUA.EDU.CN
Jun Zhu DCSZJ@MAIL.TSINGHUA.EDU.CN
Bo Zhang DCSZB@MAIL.TSINGHUA.EDU.CN
Dept. of Comp. Sci. & Tech., LITS Lab, TNList Lab, Tsinghua University, Beijing 100084, China

Abstract

Existing max-margin matrix factorization (M^3F) methods either are computationally inefficient or need a model selection procedure to determine the number of latent factors. In this paper we present a probabilistic M^3F model that admits a highly efficient Gibbs sampling algorithm through data augmentation. We further extend our approach to incorporate Bayesian nonparametrics and build accordingly a truncation-free nonparametric M^3F model where the number of latent factors is literally unbounded and inferred from data. Empirical studies on two large real-world data sets verify the efficacy of our proposed methods.

1. Introduction

Matrix factorization has been a key technique in learning latent factor models for many applications such as collaborative prediction (Srebro et al., 2005; Salakhutdinov & Mnih, 2008; Zhou et al., 2010). Given a user-item preference matrix $Y \in \mathbb{R}^{N\times M}$, which is partially observed and usually sparse, matrix factorization aims to find a low-rank matrix $X \in \mathbb{R}^{N\times M}$ that simultaneously approximates the observed entries of Y under some loss measure (e.g., the commonly used squared error) and reconstructs the missing entries. Max-margin matrix factorization (M^3F) (Srebro et al., 2005) extends the model by adopting hinge loss, which is applicable to binary, discrete ordinal, or categorical data that are typical for a preference system, and a sparsity-inducing norm regularizer. For the binary case where $Y_{ij} \in \{\pm 1\}$ and one predicts by $\widehat{Y}_{ij} = \text{sign}(X_{ij})$, the optimization problem of M^3F is defined as

$$\min_X \|X\|_* + C \sum_{ij\in\mathcal{I}} h\left(Y_{ij}X_{ij}\right), \tag{1}$$

where $\|X\|_*$ is the nuclear norm of X, $\mathcal{I}$ is the indices of the observed entries and $h(x) \triangleq \max(0, 1-x)$ is the hinge loss. Problem (1) can be equivalently formulated as a semi-definite program (SDP) and learned by standard SDP solvers, but it is unfortunately very slow and scales to only thousands of users and items.

An alternative M^3F model based on a variational formulation of the nuclear norm is then proposed in (Rennie & Srebro, 2005) and it solves an equivalent problem on the factorized form $X = UV^\top$ instead:

$$\min_{U,V} \frac{1}{2}\left(\|U\|_F^2 + \|V\|_F^2\right) + C \sum_{ij\in\mathcal{I}} h\left(Y_{ij}U_iV_j^\top\right), \tag{2}$$

where $U \in \mathbb{R}^{N\times K}$ and $V \in \mathbb{R}^{M\times K}$ are interpreted as the user coefficient matrix and the item factor matrix respectively, and K is the number of latent factors. We use U_i to denote the ith row of U, and V_j likewise. By replacing hinge loss with a smooth surrogate, a gradient descent solver has been developed and it scales to millions of users and items. However the solver works with a truncated problem where K is pre-specified. Alternatively, (Xu et al., 2012) suggests a block-wise coordinate descent algorithm that directly works with hinge loss by use of SVMs for each row of U and V. This method achieves comparable accuracy with the gradient descent solver yet is almost as time-consuming. (Xu et al., 2012) also introduces a nonparametric model for M^3F which automatically resolves the unknown number of latent factors. However both its dependency on a truncation level for practical inference and the time-consuming SVM steps therein call for further improvement.

In this paper, we present a novel interpretation of the M^3F problem, that is to formalize the deterministic regularized risk minimization (RRM) problem (2) as an equivalent *maximum a posteriori* (MAP) estimation problem, and by use of the data augmentation

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

techniques recently developed for SVMs (Polson & Scott, 2011), we are able to perform simple yet highly efficient MCMC sampling and thus drastically increase the efficiency of solving M^{3}F problems. Furthermore, to bypass the model selection issue of the *factorized* M^{3}F models (i.e. selecting the number of latent factors K), we extend our probabilistic formulation to incorporate Bayesian nonparametrics and build thereupon a nonparametric M^{3}F model, which again enjoys a simple and efficient MCMC sampling algorithm. Compared with the previous nonparametric M^{3}F (Xu et al., 2012), which resorts to variational approximation with truncated mean-field assumptions, our sampling algorithm is both assumption-free and truncation-free.

The paper is structured as follows. Section 2 formulates M^{3}F as a MAP estimation problem and presents the MCMC sampling algorithm via data augmentation; Section 3 presents the nonparametric M^{3}F extension; Section 4 presents empirical results on two prevalent collaborative filtering data sets and demonstrate efficiency improvement. Finally, Section 5 concludes.

2. A Probabilistic Formulation of M^{3}F

We start with a discussion on the generic formulation of RRM as MAP estimation.

2.1. RRM as MAP, A New Look

Given a set of training data $\mathcal{X} = \{\mathcal{X}_n\}_{n=1}^N$, many machine learning problems, including M^{3}F, can be cast as solving a RRM problem generally written as

$$\min_{\mathcal{M}} \Omega(\mathcal{M}) + C \sum_{n=1}^{N} \mathcal{R}(\mathcal{M}; \mathcal{X}_n) \tag{3}$$

where we denote the model (parameters) by $\mathcal{M}$; $\Omega(\mathcal{M})$ is the regularizer which is critical to save the model from over-fitting; $\sum_{n=1}^N \mathcal{R}(\mathcal{M}; \mathcal{X}_n)$ is the empirical loss; and C is the balancing factor, or regularization constant. Normally for supervised tasks where training labels are available, $\mathcal{X}_n = (\mathbf{x}_n, y_n)$ and

$$\mathcal{R}(\mathcal{M}; \mathcal{X}_n) = L(y_n, f(\mathcal{M}; \mathbf{x}_n)) \tag{4}$$

where $f(\mathcal{M}; \mathbf{x})$ is termed the *discriminant function*, which gives a prediction score s, and $L(y, s)$ the *loss function*.

Generally RRM is a deterministic optimization problem without any resort to a probabilistic background. For example, in the case of M^{3}F for binary data, we have $\mathcal{M} = (U, V)$, $\mathcal{X} = \{((i,j), Y_{ij}) | ij \in \mathcal{I}\}$, and

$$\begin{aligned} \Omega(U,V) &= \frac{1}{2}\left(\|U\|_F^2 + \|V\|_F^2\right) \\ s = f(U,V;(i,j)) &= U_i V_j^\top,\ L(Y_{ij}, s) = h(Y_{ij}s) \end{aligned} \tag{5}$$

In contrast, MAP estimation is backed up by Bayesian inference methodology and, given a prior distribution $p_0(\mathcal{M})$ and a likelihood term $\mathcal{L}(\mathcal{M}|\mathcal{X}) \triangleq p(\mathcal{X}|\mathcal{M})$[1], solves for the optimal model by maximizing the posterior distribution $p(\mathcal{M}|\mathcal{X}) \propto p_0(\mathcal{M})\mathcal{L}(\mathcal{M}|\mathcal{X})$. Quite often we adopt the i.i.d. assumption on the data generation process so that the likelihood factorizes as $p(\mathcal{X}|\mathcal{M}) = \prod_{n=1}^N p(\mathcal{X}_n|\mathcal{M})$ and hence, by denoting $\mathcal{L}(\mathcal{M}|\mathcal{X}_n) \triangleq p(\mathcal{X}_n|\mathcal{M})$, the problem reads

$$\max_{\mathcal{M}} p_0(\mathcal{M}) \prod_{n=1}^{N} \mathcal{L}(\mathcal{M}|\mathcal{X}_n). \tag{6}$$

To better disclose the correspondence between RRM (3) and MAP (6), we introduce two new concepts, namely the *delegate prior* and the *delegate likelihood* (abbreviated as dele-prior and dele-likelihood in the sequel). Given a prior-likelihood pair $(p_0, \mathcal{L})$, a dele-prior $\dot{p}_0(\mathcal{M})$ can be any non-negative function defined solely on $\mathcal{M}$ while a dele-likelihood $\dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_n)$ can be any non-negative function defined jointly on $\mathcal{X}_n$ and $\mathcal{M}$ (although viewed as a function merely of $\mathcal{M}$ just as is $\mathcal{L}$), as long as the original "genuine" prior-likelihood pair can be uniquely recovered as

$$\mathcal{L}(\mathcal{M}|\mathcal{X}_n) = \frac{\dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_n)}{\zeta_n(\mathcal{M})},\ p_0(\mathcal{M}) = \frac{1}{\zeta_0}\dot{p}_0(\mathcal{M}) \prod_{n=1}^{N} \zeta_n(\mathcal{M}),$$

where the normalizing factors are $\zeta_n(\mathcal{M}) \triangleq \int \dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_n) d\mathcal{X}_n$ and $\zeta_0 \triangleq \int \dot{p}_0(\mathcal{M}) \prod_{n=1}^N \zeta_n(\mathcal{M}) d\mathcal{M}$.

Note that although $\dot{\mathcal{L}}$ and $\mathcal{L}$ appear only different by a normalizing factor when viewed as a probability of $\mathcal{X}_n$, they can be completely different when viewed as a function of $\mathcal{M}$. As for $\dot{p}_0$ and p_0, since scaling $\dot{p}_0$ by any positive constant carries no effect on the resulting genuine prior p_0, we can normalize $\dot{p}_0$ if necessary.

For any qualified delegate prior-likelihood pair $(\dot{p}_0, \dot{\mathcal{L}})$ and its uniquely induced genuine pair $(p_0, \mathcal{L})$, or the other way round, for any genuine pair and all of its compatible delegate pairs, the MAP problem (6) remains intact since

$$p_0(\mathcal{M}) \prod_{n=1}^{N} \mathcal{L}(\mathcal{M}|\mathcal{X}_n) \propto \dot{p}_0(\mathcal{M}) \prod_{n=1}^{N} \dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_n).$$

The delegate prior-likelihood pair provides an alternative, presumably easier way to factorize the posterior since the dele-likelihood does not necessarily comply with the normalization constraint for $\mathcal{X}_n$. Now we may easily convert the RRM problem (3) into MAP estimation by setting the delegate pair as

$$\dot{p}_0(\mathcal{M}) = e^{-\Omega(\mathcal{M})},\ \dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_n) = e^{-C\mathcal{R}(\mathcal{M};\mathcal{X}_n)} \tag{7}$$

[1] We consider generative models here for the ease of presentation. For discriminative models, the likelihood becomes $p(\{y_n\}|\mathcal{M}, \{\mathbf{x}_n\})$ but our discussion applies as well.

and solving the following delegate form of MAP

$$\max_{\mathcal{M}} \dot{p}_0(\mathcal{M}) \prod_{n=1}^{N} \dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_n). \tag{8}$$

Directly solving problem (8) does not save us any effort. Yet as we shall demonstrate below, given a properly augmented representation, we may resort to existing probabilistic techniques, e.g., MCMC sampling methods, to perform highly efficient approximate inference that achieves comparable or even better performances. Moreover, this probabilistic interpretation naturally leads to a nonparametric Bayesian extension of the model in consideration.

Before a full exposition of the new algorithms for M^{3}F and its nonparametric extension, we provide more insights on this probabilistic interpretation. Specifically, given a prior-likelihood pair $(p_0, \mathcal{L})$, we can calculate the posterior distribution according to Bayes' theorem, or alternatively by solving a functional optimization problem (Zhu et al., 2013b)

$$\min_{q(\mathcal{M})\in\mathbb{P}} \mathrm{KL}(q(\mathcal{M})\|p_0(\mathcal{M})) - \sum_{n=1}^{N} \mathbb{E}_q[\log \mathcal{L}(\mathcal{M}|\mathcal{X}_n)], \tag{9}$$

where $\mathbb{P}$ is a space of valid probability distributions; $\mathrm{KL}(q(\mathcal{M})\|p_0(\mathcal{M}))$ is the KL-divergence; and $\mathbb{E}_q[\log \mathcal{L}(\mathcal{M}|\mathcal{X}_n)] = \int \log \mathcal{L}(\mathcal{M}|\mathcal{X}_n) q(\mathcal{M}) d\mathcal{M}$. Note that to distinguish from the posterior $p(\mathcal{M}|\mathcal{X})$ by Bayes' rule, we use $q(\mathcal{M})$ to represent a posterior distribution derived from our generic inference procedure.

By use of any compatible delegate pair, we rewrite problem (9) in an equivalent yet more general way as[2]

$$\min_{q(\mathcal{M})\in\mathbb{P}} \mathrm{KL}(q(\mathcal{M})\|\dot{p}_0(\mathcal{M})) - \sum_{n=1}^{N} \mathbb{E}_q[\log \dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_n)] \tag{10}$$

by ignoring a constant term $\log \zeta_0$. Now if we substitute the RRM-induced delegate pair (7) into (10) and denote $\widehat{\Omega}(q(\mathcal{M})) \triangleq \mathrm{KL}(q(\mathcal{M})\|\hat{p}_0(\mathcal{M}))$ where $\hat{p}_0(\mathcal{M}) \triangleq e^{-\Omega(\mathcal{M})}/Z$ is the normalized dele-prior and $\widehat{\mathcal{R}}(q(\mathcal{M}); \mathcal{X}_n) \triangleq \mathbb{E}_q[\mathcal{R}(\mathcal{M}; \mathcal{X}_n)]$, we obtain the following equivalence

$$\begin{aligned} &\operatorname*{argmin}_{\mathcal{M}} \quad \Omega(\mathcal{M}) + C \sum_{n=1}^{N} \mathcal{R}(\mathcal{M}; \mathcal{X}_n) \\ = &\operatorname*{argmax}_{\mathcal{M}} \operatorname*{argmin}_{q(\mathcal{M})\in\mathbb{P}} \widehat{\Omega}(q(\mathcal{M})) + C \sum_{n=1}^{N} \widehat{\mathcal{R}}(q(\mathcal{M}); \mathcal{X}_n). \end{aligned} \tag{11}$$

The significance of this alternative representation is that it inspires an interesting observation. That is, given a properly defined deterministic RRM problem on $(\mathcal{M}, \mathcal{X})$, we can actually solve it in two successive phases, the first one seeking an optimal distribution $\tilde{q}(\mathcal{M})$ by solving an induced functional minimization problem defined on $(q(\mathcal{M}), \mathcal{X})$, and the second finding thereupon an optimal point estimate by reading out the most probable model according to $\tilde{q}(\mathcal{M})$.

When viewed individually, the first phase itself naturally suggests a *probabilistic* extension to the original *deterministic* risk minimization problem $(\Omega, \mathcal{R})$:

$$\min_{q(\mathcal{M})\in\mathbb{P}} \mathrm{KL}(q(\mathcal{M})\|\pi(\mathcal{M})) + C \sum_{n=1}^{N} \mathbb{E}_q[\mathcal{R}(\mathcal{M}; \mathcal{X}_n)] \tag{12}$$

where we can set $\pi(\mathcal{M}) = \hat{p}_0(\mathcal{M})$ to retain the equivalence (11), or specify $\pi(\mathcal{M})$ to be any other proper distribution that we believe serves a good regularizer.

2.2. A Probabilistic Formulation of M^{3}F

We now apply the above generic discussions to the specific case of M^{3}F. For M^{3}F with binary preference scores, it suffices to substitute the definition of $(\Omega, \mathcal{R})$ (5) into the RRM-induced delegate pair (7). Below we concentrate on the more common case of M^{3}F with ordinal ratings, where $Y_{ij} \in \{1, 2, \ldots, L\}$.

As in (Srebro et al., 2005), we introduce thresholds $\theta_0 \le \theta_1 \le \cdots \le \theta_{L-1}$, where $\theta_0 = -\infty$, to discretize $\mathbb{R}$ into L intervals. Hence the model is updated as $\mathcal{M} = (U, V, \boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{L-1})^\top$ and the prediction rule is changed accordingly to $\widehat{Y}_{ij} = \max\{r | U_i V_j^\top \ge \theta_{r-1}, r = 1, \ldots, L\}$. For hard-margin, we would require $\theta_{Y_{ij}-1} + \ell \le U_i V_j^\top \le \theta_{Y_{ij}} - \ell$, while in a soft-margin setting, we define the discriminant function and the loss function to be[3]

$$\mathbf{s} = f(U, V, \boldsymbol{\theta}; (i, j)) = \boldsymbol{\theta} - (U_i V_j^\top)\mathbf{1}_{L-1}, \tag{13}$$

$$L(Y_{ij}, \mathbf{s}) = \sum_{r=1}^{L-1} h_\ell(T_{ij}^r s_r) \tag{14}$$

where $T_{ij}^r \triangleq \begin{cases} +1 & \text{for } r \ge Y_{ij} \\ -1 & \text{for } r < Y_{ij} \end{cases}$ and $h_\ell(x) \triangleq \max(0, \ell - x)$ is the generalized hinge loss with margin parameter ℓ. When $\ell \ge 1$, the loss thus defined is an upper bound to the sum of absolute differences between the predicted ratings and the true ratings, a loss measure closely related to Normalized Mean Absolute Error (NMAE) (Marlin & Zemel, 2004; Srebro et al., 2005).

Furthermore, we can learn a more flexible model to capture users' diverse rating criteria by replacing user-common thresholds $\boldsymbol{\theta}$ with user-specific ones $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{i(L-1)})^\top$. And we may as well regularize these thresholds $\boldsymbol{\theta} = \boldsymbol{\theta}_{1:N}$ with

$$\Omega(\boldsymbol{\theta}) = \frac{1}{2\varsigma^2} \sum_{i=1}^{N} \|\boldsymbol{\theta}_i - \boldsymbol{\rho}\|_2^2 \tag{15}$$

[2] We assume a normalized $\dot{p}_0(\mathcal{M})$ here.

[3] There is a score for each of the $L-1$ thresholds and they collectively form the prediction score *vector* $\mathbf{s}$.

where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_{L-1})^\top$ and $\rho_1 < \cdots < \rho_{L-1}$ are specified as a prior guidance towards an ascending sequence of large-margin thresholds. Then the overall regularizer becomes $\Omega(\mathcal{M}) = \Omega(U, V) + \Omega(\boldsymbol{\theta})$. Note that when $\varsigma \to \infty$, the regularizer on $\boldsymbol{\theta}$ plays no part.

Regularizer and loss fully specified, we may again follow the generic discussions above and obtain the following probabilistic formulation[4]:

$$\dot{p}_0(U, V, \boldsymbol{\theta}) = e^{-\frac{1}{2\sigma^2}\left(\|U\|_F^2 + \|V\|_F^2\right) - \frac{1}{2\varsigma^2}\sum_{i=1}^N \|\boldsymbol{\theta}_i - \boldsymbol{\rho}\|_2^2}$$
$$\propto \prod_{i=1}^N \mathcal{N}(U_i|\mathbf{0}, \sigma^2 I)\mathcal{N}(\boldsymbol{\theta}_i|\boldsymbol{\rho}, \varsigma^2 I) \cdot \prod_{j=1}^M \mathcal{N}(V_j|\mathbf{0}, \sigma^2 I) \quad (16)$$

$$\dot{\mathcal{L}}(U, V, \boldsymbol{\theta}|(i, j), Y_{ij}) = e^{-C\sum_{r=1}^{L-1} h_\ell\left(T_{ij}^r(\theta_{ir} - U_i V_j^\top)\right)}. \quad (17)$$

2.3. Connections with Previous Methods

In the literature of PAC-Bayes learning theory, the loss (or risk) term in problem (12) corresponds to that of a *Gibbs* classifier (McAllester, 2003; Germain et al., 2009), which is a stochastic classifier that randomly chooses a classifier $\mathcal{M}$ according to $q(\mathcal{M})$ to classify a data sample $\mathbf{x}$. An alternative formulation of inferring a posterior distribution of classifiers that has received much attention is the one induced from an *expected* classifier, which can be generally written as

$$\min_{q(\mathcal{M})\in\mathbb{P}} \mathrm{KL}(q(\mathcal{M})\|\pi(\mathcal{M})) + C\sum_{n=1}^N L(y_n, \mathbb{E}_q[f(\mathcal{M}; \mathbf{x}_n)]). \quad (18)$$

Maximum entropy discrimination (MED) (Jaakkola et al., 1999) represents one such example where the loss function L is hinge loss. MED has been adopted in various max-margin models, including max-margin supervised topic models (Zhu et al., 2009) and the probabilistic formulation of max-margin matrix factorization (Xu et al., 2012). It's obvious that the two problems (12) and (18) only differ in their choice of the loss term, with (18) choosing *loss of expectation* while (12) *expectation of loss*. Actually we have

$$\mathbb{E}_q[L(y, f(\mathcal{M}; \mathbf{x}))] \geq L(y, \mathbb{E}_q[f(\mathcal{M}; \mathbf{x})]) \quad (19)$$

given that $L(\cdot, s)$ is a convex function, e.g., hinge loss, squared loss, the loss function of M^{3}F for ordinal ratings (14), etc. Therefore our new formulation gives a more relaxed model while at the same time is much easier to solve (say, through Bayes' theorem) compared with problem (18), for which approximate variational methods are very often required, along with additional assumptions on the posterior distribution (Zhu et al., 2009; Xu et al., 2012). Note that a Gibbs max-margin topic model has been presented in (Zhu et al., 2013a) with data augmentation; And our work differs by presenting a different viewpoint as detailed above and dealing with the challenging problem of matrix factorization.

[4] We introduce an additional parameter σ for the variance of U and V. It *is* superfluous here but will be useful later in the nonparametric M^{3}F model with IBP prior.

2.4. Data Augmentation for M^{3}F

We now present a simple and efficient algorithm for learning M^{3}F within its probabilistic formulation. Our algorithm builds on the statistical idea of data augmentation (Tanner & Wong, 1987; van Dyk & Meng, 2001), whose general principle is to introduce auxiliary variables so as to facilitate Bayesian inference on the original variables of interest.

Specifically in our case, the form of the dele-likelihood $\dot{\mathcal{L}}$ (17) is very hard to manipulate due to the "max" operator inherited from the hinge loss thereof. Fortunately, it is discovered in (Polson & Scott, 2011) that $e^{-2\max(u,0)}$ enjoys the representation as a location-scale mixture of Gaussians, namely

$$e^{-2\max(u,0)} = \int_0^\infty \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{(u+\lambda)^2}{2\lambda}} d\lambda = \int_0^\infty \phi(u|-\lambda, \lambda) d\lambda$$

where $\phi(\mu|\cdot, \cdot)$ is the normal density function. This enables us to augment the original model $\mathcal{M} = (U, V, \boldsymbol{\theta})$ by introducing auxiliary variables $\boldsymbol{\lambda}$ likewise.

Specifically, let $\Delta_{ij}^r \triangleq \frac{C}{2}(\ell - T_{ij}^r(\theta_{ir} - U_i V_j^\top))$, $D = \mathbb{R}_+^{L-1}$, $\boldsymbol{\Delta}_{ij} = (\Delta_{ij}^1, \ldots, \Delta_{ij}^{L-1})^\top$ and $\boldsymbol{\lambda}_{ij} = (\lambda_{ij1}, \ldots, \lambda_{ij(L-1)})^\top$. Then for each delegate likelihood (17), we have $\dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_{ij}) = \prod_{r=1}^{L-1} \exp\{-2\max(\Delta_{ij}^r, 0)\}$ and therefore,

$$\dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_{ij}) = \int_D \phi(\boldsymbol{\Delta}_{ij}| - \boldsymbol{\lambda}_{ij}, \mathrm{diag}(\boldsymbol{\lambda}_{ij})) d\boldsymbol{\lambda}_{ij} \quad (20)$$

Eq. (20) suggests an augmented model $\mathcal{M}' = (\mathcal{M}, \boldsymbol{\lambda})$ with posterior $q(\mathcal{M}') \propto \dot{p}_0(\mathcal{M}) \prod_{ij\in\mathcal{I}} \dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_{ij}, \boldsymbol{\lambda}_{ij})$ where

$$\dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_{ij}, \boldsymbol{\lambda}_{ij}) \triangleq \phi(\boldsymbol{\Delta}_{ij}| - \boldsymbol{\lambda}_{ij}, \mathrm{diag}(\boldsymbol{\lambda}_{ij})). \quad (21)$$

Note that $\dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_{ij}, \boldsymbol{\lambda}_{ij}) = \zeta_{ij}(\mathcal{M}) p(\mathcal{X}_{ij}, \boldsymbol{\lambda}_{ij}|\mathcal{M})$ and thus is *not*, by definition, a valid "dele-likelihood" $\dot{\mathcal{L}}(\mathcal{M}'|\mathcal{X}_{ij})$ for the augmented model $\mathcal{M}'$.

This augmented representation favors Gibbs sampling in that the Gaussian form of $\dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_{ij}, \boldsymbol{\lambda}_{ij})$ appears "conjugate" to the Gaussian delegate prior $\dot{p}_0(\mathcal{M})$ (16) with respect to U_i, V_j and $\boldsymbol{\theta}_i$ individually, and furthermore, each auxiliary variable λ_{ijr} in its inverse can be shown to follow an inverse Gaussian (Polson & Scott, 2011) for its conditional distribution, hence implying simple conditional distributions. Below, we summarize the conditional distributions and the derivation is similar to that for SVMs (Polson & Scott, 2011).

For auxiliary variables, λ_{ijr} are conditionally independent of each other given $(\mathcal{M}, \mathcal{X})$ and we have

$$q(\lambda_{ijr}^{-1}|\mathcal{M}' \setminus \boldsymbol{\lambda}) = \mathcal{IG}(|\Delta_{ij}^r|^{-1}, 1), \tag{22}$$

an inverse Gaussian distribution from which samples can be efficiently drawn.

For item factors, V_j are also conditionally independent and $q(V_j|\mathcal{M}' \setminus V) = \mathcal{N}(b_j, B_j)$ where

$$\begin{aligned} B_j^{-1} &= \frac{1}{\sigma^2} I + \sum_{i|ij\in\mathcal{I}} \left(\frac{C^2}{4} \sum_{r=1}^{L-1} \frac{1}{\lambda_{ijr}} \right) U_i^\top U_i \\ b_j &= -B_j \sum_{i|ij\in\mathcal{I}} \sum_{r=1}^{L-1} \left(\frac{C}{2} T_{ij}^r + \frac{C^2}{4} \frac{T_{ij}^r \ell - \theta_{ir}}{\lambda_{ijr}} \right) U_i \end{aligned} \tag{23}$$

Similar results apply to user factors U_i and are omitted to save space.

Finally, thresholds θ_{ir} are again conditionally independent and $q(\theta_{ir}|\mathcal{M}' \setminus \boldsymbol{\theta}) = \mathcal{N}(a_{ir}, A_{ir})$ where

$$\begin{aligned} A_{ir}^{-1} &= \frac{1}{\varsigma^2} + \frac{C^2}{4} \sum_{j|ij\in\mathcal{I}} \frac{1}{\lambda_{ijr}} \\ a_{ir} &= A_{ir} \left(\frac{\rho_r}{\varsigma^2} + \sum_{j|ij\in\mathcal{I}} \left(\frac{C}{2} T_{ij}^r + \frac{C^2}{4} \frac{T_{ij}^r \ell + U_i V_j^\top}{\lambda_{ijr}} \right) \right) \end{aligned} \tag{24}$$

With the above conditional distributions, we can develop a Gibbs sampling algorithm for the augmented model $q(\mathcal{M}, \boldsymbol{\lambda})$ by alternately drawing samples from each of the conditional distributions with random initialization. By ignoring $\boldsymbol{\lambda}$ we implicitly obtain the target posterior $q(\mathcal{M})$.

3. A Nonparametric M^{3}F with IBP Prior and Data Augmentation

Solving M^{3}F for U and V (2) instead of directly for X (1) has resulted in much more scalable methods (Rennie & Srebro, 2005; Xu et al., 2012). One resulting problem nevertheless, is to explicitly handle the latent factor dimension, i.e. the number of columns, K, of the two matrices. A typical solution relies on some general model selection procedure, e.g., cross-validation, which enumerates and compares many candidate models with different values of K and thus can be computationally expensive.

To solve this problem, (Xu et al., 2012) introduces a probabilistic model for M^{3}F that is induced from *expected* classifiers (18) and built accordingly a nonparametric M^{3}F model termed infinite probabilistic M^{3}F (iPM^{3}F) which automatically resolves the unknown number of latent factors. However their formulation was rooted in the MED framework and consequently resorted to a complicated approximate variational learning algorithm with mean-field assumptions. In order for a practical solution, (Xu et al., 2012) further set an upper bound, namely the truncation level, to the number of latent factors. Both the mean-field assumptions and the truncation level introduce extra bias into the posterior inference. And what's more, it requires some domain knowledge to properly set the truncation level: a higher level indicates more parameters and thus more time for solution while a lower level puts model-complexity sufficiency at risk and is prone to hamper the model's "infinite" flexibility.

Below, we propose an alternative nonparametric Bayesian M^{3}F model (termed Gibbs iPM^{3}F) by adopting the probabilistic formulation induced from *Gibbs* classifiers (12) instead. Again, by use of data augmentation, we design efficient Gibbs sampling algorithms which is both assumption-free and truncation-free.

3.1. Gibbs iPM^{3}F

Unlike the parametric Gibbs M^{3}F which is induced from a deterministic RRM problem, we directly build our nonparametric model from a probabilistic setting (12). Specifically, we reuse the empirical loss as defined by Eq. (13) and (14) since they naturally fit here. While for the dele-prior $\pi(\mathcal{M})$, it should not only be flexible enough to allow Bayesian inference on factor matrices with an *unbounded* number of columns, but, what's even more important, be favorable to sparse matrices as well so that only a finite number of features would be "active" for any finite data set.

The Indian buffet process (IBP) (Griffiths & Ghahramani, 2005) appears to feed our need for this case. Think of a binary matrix Z as recoding customers' behavior of sampling dishes from an infinite long buffet. Then IBP specifies a stochastic process that generates binary matrices Z as follows:

1. The first customer samples the first Poisson(α) number of dishes;
2. The ith customer first samples dishes that have already been taken, according to their popularity m_k/i where m_k is the number of previous customers who have sampled that dish; then he tries a Poisson(α/i) number of new dishes.

The process above induces a distribution for the *lof*-equivalent class of binary matrices. We denote this distribution by IBP(α) and define Gibbs iPM^{3}F to be solving problem (12) where we replace U by Z and specify the normalized dele-prior $\pi(\mathcal{M})$ as

$$\pi(Z, V, \theta) = \text{IBP}(Z|\alpha) \cdot \prod_{j=1}^{M} \mathcal{N}(V_j|\mathbf{0}, \sigma^2 I) \cdot \prod_{i=1}^{N} \mathcal{N}(\boldsymbol{\theta}_i|\boldsymbol{\rho}, \varsigma^2 I)$$

3.2. A Gibbs Sampling Algorithm

Since the dele-likelihood in Gibbs iPM^{3}F remains the same as Gibbs M^{3}F, it is expected that exactly the same data augmentation technique (20) can be applied here. Moreover, since the dele-prior of V and $\boldsymbol{\theta}$ are also simply reused from Gibbs M^{3}F, their conditional distributions would remain the same as in Eq. (23)&(24) given that we replace U_i with Z_i.

For the binary latent feature matrix Z, we follow the uncollapsed Gibbs sampling (Doshi-Velez et al., 2009) where V is not marginalized over but kept in the conditions. Specifically, for existing features, we have the conditional distribution

$$q(Z_{ik}|\mathcal{M}' \setminus Z_{ik}) \propto \pi(Z_{ik}|Z_{-(ik)}) \prod_{j|ij\in\mathcal{I}} \dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_{ij}, \boldsymbol{\lambda}_{ij}) \quad (25)$$

where $\pi(Z_{ik}|Z_{-(ik)}) = \text{Bernoulli}(\sum_{j\neq i} Z_{jk}/N)$ according to the *exchangeable* IBP and $\dot{\mathcal{L}}(\mathcal{M}|\mathcal{X}_{ij}, \boldsymbol{\lambda}_{ij})$ is just as defined in Eq. (21) with U replaced by Z.

While for new features $Z_i^\nu = \mathbf{1}_{k_i}^\top$, we equivalently sample $k_i \in \mathbb{Z}_{\geq 0}$ and adopt the partially collapsed sampler where the new latent features $V^{i\nu} \in \mathbb{R}^{M\times k_i}$ are integrated out and thus obtain

$$\begin{aligned} q(Z_i^\nu|\mathcal{M}') &= \int q(Z_i^\nu, V^{i\nu}|\mathcal{M}')dV^{i\nu} \\ &\propto \pi(Z_i^\nu|Z) \prod_{j|ij\in\mathcal{I}} \int \pi(V_j^{i\nu})\dot{\mathcal{L}}(\mathcal{M}, Z_i^\nu, V_j^{i\nu}|\mathcal{X}_{ij}, \boldsymbol{\lambda}_{ij})dV_j^{i\nu} \\ &\propto \text{Poisson}(k_i|\alpha/N) \prod_{j|ij\in\mathcal{I}} \frac{|\Sigma_{ijk_i}|^{1/2}}{\sigma^{k_i}} e^{\frac{1}{2}\boldsymbol{\omega}_{ijk_i}^\top \Sigma_{ijk_i}^{-1}\boldsymbol{\omega}_{ijk_i}} \end{aligned} \quad (26)$$

where $\Sigma_{ijk_i}^{-1} = \frac{1}{\sigma^2} I_{k_i\times k_i} + \big(\sum_{r=1}^{L-1} \frac{C^2}{4\lambda_{ijr}}\big)\mathbf{1}_{k_i\times k_i}$ and $\boldsymbol{\omega}_{ijk_i} = -\Big(\frac{C}{2}\sum_{r=1}^{L-1} T_{ij}^r\big(1+\frac{\Delta_{ij}^r}{\lambda_{ijr}}\big)\Big)\Sigma_{ijk_i}\mathbf{1}_{k_i}$.

Then conditioned on the newly sampled Z_i^ν (or k_i), we draw the corresponding new features $V^{i\nu}$

$$\begin{aligned} q(V_j^{i\nu}|\mathcal{M}', Z_i^\nu) &\propto \pi(V_j^{i\nu})\dot{\mathcal{L}}(\mathcal{M}, Z_i^\nu, V_j^{i\nu}|\mathcal{X}_{ij}, \boldsymbol{\lambda}_{ij})^{\mathbb{I}_{ij\in\mathcal{I}}} \\ &\propto \begin{cases} \mathcal{N}(\boldsymbol{\omega}_{ijk_i}, \Sigma_{ijk_i}), & ij \in \mathcal{I} \\ \mathcal{N}(\mathbf{0}, \sigma^2 I), & ij \notin \mathcal{I} \end{cases} \end{aligned} \quad (27)$$

4. Experiments and Discussions

We conduct experiments on the MovieLens 1M and the EachMovie data sets, and compare our results with M^{3}F (*Smooth Hinge*, truncated) (Rennie & Srebro, 2005), bcd M^{3}F("bcd" for "block-wise coordinate descent", truncated) (Xu et al., 2012) and iPM^{3}F (truncated-mean-field, infinite) (Xu et al., 2012).

Data sets: The MovieLens data set contains 1,000,209 anonymous ratings (ranging from 1 to 5) of 3,952 movies made by 6,040 users, among which 3,706 movies are actually rated and every user has at least 20 ratings. The EachMovie data set contains 2,811,983 ratings of 1,628 movies made by 72,916 users, among which 1,623 movies are actually rated and 36,656 users has at least 20 ratings. As in (Marlin & Zemel, 2004; Rennie & Srebro, 2005), we discarded users with fewer than 20 ratings, leaving us with 2,579,985 ratings. There are 6 possible rating values, $\{0, 0.2, \ldots, 1\}$ and we mapped them to $\{1, 2, \ldots, 6\}$.

Protocol: As in (Marlin & Zemel, 2004; Rennie & Srebro, 2005; Xu et al., 2012), we adopt the *all-but-one* protocol to construct training sets and test sets. And we consider both *weak* and *strong* generalization, where *weak* indicates *all* users contribute to the learning of the latent factors while *strong* transfers the learned movie latent factors from one group of users to another. As in previous methods, we randomly partition the users into 5,000 and 1,040 for *weak* and *strong* in MovieLens, and 30,000 and 6,565 in EachMovie. We repeat the random partition thrice, test our model against each of them and report the averaged Normalized Mean Absolute Error (NMAE).

Implementation details[5]**:** We perform cross-validation to choose the best regularization constant C from the same 11 candidate values that are log-evenly distributed between $0.1^{3/4}$ and 0.1^2 as in (Xu et al., 2012). According to (Rennie & Srebro, 2005), factor numbers higher than 50 yield similar performances and hence they choose $K = 100$ as a compromise between model capacity and computational complexity. Therefore we also set the truncation level K to be 100 for iPM^{3}F and all the parametric M^{3}F methods. Other hyper-parameters are set as follows: $\alpha = 3$, $\sigma = 1$, $\ell = 9$, $\varsigma = 1.5\ell$; $\rho_1, \ldots, \rho_{L-1}$ are set to be symmetric with respect to 0, with a step-size of 3ℓ.

Point estimate: We sought point estimate because our model formulation adopts a risk term that is induced from stochastic Gibbs classifiers. More specifically, we compared both the *single* samples $\mathcal{M}^{(m)}$ drawn from each Gibbs sampling iteration and the Rao-Blackwellizedly *averaged* samples

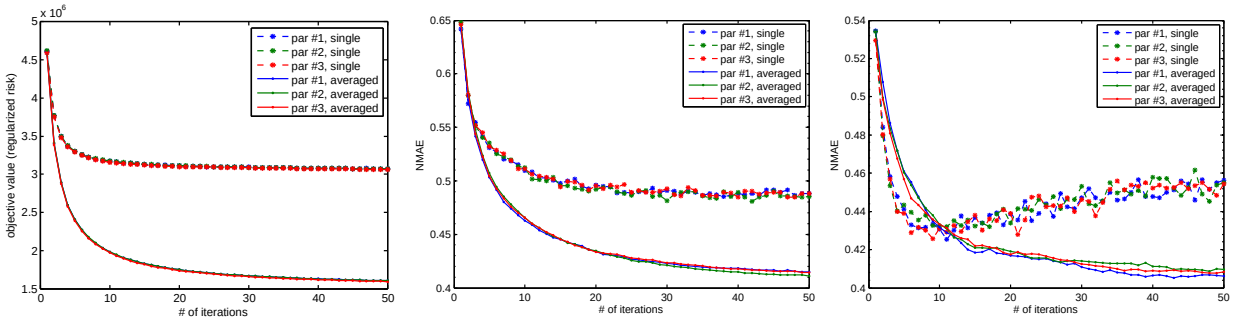
$$\bar{\mathcal{M}}^{(m)} = \frac{1}{m}\sum_{i=1}^{m} \mathcal{M}^{(i)}. \quad (28)$$

Fig. 1 illustrates the difference between these two estimates, where dashed curves represent single samples while solid curves represent averaged samples. It seems taking average of the samples not only stabilizes the results but also continuously reduces test error as well as the original objective value.

[5]The data sets and the implementation are available at https://github.com/chokkyvista/iPM3F

Table 1. Test error of different models on the MovieLens and EachMovie data sets.

	MovieLens		EachMovie	
Algorithm	**weak**	**strong**	**weak**	**strong**
M^{3}F (Rennie & Srebro, 2005)	$.4156 \pm .0037$	$.4203 \pm .0138$	$.4397 \pm .0006$	$.4341 \pm .0025$
bcd M^{3}F (Xu et al., 2012)	$.4176 \pm .0016$	$.4227 \pm .0072$	$.4348 \pm .0023$	$.4301 \pm .0034$
Gibbs M^{3}F	$.4037 \pm .0005$	$.4040 \pm .0055$	$.4134 \pm .0017$	$.4142 \pm .0059$
iPM^{3}F (Xu et al., 2012)	$.4031 \pm .0030$	$.4135 \pm .0109$	$.4211 \pm .0019$	$.4224 \pm .0051$
Gibbs iPM^{3}F	$.4080 \pm .0013$	$.4201 \pm .0053$	$.4220 \pm .0003$	$.4331 \pm .0057$



(a) (b) (c)

Figure 1. (a) Regularized risk, NMAE of (b) Gibbs M^{3}F and (c) Gibbs iPM^{3}F on the EachMovie data set.

Test error: We report NMAE error of the averaged samples for Gibbs M^{3}F and Gibbs iPM^{3}F. As shown in Table 1, Gibbs M^{3}F significantly outperforms previous parametric M^{3}F models, for both weak and strong generalization tasks. We believe this largely attributes to our additionally introduced regularizer for $\boldsymbol{\theta}$ (15); For the nonparametric models, although Gibbs iPM^{3}F only obtains comparable, or even marginally worse test performance compared with iPM^{3}F, we consider it to be the cost of exchanging accuracy for efficiency since our alternative relaxed loss term (19) favors the development of much more efficient learning algorithms.

Training time: In Table 2, the training time of M^{3}F is directly cited from (Rennie & Srebro, 2005) and it was measured on a "single 3.06GHz Pentium 4 CPU" while all other 4 methods were measured by MATLAB with single computational thread on a 4-core 3.00GHz Intel i5 CPU. In both cases, our proposed methods achieved drastic efficiency gain. Note that M^{3}F works with a derivable *Smooth Hinge* while our methods directly work with the hinge loss without solving time-consuming SVMs. Also note that for the nonparametric Gibbs iPM^{3}F, its number of active factors K is constantly changing during the sampling process, and so is the running time for each iteration, as shown in Fig. 2(b). We discuss the asymptotic computational complexity of our methods in the appendix.

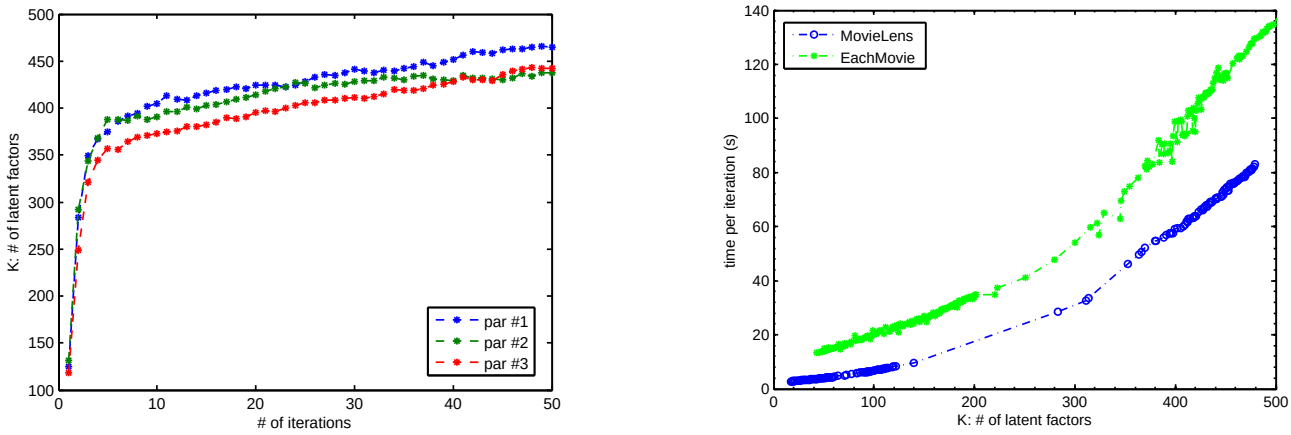
Table 2. Training time of different models.

Algorithm	**MovieLens**	**EachMovie**	**Iters**
M^{3}F	5h	15h	100
bcd M^{3}F	4h	10h	50
Gibbs M^{3}F	0.11h	0.35h	50
iPM^{3}F	4.6h	5.5h	50
Gibbs iPM^{3}F	0.68h	0.70h	50

RRM-MAP duality: Although Gibbs iPM^{3}F is directly defined from problem (12) without explicit reference to any underlying RRM as Gibbs M^{3}F, we may still find the corresponding RRM by choosing the regularizer as $\Omega'(\mathcal{M}) = -\log \pi(\mathcal{M})$ and thus obtain

$$\Omega'(\mathcal{M}) = \frac{1}{2\sigma^2}\|V\|_F - \log \mathrm{IBP}(Z|\alpha) + \frac{1}{2\varsigma^2}\sum_{i=1}^{N}\|\boldsymbol{\theta}_i - \boldsymbol{\rho}\|_2^2.$$

We can see more clearly in this form that the variance parameters σ and ς each has their own right in weighing the regularizer and cannot be offset by the regularization constant C. The benefit of acquiring this induced RRM is that it allows us to calculate the objective value and use it as a criterion of convergence. We found that for Gibbs iPM^{3}F, the induced regularized risk (calculated from single samples) does not necessarily decrease during the sampling process yet we still observed a stable trend of NMAE going down.



(a) (b)

Figure 2. (a) Number of latent factors of Gibbs iPM^{3}F on MovieLens. (b) Time per iteration as a function of K.

Convergence: It might be true that the Markov chain itself has not properly converged when we stop the iteration based on objective values and validation error, especially for the IBP-involved nonparametric model. Yet from our experience and experiments, the Gibbs sampler works quite well and the curves do display a clear trend of convergence in Fig. 1. We hope to come up with further detailed analysis in future work.

Latent dimension: Fig. 2(a) shows the running number of latent factors inferred from data by Gibbs iPM^{3}F with the three randomly constructed training sets as indicated by different colors and this clearly illustrates the flexibility of nonparametric models. The "optimal" latent dimension appears to be around 450 for MovieLens and 200 for EachMovie. We emphasize however, given any specific hyper-parameters, the corresponding Gibbs M^{3}F model always has its own optimal solution that comes with a latent dimension and a test error; And by "optimal" latent dimension we actually mean the one with the best test error.

Poisson truncation level: When sampling new latent factors k_i in Gibbs iPM^{3}F, we specify a Poisson truncation level κ as did (Doshi-Velez et al., 2009) so that k_i greater than κ get directly rejected. This would not be a problem since we find that the cost of sampling new latent factors $Z_{i\nu}$ can be reduced to linear to κ and thus we may set κ to be sufficiently large without worrying about its impact on efficiency. We defer details into the appendix. In our current implementation, we choose $\kappa = 10$.

Factor alignment: Averaging samples from Gibbs iPM^{3}F is a little bit tricker than from Gibbs M^{3}F, since we are constantly facing newly generated factors as well as nullified factors that get crossed out and missing from subsequent samples. We compared two different methods for this. The first one ignores such factor alignment and sum two samples directly as is, padding zero wherever necessary. While the second one respects such correspondence between factors and makes sure they are always properly aligned before averaging. Our experiments indicate no telling difference between these two methods.

5. Conclusions

We have presented a novel probabilistic interpretation of max-margin matrix factorization, which naturally leads to a simple and fast algorithm by exploring the ideas of data augmentation. Moreover, we generalized the ideas to present a new nonparametric Bayesian max-margin matrix factorization model, which again has a simple and efficient sampling algorithm without making any restricting assumptions on the posterior distributions or setting a truncation level to the number of latent factors as in existing variational methods.

Acknowledgments

This work is supported by the National Basic Research Program (973 Program) of China (Nos. 2013CB329403, 2012CB316301), National Natural Science Foundation of China (Nos. 91120011, 61273023), and Tsinghua University Initiative Scientific Research Program (No. 20121088071).

References

Doshi-Velez, F., Miller, K. T., van Gael, J., and Teh, Y. W. Variational inference for the Indian buffet process. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 12, 2009.

Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. PAC-Bayesian learning of linear classifiers. In *International Conference on Machine Learning (ICML)*, pp. 353–360, 2009.

Griffiths, T. and Ghahramani, Z. Infinite latent feature models and the Indian buffet process. Technical report, Gatsby Computational Neuroscience Unit, 2005.

Jaakkola, T., Meila, M., and Jebara, T. Maximum entropy discrimination. In *Advances in Neural Information Processing Systems (NIPS)*, 1999.

Marlin, B. and Zemel, R. S. The multiple multiplicative factor model for collaborative filtering. In *International Conference on Machine Learning (ICML)*, 2004.

McAllester, D. PAC-Bayesian stochastic model selection. *Machine Learning*, 51:5–21, 2003.

Polson, N. G. and Scott, S. L. Data augmentation for support vector machines. *Bayesian Analysis*, 6(1): 1–24, 2011.

Rennie, J. D. M. and Srebro, N. Fast maximum margin matrix factorization for collaborative prediction. In *International Conference on Machine Learning (ICML)*, 2005.

Salakhutdinov, R. and Mnih, A. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *International Conference on Machine Learning (ICML)*, 2008.

Srebro, N., Rennie, J. D. M., and Jaakkola, T. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*, 2005.

Tanner, M. A. and Wong, W. H. The calculation of posterior distributions by data augmentation. *Journal of the Americal Statistical Association (JASA)*, 82(398):528–540, 1987.

van Dyk, D. and Meng, X. The art of data augmentation. *Journal of Computational and Graphical Statistics (JCGS)*, 10(1):1–50, 2001.

Xu, M., Zhu, J., and Zhang, B. Nonparametric maximum margin matrix factorization for collaborative prediction. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

Zhou, M., Wang, C., Chen, M., Paisley, J., Dunson, D., and Carin, L. Nonparametric Bayesian matrix completion. In *Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pp. 213–216, 2010.

Zhu, J., Ahmed, A., and Xing, E. P. MedLDA: Maximum margin supervised topic models for regression and classification. In *International Conference on Machine Learning (ICML)*, 2009.

Zhu, J., Chen, N., Perkins, H., and Zhang, B. Gibbs max-margin topic models with fast sampling algorithms. In *International Conference on Machine Learning (ICML)*, 2013a.

Zhu, J., Chen, N., and Xing, E. P. Bayesian inference with posterior regularization and applications to infinite latent svms. *arXiv Report*, `arXiv:1210.1766v2`, 2013b.