Gated Probabilistic Matrix Factorization: Learning Users' Attention from Missing Values

Shohei Ohsawa, Yachiko Obara, Takayuki Osogami
IBM Research – Tokyo
19–21 Nihonbashi, Hakozaki-cho, Chuo-ku, Tokyo, Japan
{ohsawrks, obara, osogami}@jp.ibm.com

Abstract

Recommender systems rely on techniques of predicting the ratings that users would give to yet unconsumed items. Probabilistic matrix factorization (PMF) is a standard technique for such prediction and makes a prediction on the basis of an underlying probabilistic generative model of the behavior of users. We investigate a new model of users' consumption and rating, where a user tends to consume an item that emphasizes those features that the user seeks to enjoy, and the ratings of the users are more strongly affected by those features than others. We incorporate this new user model into PMF and show that the resulting method, *Gated PMF* (GPMF), improves the predictive accuracy by several percent on standard datasets. GPMF is widely applicable, as it is trained only with the ratings given by users and does not rely on any auxiliary data.

1 Introduction

Providing effective recommendations is critical in online stores, social media, and other large systems where users cannot easily find what they want (products, friends, or other items). Many recommender systems ask their users to rate the items to improve the quality of future recommendation. Here, the ratings provided by users are used to predict the ratings that (those or other) users would give to yet unrated items. Improving the quality of such prediction (and then recommendation) has an immediate impact on user satisfaction and the revenue of service providers.

Probabilistic matrix factorization (PMF) [Mnih and Salakhutdinov, 2007] is a widely used technique of predicting such ratings and has been particularly influential in the literature. PMF is based on a generative model where the rating given by a user on an item follows a distribution that depends on the latent *preferences* of the user and the latent *features* of the item. Here, a user tends to give high ratings on those items whose features match the preferences of the user.

The key information that has been used to improve the quality of prediction made by PMF and other approaches is the dependency between which items a user rates and what ratings the user gives. For example, users who have rated common items (e.g., watched common movies) tend to provide ratings that are similar to each other [Mnih and Salakhutdinov, 2007].

What has been ignored in the prior work is the particular dependency between why a user consumes an item and how that affects the user's rating. This paper investigates our own hypothesis, which is motivated by observations in the literature of economics [Davenport and Beck, 2001; Schoormans and Robben, 1997], about how a user consumes and rates items. A user looks for an item that has the particular features that the user wants to enjoy and tends to consume those items that emphasize such features (e.g., by printing those features on their packages). When the user rates the consumed item, the rating is more strongly affected by the quality of those features that are emphasized by the item and that have caught the attention of the user. For example, a user might choose to watch a movie simply because it has been advertised to be funny. We expect that this user's rating on this movie will be strongly affected, either negatively or positively, by whether the movie is funny in the way that the user has expected.

Here, we propose *Gated PMF* (GPMF), which extends PMF to take into account the dependency between why a user consumes an item and how that affects the rating. Similar to PMF, GPMF is based on a probabilistic generative model, which we refer to as the *consumption-rating model* (CRM). In the CRM, a user consumes an item with some probability that depends on what features the user seeks (i.e., *attention* of the user) and what features the item emphasizes (i.e., *attraction* of the item). The attention and attraction then modify the distribution of the user's rating, which would otherwise depend solely on the user's *preferences* and the item's *features*. More specifically, the attention and attraction select some of the dimensions of the preferences and features so that the selected dimensions have more impact on the rating than others. GPMF and the CRM, motivated by our own hypothesis, constitute the first contribution of this paper.

We then show the effectiveness of GPMF through numerical experiments with datasets from the real world. Specifically, we found that GPMF improves the predictive accuracy upon PMF by several percent depending on the settings. These experimental results supporting our hypothesis constitute our second contribution.

After we discuss related work, we introduce GPMF.

Specifically, we show how to learn the parameters of the CRM and discuss how GPMF implements the ideas from our hypothesis. We then show the results of numerical experiments.

2 Gated Probabilistic Matrix Factorization

We present GPMF in this section. After we give a formal definition of the problem we deal with, we propose CRM, upon which GPMF is based. We then construct the GPMF with the CRM. After that, we discuss how the GPMF implements the ideas in our hypothesis.

2.1 Problem Definition

We study the problem of predicting the values of users' ratings on items. Users have consumed and rated some of the items, and we assume that we can observe and know those ratings. We then seek to predict the users' ratings on items that have not yet been rated. Let N be the number of users and M be the number of items. A rating is a real value in $[-1, 1]$.

Let $\mathbf{R} \in [-1, 1]^{N\times M}$ be a matrix, which we refer to as the *observed rating matrix*. The (i, j)-th element of $\mathbf{R}$ denotes the observed rating that the i-th user has given to the j-th item. If the item has not been rated by the user, the observed rating is undefined. Let $\mathbf{C} \in \{0, 1\}^{N\times M}$ be an indicator matrix whose (i, j)-element, c_{ij}, is 1 if the i-th user has consumed (and rated) the j-th item (i.e., r_{ij} is defined) and 0 otherwise. We refer to $\mathbf{C}$ as the *observed consumption matrix*.

We consider a stochastic generative model for users to consume and rate items. Let $\mathbf{R}^*$ be a random matrix of size $N \times M$, which we refer to as the *real rating matrix*. Its (i, j)-element, r^*_{ij}, is a random variable, having the support on $[-1, 1]$, that denotes the rating that the i-th user gives to the j-th item given that the user consumes the item. Let $\mathbf{C}^*$ be a random matrix of size $N \times M$, which we refer to as the *real consumption matrix*. Its (i, j)-element, c^*_{ij}, is a random variable, having the support on $\{0, 1\}$, that denotes whether the j-th item is consumed by the i-th user ($c_{ij} = 1$) or not ($c_{ij} = 0$). Here, $\mathbf{R}$ is a sample from $\mathbf{R}^*$, and $\mathbf{C}$ is from $\mathbf{C}^*$.

Our goal is to give an *estimated rating matrix*, $\tilde{\mathbf{R}}$, in such a way that its (i, j)-element comes close to the rating that the i-th user gives to the j-th item. The quality of $\tilde{\mathbf{R}}$ is evaluated on the basis of $\mathbf{R}_{\text{test}}$ and $\mathbf{C}_{\text{test}}$, a *realized rating matrix* and a *realized consumption matrix*, that will be sampled from $\mathbf{R}^*$ and $\mathbf{C}^*$, respectively, after estimating $\tilde{\mathbf{R}}$. Specifically, we seek to minimize the following:

$$\mathbb{E}_{\mathbf{R}^*,\mathbf{C}^*}\left[\left\|\mathbf{C}_{\text{test}} \circ \left(\tilde{\mathbf{R}} - \mathbf{R}_{\text{test}}\right)\right\|^2_{\text{F}}\right], \tag{1}$$

where $\|\cdot\|_{\text{F}}$ denotes the Frobenius norm of a matrix, and the expectation is with respect to the distributions of $\mathbf{R}^*$ and $\mathbf{C}^*$, which are unknown to us.

2.2 The Consumption-Rating Model

Here we describe the generative model behind GPMF: the *consumption-rating model* (CRM). CRM naturally extends the generative model behind PMF by explicitly modeling the phenomenon that a user consumes an item before he rates it.

The rating model of PMF [Mnih and Salakhutdinov, 2007] is a probabilistic model that extends the singular value decomposition (SVD) to take into account missing values. Let L be the dimension of the feature space. The rating model of PMF, shown in Figure 1(a) as a graphical model, models the observed rating r^* with an inner product of two vectors, a user preference $\mathbf{u}_i \in \mathbb{R}^L$ and an item feature $\mathbf{v}_j \in \mathbb{R}^L$:

$$p(r^*_{ij} = r_{ij}|\mathbf{u}_i, \mathbf{v}_j, \sigma^2_r) = \mathcal{N}(r_{ij}|\mathbf{u}_i^{\mathrm{T}}\mathbf{v}_j, \sigma^2_r), \tag{2}$$

where $\sigma^2_r \in \mathbb{R}^+$ is a variance of r^* ($\sigma^2_\cdot \in \mathbb{R}^+$ denotes a variance in this paper) and $\mathcal{N}(\cdot|\cdot)$ is a probability density function (PDF) of a normal distribution. We denote two matrices containing all the user preferences and the item features, respectively, with $\mathbf{U} \equiv (\mathbf{u}_1, \dots, \mathbf{u}_N)$ and $\mathbf{V} \equiv (\mathbf{v}_1, \dots, \mathbf{v}_M)$. For brevity, we omit random variables in PDFs when those random variables are clear from the context. For example, we will write $p(r_{ij})$ to mean $p(r^*_{ij} = r_{ij})$. The rating model of PMF models the likelihood of $\mathbf{U}$ and $\mathbf{V}$ by ignoring missing values and assumes that each element r^*_{ij} of $\mathbf{R}^*$ is conditionally independent of each other given $\mathbf{U}$, $\mathbf{V}$, and σ^2_r. That is,

$$p(\mathbf{R}|\mathbf{C}, \mathbf{U}, \mathbf{V}, \sigma^2_r) = \prod_{i=1}^{N}\prod_{j=1}^{M}\left[p(r_{ij}|\mathbf{u}_i, \mathbf{v}_j, \sigma^2_r)\right]^{c_{ij}}. \tag{3}$$

We now describe our CRM, whose graphical model is shown in Figure 1(b). The key elements of the CRM are the attention of a user and the attraction of an item, which are respectively denoted by L-dimensional vectors $\mathbf{g}_i$ and $\mathbf{h}_j$. Namely, $\mathbf{g}_i$ denotes what features of items a user seeks to enjoy and $\mathbf{h}_j$ denotes what features an item emphasizes.

CRM models c^*_{ij} with a Bernoulli distribution whose mean is an inner product of the two vectors:

$$p(c_{ij}|\mathbf{g}_i, \mathbf{h}_j) = \mathcal{B}(c_{ij}|\mathbf{g}_i^{\mathrm{T}}\mathbf{h}_j), \tag{4}$$

where $\mathcal{B}(\cdot|q) \equiv q^x(1 - q)^{1-x}$ for $x \in \{0, 1\}$ and $0 < q < 1$. The elements of $\mathbf{g}_i$ and $\mathbf{h}_j$ independently follow Beta distributions:

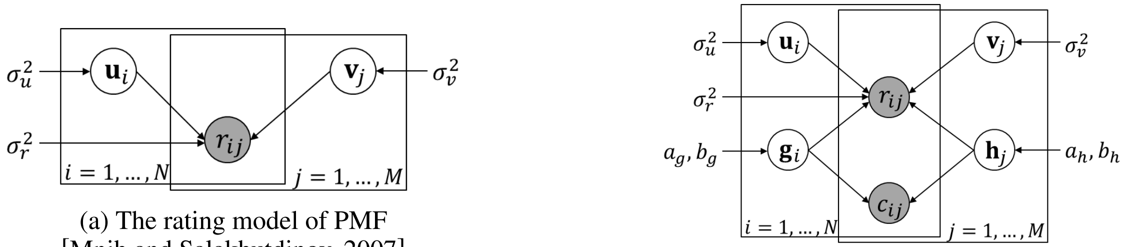
$$p(\mathbf{g}_i|a_g, b_g) = \prod_{k=1}^{L}\mathrm{Beta}(g_{ik}|a_g, b_g), \tag{5}$$

$$p(\mathbf{h}_j|a_h, b_h) = \prod_{k=1}^{L}\mathrm{Beta}(h_{jk}|a_h, b_h), \tag{6}$$

where $\mathrm{Beta}(q|a, b) \equiv q^a(1 - q)^b/Z(a + 1, b + 1)$ for $q \in (0, 1)$, $a > 0$, and $b > 0$, and Z denotes the Beta function. We denote two matrices containing all the attention and the attraction, respectively with $\mathbf{G} = (\mathbf{g}_1, \dots, \mathbf{g}_N)$ and $\mathbf{H} = (\mathbf{h}_1, \dots, \mathbf{h}_M)$.

In our rating model, the rating value follows normal distribution whose mean is a "gated" inner product of the user preference $\mathbf{u}_i$ and the item feature $\mathbf{v}_j$. By "gated" we mean that the vectors are weighted by a diagonal matrix that depends on the consumption model. That is, the PDF of the rating value is given by

$$p(r_{ij}|\mathbf{u}_i, \mathbf{v}_j, \mathbf{g}_i, \mathbf{h}_j, \sigma^2_r) = \mathcal{N}\left(r_{ij}|\mathbf{u}_i^{\mathrm{T}}\mathbf{\Gamma}_{ij}\mathbf{v}_j, \sigma^2_r\right), \tag{7}$$



(a) The rating model of PMF [Mnih and Salakhutdinov, 2007]

(b) The consumption–rating model of GPMF

Figure 1: The graphical models of each probabilistic model of PMF and GPMF. Gray circles and white circles indicate observed variables and latent variables, respectively. Symbols without circles indicate constants. Arrows between symbols indicate probabilistic dependency (e.g., X generates Y). Rectangle with suffixes indicates a block, which consists of multiple elements.

where

$$\boldsymbol{\Gamma}_{ij} = \frac{1}{\mathbf{g}_i^{\mathrm{T}}\mathbf{h}_j}\begin{pmatrix} g_{1i}h_{1j} & & 0 \\ & \ddots & \\ 0 & & g_{Li}h_{Lj} \end{pmatrix}. \tag{8}$$

For $k \in \{1, \ldots, L\}$, the element $g_{ki}h_{kj}$ takes a large value when both attention and attraction have large values. The coefficient $1/\mathbf{g}_i^{\mathrm{T}}\mathbf{h}_j$ guarantees that the trace of $\boldsymbol{\Gamma}_{ij}$ is 1. Hence, $\boldsymbol{\Gamma}_{ij}$ determines the relative importance of each axis in the vector space for the i-th user and the j-th item. The distribution of r_{ij} is equivalent to the rating model of PMF if $\boldsymbol{\Gamma}_{ij} = L^{-1}\mathbf{I}$ because the mean is proportional to $\mathbf{u}_i^{\mathrm{T}}\mathbf{v}_j$. The priors of the parameters $\mathbf{u}_i$ and $\mathbf{v}_j$ follow zero-mean normal distribution and their PDFs are denoted by

$$p(\mathbf{u}_i|\sigma_u^2) = \mathcal{N}(\mathbf{u}_i|0, \sigma_u^2\mathbf{I}),\; p(\mathbf{v}_j|\sigma_v^2) = \mathcal{N}(\mathbf{v}_j|0, \sigma_v^2\mathbf{I}). \tag{9}$$

Similar to the rating model of PMF, we assume that elements of $\mathbf{C}$ are conditionally independent of each other given $\mathbf{G}$ and $\mathbf{H}$ and that elements of $\mathbf{R}$ are conditionally independent of each other given $\mathbf{U}$ and $\mathbf{V}$. The PDFs of $\mathbf{C}$ and $\mathbf{R}$ are then denoted as follows:

$$p(\mathbf{C}|\mathbf{G}, \mathbf{H}) = \prod_{i=1}^{N}\prod_{j=1}^{M} p(c_{ij}|\mathbf{g}_i, \mathbf{h}_j), \tag{10}$$

$$\begin{aligned} p(\mathbf{R}|\mathbf{C}, \mathbf{U}, \mathbf{V}, &\mathbf{G}, \mathbf{H}, \sigma_r^2) \\ &= \prod_{i=1}^{N}\prod_{j=1}^{M} p(r_{ij}|\mathbf{u}_i, \mathbf{v}_j, \mathbf{g}_i, \mathbf{h}_j, \sigma_r^2)^{c_{ij}}, \end{aligned} \tag{11}$$

where the right-hand sides are given by Eqs. (4) and (7).

In summary, the proposed CRM is characterized by the following PDF:

$$\begin{aligned} p(&\mathbf{R}, \mathbf{U}, \mathbf{V}, \mathbf{C}, \mathbf{G}, \mathbf{H}|\boldsymbol{\Omega}) \\ &= \; p(\mathbf{R}|\mathbf{C}, \mathbf{U}, \mathbf{V}, \mathbf{G}, \mathbf{H}, \sigma_r^2)\, p(\mathbf{U}|\sigma_u^2)\, p(\mathbf{V}|\sigma_v^2) \\ &\quad\; p(\mathbf{C}|\mathbf{G}, \mathbf{H})\, p(\mathbf{G}|a_g, b_g)\, p(\mathbf{H}|a_h, b_h), \end{aligned} \tag{12}$$

where $\boldsymbol{\Omega} \equiv (\sigma_r^2, \sigma_u^2, \sigma_v^2, a_g, b_g, a_h, b_h)$ denotes the hyperparameters and the factors in the right-hand side are given by Eqs. (5), (6), (9), (10), (11).

The main feature differentiating CRM from the rating model of PMF is the dependency between rating and consumption that are modeled with two latent factors, attention and attraction. In our consumption model, $\mathbf{g}_i$ and $\mathbf{h}_j$ represent the mean value of c with their inner product, $\mathbf{g}_i^{\mathrm{T}}\mathbf{h}_j$. When $\mathbf{g}_i$ and $\mathbf{h}_j$ have elements that take large values in common dimensions, the i-th user tends to consume the j-th item. In our rating model, $\mathbf{g}_i$ and $\mathbf{h}_j$ define the gate matrix $\boldsymbol{\Gamma}_{ij}$, which is diagonal. Intuitively, each element in the gate matrix scales the vector space where the inner product of $\mathbf{u}_i$ and $\mathbf{v}_j$ is evaluated. In other words, the value of $\mathbf{u}_i^T\boldsymbol{\Gamma}_{ij}\mathbf{v}_j$ is strongly affected by the values of $\mathbf{u}_i$ and $\mathbf{v}_j$ in the dimensions where $\boldsymbol{\Gamma}_{ij}$ has large values. Hence, CRM improves the predictive accuracy by using the consumption behavior to estimate the importance of the features in rating.

While we could consider a simpler method that applies PMF twice, such a method (*two-phase PMF*) is inferior to GPMF, for the following reasons. In two-phase PMF, the first PMF predicts consumption values and the second PMF predicts rating values. The two drawbacks here are that first, two-phase PMF does not consider how to connect consumption values and rating values, and second, two-phase PMF treats the distribution of consumption value as normal distribution even though the individual consumption value only takes the binary value. In the experiment, we show that GPMF outperforms two-phase PMF.

2.3 Learning and Predicting with the GPMF

We now derive the loss function on the basis of maximum a posteriori (MAP) estimation. We define $\boldsymbol{\Theta} \equiv \{\mathbf{U}, \mathbf{V}, \mathbf{G}, \mathbf{H}\}$ as a set of parameters to optimize. We maximize the posterior distribution of $\boldsymbol{\Theta}$ with respect to the observed ratings $\mathbf{R}$ and missing values $\mathbf{C}$. We seek to obtain the optimal parameter $\hat{\boldsymbol{\Theta}} \equiv \mathrm{argmax}_{\boldsymbol{\Theta}} p(\boldsymbol{\Theta}|\mathbf{R}, \mathbf{C}, \boldsymbol{\Omega})$ by minimizing the following loss function, E, with respect to $\boldsymbol{\Theta}$:

$$E \equiv -\log p(\boldsymbol{\Theta}|\mathbf{R}, \mathbf{C}, \boldsymbol{\Omega}). \tag{13}$$

The loss function E can be written as $E = E_C + E_R$, where E_C and E_R are the loss function of the consumption model and the rating model, respectively. Specifically, each loss function can be derived as follows.

$$E_C \equiv \left\| \mathbf{C} \circ \mathrm{Log}\tilde{\mathbf{C}} + (\mathbf{1} - \mathbf{C}) \circ \mathrm{Log}(\mathbf{1} - \tilde{\mathbf{C}}) \right\|_1 + a_g \left\| \mathrm{Log}\mathbf{G} \right\|_1 + b_g \left\| \mathrm{Log}(\mathbf{1} - \mathbf{G}) \right\|_1 + a_h \left\| \mathrm{Log}\mathbf{H} \right\|_1 + b_h \left\| \mathrm{Log}(\mathbf{1} - \mathbf{H}) \right\|_1 + \mathrm{const.}, \tag{14}$$

$$E_R \equiv \frac{1}{2\sigma_r^2} \left\| \mathbf{C} \circ (\mathbf{R} - \tilde{\mathbf{R}}) \right\|_{\mathrm{F}}^2 + \frac{1}{2\sigma_u^2} \left\| \mathbf{U} \right\|_{\mathrm{F}}^2 + \frac{1}{2\sigma_v^2} \left\| \mathbf{V} \right\|_{\mathrm{F}}^2, \tag{15}$$

where

$$\tilde{\mathbf{C}} = \mathbf{G}^{\mathrm{T}} \mathbf{H}, \tag{16}$$

$$\tilde{\mathbf{R}} = \tilde{\mathbf{C}}^{\sharp} \circ (\mathbf{G} \circ \mathbf{U})^{\mathrm{T}} (\mathbf{H} \circ \mathbf{V}), \tag{17}$$

$\mathrm{Log}(\cdot)$ is an element-wise natural logarithm function for a matrix, $\mathbf{1}$ is a matrix with every element being 1, $\|\cdot\|_1$ is a l_1 norm, $\cdot^{\sharp}$ is an inverse matrix for Hadamard product (i.e., $\mathbf{X} \circ \mathbf{X}^{\sharp} = \mathbf{1}$ for a matrix $\mathbf{X}$), and $\mathrm{const.}$ includes the terms that depends only on the hyperparameters, $\mathbf{\Omega}$. The learning process of the GPMF proceeds by minimizing E with respect to $\mathbf{\Theta}$ by using stochastic gradient descent (SGD). In each step of SGD, we squash the value of each element of $\tilde{\mathbf{R}}$ into [-1,1] using a hyperbolic tangent function, $\tanh(\tilde{r}_{ij})$, to be consistent with the range of ratings. We perform a grid search to optimize $\mathbf{\Omega}$.

2.4 Automatic Feature Selection and Resampling

GPMF performs *feature selection* and *resampling* automatically by optimizing the loss functions. In the feature selection, from the L-dimensional features, GPMF places higher weights on the element that the user pays attention to. In the resampling, GPMF selects the items that users tend to consume and increases the weight on the rating value of those selected items. In the following, we elaborate on these roles that GPMF plays.

Feature Selection To describe why GPMF acts as a feature selector, we change variables: $\mathbf{U}_g = \mathbf{G} \circ \mathbf{U}$ and $\mathbf{V}_h = \mathbf{H} \circ \mathbf{V}$. We can interpret $\mathbf{U}_g$ and $\mathbf{V}_h$ as feature matrices filtered by attraction and attention, respectively. Substituting $\mathbf{U}_g$ and $\mathbf{V}_h$ into the CRM instead of $\mathbf{U}$ and $\mathbf{V}$, we can rewrite Eqs. (15) and (17) as

$$E_R = \frac{1}{2\sigma_r^2} \left\| \mathbf{C} \circ (\mathbf{R} - \tilde{\mathbf{R}}) \right\|_{\mathrm{F}}^2 + \frac{1}{2\sigma_u^2} \left\| \mathbf{G}^{\sharp} \circ \mathbf{U}_g \right\|_{\mathrm{F}}^2 + \frac{1}{2\sigma_v^2} \left\| \mathbf{H}^{\sharp} \circ \mathbf{V}_h \right\|_{\mathrm{F}}^2, \tag{18}$$

where

$$\tilde{\mathbf{R}} = \tilde{\mathbf{C}}^{\sharp} \circ \left(\mathbf{U}_g{}^{\mathrm{T}} \mathbf{V}_h \right). \tag{19}$$

The last two terms of Eq. (18) indicate that $\mathbf{G}$ and $\mathbf{H}$ are regularization coefficients that control the variance of the elements of $\mathbf{U}_g$ and $\mathbf{V}_h$. By optimizing the loss function, the relevant features in $\mathbf{U}_g$ and $\mathbf{V}_h$ take large variance and irrelevant features take small variance. Hence, the irrelevant features should vanish.

Resampling Next, we discuss why GPMF performs resampling in the context of minimizing mean squared error (MSE). MSE is a popular objective function that is often used to evaluate methods such as PMF for predicting ratings. The expected value of MSE with respect to $\mathbf{C}^*$ can be written as

$$\mathbb{E}_{\mathbf{C}^*}\left[\mathrm{MSE} | \mathbf{R}^* \right] = \left\| \mathbf{\Psi} \circ (\mathbf{R}^* - \tilde{\mathbf{R}}) \right\|_{\mathrm{F}}^2, \tag{20}$$

where $\mathbf{\Psi} \in \mathbb{R}^{N \times M}$, in an element-wise manner, denotes the expected value of $\mathbf{C}^*$. The expected MSE is proportional to the squared Euclidean distance between the real matrix $\mathbf{R}^*$ and the predicted matrix $\tilde{\mathbf{R}}$, as long as $\mathbf{R}$ is randomly observed from $\mathbf{R}^*$ or the entries of $\mathbf{\Psi}$ are identical. The observed ratings, however, do not follow the distribution of $\mathbf{R}^*$ because of the dependency on $\mathbf{C}^*$, and the MSE is biased by $\mathbf{\Psi}$.

Therefore, predicting consumption (i.e., which elements of $\mathbf{\Psi}$ have large values) can contribute to reducing the MSE. This is the trick GPMF implicitly utilizes. In our CRM, $\mathbf{\Psi} = \tilde{\mathbf{C}}$ holds because of Eq. (4). Therefore, Eq. (20) can be written as follows.

$$\mathbb{E}_{\mathbf{C}^*}\left[\mathrm{MSE} | \mathbf{R}^* \right] = \left\| \tilde{\mathbf{C}} \circ (\mathbf{R}^* - \tilde{\mathbf{R}}) \right\|_{\mathrm{F}}^2. \tag{21}$$

As we cannot observe $\mathbf{R}^*$ in the training phase, we use $\mathbf{C} \circ \mathbf{R}$ instead of $\mathbf{R}^*$. By Eq. (19), the objective function is then approximated as follows.

$$\mathbb{E}_{\mathbf{C}^*}\left[\mathrm{MSE} | \mathbf{R}^* \right] \approx \left\| \mathbf{C} \circ (\tilde{\mathbf{C}} \circ \mathbf{R} - \mathbf{U}_g^{\mathrm{T}} \mathbf{V}_h) \right\|_{\mathrm{F}}^2. \tag{22}$$

GPMF minimizes (22) with regularization. We can interpret that $\tilde{\mathbf{C}} \circ \mathbf{R}$ is *resampled* from $\mathbf{R}$ by taking into account whether a user is likely to consume an item.

3 Experimental Results

3.1 Data Sets and Evaluation Metrics

We evaluate GPMF using real-world datasets. Although the most popular dataset is that of Netflix, used in the paper of PMF [Mnih and Salakhutdinov, 2007], this dataset is currently unavailable due to privacy issues. Hence, we use MovieLens-100k (http://grouplens.org/datasets/movielens/) and MovieTweetings (http://github.com/sidooms/MovieTweetings); see Table 1. MovieLens-100k (ML100K) is a real-world dataset from MovieLens, which provides services of rating movies. In ML100K, users give ratings (1, 2, 3, 4, or 5) to items (i.e., movies). MovieTweetings (MTweet) is a dataset for benchmarking, crawled from social media, for several recommendation tasks in RecSys 2013 [Dooms *et al.*, 2013] and has been used in several studies [Hernandez-lobato *et al.*, 2014]. MTweet consists of rating values of 1-10 given to movies by users of Twitter. We squash the ratings into the range $[0, 1]$ such that 0 is the lowest.

3.2 Baselines and Parameter Settings

Our experiments compare GPMF against several baselines:

Table 1: Summary of the datasets. N and M indicate the number of users and items, respectively. #Ratings is the number of observed ratings and Sparsity is #Ratings per NM.

Datasets	N	M	#Ratings	Sparsity
ML100K	943	1,682	100,000	0.063
MTweet	3,871	2,217	111,566	0.013

- PMF and its extension (CPMF) [Mnih and Salakhutdinov, 2007]. CPMF exploits the consumption effect by smoothing. The difference between our method and CPMF is that we consider the effect of attention and attraction.
- MNAR-PMF [Hernandez-lobato *et al.*, 2014], which models value-based missing value mechanisms.
- Low rank completion (LRC) [Jain *et al.*, 2013], which predicts missing values of the rating matrix by assuming that the matrix has a low rank.
- Two-phase PMF, which simply applies PMF twice.

As the CRM tightly integrates a rating model and a consumption model, we evaluate only the end-to-end accuracy of GPMF on predicting ratings. Note that GPMF makes predictions about consumption only to improve the predictive accuracy of ratings.

In the training, we adjust hyperparameters by grid search. Specifically, we choose the dimension of feature space L, the learning rate of the SGD, and the regularization parameter λ from the following candidates: $10^{-3}, 10^{-2}, 10^{-1}$, and 1.

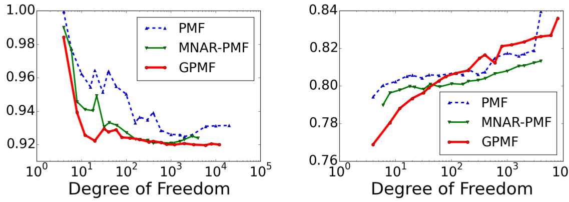
We utilize mean absolute error (MAE) and root mean squared error (RMSE) for evaluating the experimental results. Note that, in our settings, RMSE is proportional to log-likelihood, which was used in the related work [Lakshminarayanan *et al.*, 2011; Hernandez-lobato *et al.*, 2014]. We evaluate both the MAE and the RMSE by five-fold cross validation.

3.3 Experimental Results

Table 2 shows that the GPMF outperforms the baselines. Figure 2 shows more details of the comparisons, where we now vary the degree of freedom (i.e., the number of parameters to optimize, which is varied with L) for each method. Both panels in Figure 2 suggest that GPMF enables us to obtain higher predictive accuracy than PMF and MNAR-PMF when the degree of freedom is at most 30. GPMF is inferior or comparable to the baselines for high degrees of freedom. In Figure 2a, although GPMF has comparable accuracy to MNAR-PMF for high degrees of freedom ($10^2 - 4 \times 10^3$), GPMF outperforms MNAR-PMF for lower degrees of freedom ($10 - 30$). This means that GPMF is able to obtain simpler models than MNAR-PMF. GPMF consistently outperforms PMF for all degrees of freedom under consideration. In Figure 2b, GPMF outperforms PMF and MNAR-PMF for low degrees of freedom. The reason for the differing trends between the two datasets is that the MovieTweetings data can be represented by the product of matrices, each having a low rank. When the dimension of the model is larger than the rank of the data, the model tends to overfit the data and the error increases. Overall, GPMF tends to outperform PMF, primarily because GPMF selects the relevant features from noisy rating values. Note that, similar to PMF, the computational cost of GPMF scales linear with the sample size, since GPMF requires only an additional computation of consumption matrix.

Table 2: Performance of GPMF compared against several existing methods. We performed a grid search for each method, and the best performance is shown.

	ML100K		MTweet	
Method	**MAE**	**RMSE**	**MAE**	**RMSE**
LRC	0.7313	0.9384	0.6037	0.7961
PMF	0.7330	0.9248	0.6000	0.7942
CPMF	0.7305	0.9233	0.5876	0.7823
Two-phase PMF	0.7314	0.9211	0.6021	0.7935
MNAR-PMF	0.7251	0.9209	0.5962	0.7899
GPMF	**0.7207**	**0.9197**	**0.5788**	**0.7690**



(a) MovieLens-100k (b) MovieTweet

Figure 2: RMSE with varying degree of freedom for each method. The y-axis indicates RMSE. The degree of freedom in the x-axis is the number of parameters that the model optimizes.

We now discuss why GPMF achieves the higher accuracy for lower dimensional cases. GPMF aims to improve the accuracy by suppressing the samples made up of pairs of a user and an item if that item is unlikely to be consumed by that user according to the consumption model. Thus, the resampling has the effect of feature learning, analogous to clustering or dimensionality reduction, on the basis of users' attention: a user continues to consume the items of a topic that the user seeks to enjoy. By resampling, GPMF succeeds in learning a *good* representation of the features even in low dimensional cases. However, attention and resampling incurs a side effect in that it cannot model the case where a user tends to consume multiple topics. This drawback results in GPMF being inferior or comparable to the PMF and MNAR-PMF in high dimensional cases. There are a few possible ways of circumventing this disadvantage. One is a multi-topical consumption model such as LDA [Blei *et al.*, 2003; Hofmann, 2004]. Another is to model the temporal dynamics of the attention. Several studies [McAuley and Leskovec, 2013] have taken into account users' preferences that vary over time. We are planning to overcome the drawback of GPMF by utilizing some of the ideas in these approaches in our future work.

4 Related Work

Although our particular model of users' consumption and rating is new, prior works have studied other models of consumption behavior to improve the predictive accuracy of ratings. A particularly attractive property of GPMF that distinguishes it from existing techniques is that a regularization term in the objective function of training GPMF is naturally derived from the model of user's consumption. Here, we review the prior work related to ours primarily from two perspectives: one on consumption models and the other on regularization. We will also briefly review the prior work on attention-based models, which have been studied outside the literature on recommender systems but motivated our study on improving recommender systems by modeling users' attention.

Modeling Consumption Behavior Our method regularizes user preferences and item features on the basis of consumption behavior. Previous works use consumption behavior for smoothing user preferences and predicted rating values. Constrained PMF (CPMF) [Mnih and Salakhutdinov, 2007] models user preferences on the basis of ratings given by other users who have similar consumption behavior. There are many techniques based on the dependency between consumption and ratings such as CPT-v [Marlin and Zemel, 2009], Logit-vd [Marlin and Zemel, 2009], and MNAR-PMF [Hernandez-lobato *et al.*, 2014]. While Hu et al. (2008) aims to predict consumption from ratings, they do not apply their work for rating prediction.

Regularization GPMF adjusts the model complexity by estimating the variance of parameters in the training phase. While existing models have a regularization term in their objective function, the terms do not depend on the any implicit feedback from users. Hence, the fixed model complexity should be given prior to the training. PMF and several techniques of matrix factorization [Srebro *et al.*, 2004] have a regularization term that has fixed weight, which we should give preliminarily.

While several methods dynamically train the model complexity like ours does, they do not consider the consumption data. Salakhutdinov et al. (2008) proposed a model to adjust the model complexity by assuming hyperprior to prior distribution of parameters. Their method optimizes the parameter by Markov Chain Monte Carlo (MCMC). Lakshminarayanan et al. (2011) pointed out that PMF assumes homoscedasticity of the features and proposed a heteroscedastic model. Rendle et al. (2012) proposed a method to automatically adjust the hyperparameter of several matrix factorization methods. Although their method also fixes the number of parameters, it is able to adjust the model complexity by underestimating the variance of irrelevant features. They pointed out that a several matrix factorization method such as PMF can be generally formed by Factorization Machines [Rendle, 2010] and proposed a two-step model to tune hyperparameters by using a training set and a validation set. Nakajima et al. (2010) proposed the variational Bayesian approach to minimizing posteriori distribution of the parameters.

Although several works use auxiliary data such as social networks [Ma *et al.*, 2008; Jamali and Ester, 2011] and timestamps of ratings [Koren, 2010] to train the models, GPMF uses only observed ratings in order to ensure its applicability to standard collaborative filtering settings. Information of consumption behavior is included in the observed ratings. Pan et al. (2012) train the model from the possible range of rating values that we have to specify in advance.

Learning the Attention The attention, which we exploit in this literature, is the curious habit of humans to focus on a relevant feature from an object for a certain goal such as recognition [Rensink, 2000; Corbetta and Shulman, 2002]. The works motivated us to bring the attention into recommendation systems. In computer vision, the attention corresponds to saliency, which selects the relevant features by masking the input image. Sohn et al. (2013) proposed Gated Boltzmann machines consisting of a gate layer on top of restricted Boltzmann machines. Their method automatically selects the relevant features by masking the input. The gate corresponds to the salience of the human, and the gate separates the relevant and the irrelevant features from the data, such as numerical image data with real-world background. DRAW [Gregor *et al.*, 2015] is an image generation model that utilizes attention to model what a user sees. Xu et al. (2015) proposes the recurrent neural network-based model for image caption generation. The method generates a caption by moving attention to each object on an input image.

5 Conclusion and Future Work

We have proposed gated probabilistic matrix factorization (GPMF), which is based on a new probabilistic generative model, CRM, that models the attention of users to improve predictive accuracy on ratings. The key elements of CRM are the attention and attraction matrices, which play the role of selecting relevant features from the rating matrix. In the numerical experiment, GPMF outperformed PMF, MNAR-PMF, and other baseline methods with respect to MAE and RMSE.

GPMF is widely applicable to problems that can be reduced to rating prediction because it uses only the information of ratings given by users and does not use any auxiliary data (e.g., trust network and temporary dynamics, which would be required by other approaches). Although several researchers have proposed attention-based methods for a variety of tasks, this paper is the first study showing that an attention-based method is effective for collaborative filtering. An additional novelty of this paper is the way that we deal with missing values, making the ideas presented here of potential interest in other SVD-based methods such as biased SVD.

Acknowledgement

This research was supported by CREST, JST.

References

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *The Journal of machine Learning research*, 3:993–1022, 2003.

[Corbetta and Shulman, 2002] Maurizio Corbetta and Gordon L Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 3(3):201–215, 2002.

[Davenport and Beck, 2001] Thomas H Davenport and John C Beck. *The attention economy: Understanding the new currency of business*. Harvard Business Press, 2001.

[Dooms *et al.*, 2013] Simon Dooms, Toon De Pessemier, and Luc Martens. Movietweetings: a movie rating dataset collected from Twitter. In *Workshop on Crowdsourcing and Human Computation for Recommender Systems, CrowdRec at RecSys 2013*, 2013.

[Gregor *et al.*, 2015] Karol Gregor, Ivo Danihelka, Alex Graves, and Daan Wierstra. DRAW: A recurrent neural network for image generation. In *Proceedings of The 32th International Conference on Machine Learning (ICML-15)*, 2015.

[Hernandez-lobato *et al.*, 2014] Jose M Hernandez-lobato, Neil Houlsby, and Zoubin Ghahramani. Probabilistic matrix factorization with non-random missing data. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1512–1520, 2014.

[Hofmann, 2004] Thomas Hofmann. Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems (TOIS)*, 22(1):89–115, 2004.

[Hu *et al.*, 2008] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 263–272. IEEE, 2008.

[Jain *et al.*, 2013] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013.

[Jamali and Ester, 2011] Mohsen Jamali and Martin Ester. A transitivity aware matrix factorization model for recommendation in social networks. In *22nd International Joint Conference on Artificial Intelligence (IJCAI-11)*, 2011.

[Koren, 2010] Yehuda Koren. Collaborative filtering with temporal dynamics. *Communications of the ACM*, 53(4):89–97, 2010.

[Lakshminarayanan *et al.*, 2011] Balaji Lakshminarayanan, Guillaume Bouchard, and Cedric Archambeau. Robust bayesian matrix factorisation. In *International Conference on Artificial Intelligence and Statistics (AISTATS-11)*, pages 425–433, 2011.

[Ma *et al.*, 2008] Hao Ma, Haixuan Yang, Michael R Lyu, and Irwin King. Sorec: social recommendation using probabilistic matrix factorization. In *Proceedings of the 17th ACM conference on Information and knowledge management (CIKM-08)*, pages 931–940, 2008.

[Marlin and Zemel, 2009] Benjamin M Marlin and Richard S Zemel. Collaborative prediction and ranking with non-random missing data. In *Proceedings of the third ACM conference on Recommender systems (RecSys-09)*, pages 5–12, 2009.

[McAuley and Leskovec, 2013] Julian John McAuley and Jure Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web (WWW-13)*, pages 897–908, 2013.

[Mnih and Salakhutdinov, 2007] Andriy Mnih and Ruslan Salakhutdinov. Probabilistic matrix factorization. In *Advances in neural information processing systems*, pages 1257–1264, 2007.

[Nakajima and Sugiyama, 2010] Shinichi Nakajima and Masashi Sugiyama. Implicit regularization in variational bayesian matrix factorization. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 815–822, 2010.

[Pan *et al.*, 2012] Weike Pan, Evan Wei Xiang, and Qiang Yang. Transfer learning in collaborative filtering with uncertain ratings. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI-12)*, 2012.

[Rendle, 2010] Steffen Rendle. Factorization machines. In *2010 IEEE 10th International Conference on Data Mining (ICDM-10)*, pages 995–1000, 2010.

[Rendle, 2012] Steffen Rendle. Learning recommender systems with adaptive regularization. In *Proceedings of the 5th ACM international conference on Web search and data mining (WSDM-12)*, pages 133–142, 2012.

[Rensink, 2000] Ronald A Rensink. The dynamic representation of scenes. *Visual cognition*, 7(1-3):17–42, 2000.

[Salakhutdinov and Mnih, 2008] Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th international conference on Machine learning (ICML-08)*, pages 880–887, 2008.

[Schoormans and Robben, 1997] Jan PL Schoormans and Henry SJ Robben. The effect of new package design on product attention, categorization and evaluation. *Journal of Economic Psychology*, 18(2):271–287, 1997.

[Sohn *et al.*, 2013] Kihyuk Sohn, Guanyu Zhou, Chansoo Lee, and Honglak Lee. Learning and selecting features jointly with point-wise gated Boltzmann machines. In *Proceedings of The 30th International Conference on Machine Learning (ICML-13)*, pages 217–225, 2013.

[Srebro *et al.*, 2004] Nathan Srebro, Jason Rennie, and Tommi S Jaakkola. Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pages 1329–1336, 2004.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of The 32th International Conference on Machine Learning (ICML-15)*, 2015.