Bayesian Matrix Completion via Adaptive Relaxed Spectral Regularization

Yang Song[†] and **Jun Zhu**[‡]
[†]Department of Physics, Tsinghua University, yang.song@zoho.com
[‡]Department of Computer Science & Tech., State Key Lab of Intell. Tech. & Sys.; CBICR Center;
Tsinghua National Lab for Information Science and Tech., Tsinghua University, dcszj@mail.tsinghua.edu.cn

Abstract

Bayesian matrix completion has been studied based on a low-rank matrix factorization formulation with promising results. However, little work has been done on Bayesian matrix completion based on the more direct spectral regularization formulation. We fill this gap by presenting a novel Bayesian matrix completion method based on spectral regularization. In order to circumvent the difficulties of dealing with the orthonormality constraints of singular vectors, we derive a new equivalent form with relaxed constraints, which then leads us to design an adaptive version of spectral regularization feasible for Bayesian inference. Our Bayesian method requires no parameter tuning and can infer the number of latent factors automatically. Experiments on synthetic and real datasets demonstrate encouraging results on rank recovery and collaborative filtering, with notably good results for very sparse matrices.

Introduction

Matrix completion has found applications in many situations, such as collaborative filtering. Let $Z_{m\times n}$ denote the data matrix with m rows and n columns, of which only a small number of entries are observed, indexed by $\Omega \subset [m] \times [n]$. We denote the possibly noise corrupted observations of Z on Ω as $P_\Omega(X)$, where P_Ω is a projection operator that retains entries with indices from Ω and replaces others with 0. The matrix completion task aims at completing missing entries of Z based on $P_\Omega(X)$, under the low-rank assumption $\mathrm{rank}(Z) \ll \min(m, n)$. When a squared-error loss is adopted, it can be written as solving:

$$\min_Z \frac{1}{2\sigma^2} \|P_\Omega(X - Z)\|_F^2 + \lambda \,\mathrm{rank}(Z), \qquad \text{(P0)}$$

where $\|P_\Omega(A)\|_F^2 = \sum_{(i,j)\in\Omega} a_{ij}^2$; λ is a positive regularization parameter; and σ^2 is the noise variance.

Unfortunately, the term $\mathrm{rank}(Z)$ makes P0 NP-hard. Therefore, the nuclear norm $\|Z\|_*$ has been widely adopted as a convex surrogate (Fazel 2002) to the rank function to turn P0 to a convex problem:

$$\min_Z \frac{1}{2\sigma^2} \|P_\Omega(X - Z)\|_F^2 + \lambda \|Z\|_* . \qquad \text{(P1)}$$

Though P1 is convex, the definition of nuclear norm makes the problem still not easy to solve. Based on a variational formulation of the nuclear norm, it has been popular to solve an equivalent and easier low-rank matrix factorization (MF) form of P1:

$$\min_{A,B} \frac{1}{2\sigma}$$

t 200) t e e t e s t
t e t e s e e t t e e
e t e (2012 201) te
e t s t ( ) et
es et st s e e e
t st t e e e e t e ee t s t
e t s e 1 t tt e t t t
et s ee t e t e s e t s es

ese *spectral regularization* algorithms require optimization on a Stiefel manifold (Stiefel 1935; James 1976), which is defined as the set of k-tuples $(\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_k)$ of orthonormal vectors in $\mathbb{R}^n$. This is the main difficulty that has prevented the attempts, if any, to develop Bayesian methods based on the spectral regularization formulation.

Though matrix completion via spectral regularization is not easy, there are potential advantages over the matrix factorization approach. One of the benefits is the direct control over singular values. By imposing various priors on singular values, we can incorporate abundant information to help matrix completion. For example, Todeschini et al. (Todeschini, Caron, and Chavent 2013) put sparsity-inducing priors on singular values, naturally leading to hierarchical adaptive nuclear norm (HANN) regularization, and they reported promising results.

In this paper, we aim to develop a new formulation of the nuclear norm, hopefully having the same simplicity as MF and retaining some good properties of spectral regularization. The idea is to prove the *orthonormality insignifi-*

cance property of P1. Based on the new formulation, we develop a novel Bayesian model via a sparsity-inducing prior on singular values, allowing various dimensions to have different regularization parameters and automatically infer them. This involves some natural modifications to our new formulation to make it more flexible and adaptive, as people typically do in Bayesian matrix factorization. Empirical Bayesian methods are then employed to avoid parameter tuning. Experiments about rank recovery on synthetic matrices and collaborative filtering on some popular benchmark datasets demonstrate competitive results of our method in comparison with various state-of-the-art competitors. Notably, experiments on synthetic data show that our method performs considerably better when the matrices are very sparse, suggesting the robustness offered by using sparsity-inducing priors.

Relaxed Spectral Regularization

Bayesian matrix completion based on matrix factorization is relatively easy, with many examples (Lim and Teh 2007; Salakhutdinov and Mnih 2008). In fact, we can view (1) as a *maximum a posterior* (MAP) estimate of a simple Bayesian model, whose likelihood is Gaussian, i.e., for $(i,j) \in \Omega$, $X_{ij} \sim \mathcal{N}((AB^\intercal)_{ij}, \sigma^2)$, and the priors on A and B are also Gaussian, i.e., $p(A) \propto \exp(-\lambda \|A\|_F^2/2)$ and $p(B) \propto \exp(-\lambda \|B\|_F^2/2)$. It is now easy to do the posterior inference since the prior and likelihood are conjugate.

However, the same procedure faces great difficulty when we attempt to develop Bayesian matrix completion based on the more direct spectral regularization formulation P1. This is because the prior $p(Z) \propto \exp(-\lambda \|Z\|_*)$ is not conjugate to the Gaussian likelihood (or any other common likelihood). To analyze $p(Z)$ more closely, we can conduct singular value decomposition (SVD) on Z to get $Z = \sum_{k=1}^r d_k \mathbf{u}_k \mathbf{v}_k^\intercal$, where $\vec{d} := \{d_k : k \in [r]\}$ are singular values; $U := \{\mathbf{u}_k : k \in [r]\}$ and $V := \{\mathbf{v}_k : k \in [r]\}$ are orthonormal singular vectors on Stiefel manifolds. Though we can define a factorized prior $p(Z) = p(\vec{d})p(U)p(V)$, any prior on U or V (e.g., the uniform Haar prior (Todeschini, Caron, and Chavent 2013)) needs to deal with a Stiefel manifold, which is highly nontrivial.

In fact, handling distributions embedded on Stiefel manifolds still remains a largely open problem, though some results (Byrne and Girolami 2013; Hoff 2009; Dobigeon and Tourneret 2010) exist in the literature of directional statistics. Fortunately, as we will prove in Theorem 1 that the orthonormality constraints on U and V are not necessary for spectral regularization. Rather, the unit sphere constraints $\|\mathbf{u}_k\| \le 1$ and $\|\mathbf{v}_k\| \le 1$, for all $k \in [r]$, are sufficient to get the same optimal solutions to P1. We call this phenomenon *orthonormality insignificance*. We will call spectral regularization with orthonormality constraints relaxed by unit sphere constraints *relaxed spectral regularization*.

Orthonormality insignificance for spectral regularization

We now present an equivalent formulation of the spectral regularization in P1 by proving its orthornormality insignificance property.

With the SVD of Z, we first rewrite P1 equivalently as P1′ to show all constraints explicitly:

$$\min_{\vec{d},U,V} \frac{1}{2\sigma^2} \left\| P_\Omega\left(X - \sum_{k=1}^r d_k \mathbf{u}_k \mathbf{v}_k^\intercal\right)\right\|_F^2 + \lambda \sum_{k=1}^r d_k \quad \text{(P1′)}$$
$$s.t. \quad d_k \ge 0, \quad \|\mathbf{u}_k\| = 1, \quad \|\mathbf{v}_k\| = 1, \quad \forall k \in [r]$$
$$\mathbf{u}_i^\intercal \mathbf{u}_j = 0, \quad \mathbf{v}_i^\intercal \mathbf{v}_j = 0, \quad \forall i,j \in [r] \text{ and } i \ne j,$$

where $r = \min(m,n)$. Then, we can have an equivalent formulation of P1 as summarized in Theorem 1, which lays the foundation for the validity of relaxed spectral regularization.

Theorem 1. *Let s be the optimal value of P1 (P1′), and let t be the optimal value of P2 as defined below:*

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\vec{d}} \frac{1}{2\sigma^2} \left\| P_\Omega\left(X - \sum_{k=1}^r d_k \boldsymbol{\alpha}_k \boldsymbol{\beta}_k^\intercal\right)\right\|_F^2 + \lambda \sum_{k=1}^r d_k \quad \text{(P2)}$$
$$s.t. \quad d_k \ge 0, \quad \|\boldsymbol{\alpha}_k\|_2 \le 1, \quad \|\boldsymbol{\beta}_k\|_2 \le 1, \quad \forall k \in [r],$$

Then, we have $s = t$. Furthermore, suppose an optimal solution for P2 is $(\vec{d}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$, then $Z^* = \sum_{k=1}^r d_k^* \boldsymbol{\alpha}_k^* \boldsymbol{\beta}_k^{*\intercal}$ is also an optimal solution for P1. Similarly, for any optimal solution $Z^\dagger$ for P1, there exists a decomposition $Z^\dagger = \sum_{k=1}^r d_k^\dagger \boldsymbol{\alpha}_k^\dagger \boldsymbol{\beta}_k^{\dagger\intercal}$ optimal for P2.

Sketch of the proof. Let $Z^* = \sum_{k=1}^r d_k^* \boldsymbol{\alpha}_k^* \boldsymbol{\beta}_k^{*\intercal}$ be an optimal solution for P2 with the optimal value t. Since P1′ is basically the same optimization problem as P2 with stricter constraints, we have $s \ge t$.

Conduct singular value decomposition to obtain $Z^* = \sum_{k=1}^r \sigma_k^* \mathbf{u}_k^* \mathbf{v}_k^{*\intercal}$ and we can prove that $\|Z^*\|_* = \sum_{k=1}^r \sigma_k^* \le \sum_{k=1}^r d_k^*$. If $\sum_{k=1}^r \sigma_k^* < \sum_{k=1}^r d_k^*$, then we can plug Z^* into P1 to get a smaller value than t, contradicting $s \ge t$. As a result, $\sum_{k=1}^r \sigma_k^* = \sum_{k=1}^r d_k^*$ and $s = t$.

Furthermore, since $s = t$ and plugging Z^* into P1 can lead to a value at least as small as t, we conclude that Z^* is also an optimal solution for P1. Let $Z^\dagger$ be any optimal solution for P1, we can also prove that there is a decomposition $Z^\dagger = \sum_{k=1}^r d_k^\dagger \boldsymbol{\alpha}_k^\dagger \boldsymbol{\beta}_k^{\dagger\intercal}$ to be an optimal solution for P2.

The formal proof and some remarks are provided in the supplementary material[1]. □

Now we have justified the orthonormality insignificance property of spectral regularization. As a result, P2 serves to be another equivalent form of P1, similar to the role played by MF. This *relaxed spectral regularization* formulation lies somewhere between MF and spectral regularization, since it has more (but easily solvable) contraints than MF and still retains the form of SVD. As discussed before, it is easier to conduct Bayesian inference on a posterior without strict orthonormality constraints, and therefore models on relaxed spectral regularization are our focus of investigation.

In addition, Theorem 1 can be generalized to *arbitrary* loss besides squared-error loss, which means it is as widely applicable as MF. See Remark 2 in the supplementary material for more details.

[1] See http://bigml.cs.tsinghua.edu.cn/~yangsong/papers/AAAI/GASR-supp.pdf

Adaptive relaxed spectral regularization

Based on the relaxed spectral regularization formulation in Theorem 1, a Bayesian matrix completion algorithm similar to BPMF can be directly derived. Let the priors of $\boldsymbol{\alpha}_k$, $\boldsymbol{\beta}_k$ be uniform Haar priors within unit spheres; and the prior of d_k's to be exponential distributions, then the posterior has exactly the same form as P2. Such an algorithm should have similar performance to BPMF.

Instead of building models on P2 exactly, we consider another modified form where each d_k has its own positive regularization parameter γ_k. Obviously this is a generalization of relaxed spectral regularization and admits it as a special case. We define the adaptive relaxed spectral regularization problem as

$$\min_{\vec{d},U,V} \frac{1}{2\sigma^2} \left\| P_\Omega \left(X - \sum_{k=1}^{r} d_k \mathbf{u}_k \mathbf{v}_k^\intercal \right) \right\|_F^2 + \sum_{k=1}^{r} \gamma_k d_k \quad \text{(P3)}$$
$$s.t. \quad d_k \geq 0, \quad \|\mathbf{u}_k\|_2 \leq 1, \quad \|\mathbf{v}_k\|_2 \leq 1, \quad k \in [r].$$

Such a variation is expected to be more flexible and better at bridging the gap between the nuclear norm and the rank functions, thus being more capable of approximating rank regularization than the standard nuclear norm. Similar ideas arose in (Todeschini, Caron, and Chavent 2013) before and was called *hierarchical adaptive nuclear norm* (HANN). But note that although we propose a similar approach to HANN, our regularization is substantially different because of the relaxed constraints.

However, P3 may be harder to solve than the original P2 due to the difficulty in hyperparameter tuning, since adaptive regularization introduces dramatically more hyperparameters. We will build hierarchical priors for these hyperparameters and derive a Bayesian algorithm for solving P3 and inferring hyperparameters simultaneously in the following section.

Bayesian Matrix Completion via Adaptive Relaxed Spectral Regularization

Probabilistic model

We now turn P3 into an equivalent MAP estimation task. Naturally, the squared-error loss in P3 corresponds to the negative logarithm of the Gaussian likelihood, $X_{ij} \sim \mathcal{N}(\sum_{k=1}^{r} d_k u_{ki} v_{kj}, \sigma^2)$, where u_{ki} denotes the ith term in $\mathbf{u}_k$; likewise for v_{kj}. For priors, we adopt uniform Haar priors on U and V within unit spheres, and exponential priors on $\vec{d}$, as summarized below:

$$\tilde{p}(\mathbf{u}_k) = \begin{cases} 1, & \|\mathbf{u}_k\| \leq 1 \\ 0, & \|\mathbf{u}_k\| > 1 \end{cases}, \quad \forall k \in [r]$$

$$\tilde{p}(\mathbf{v}_k) = \begin{cases} 1, & \|\mathbf{v}_k\| \leq 1 \\ 0, & \|\mathbf{v}_k\| > 1 \end{cases}, \quad \forall k \in [r]$$

$$p(d_k \mid \gamma_k) = \gamma_k e^{-\gamma_k d_k}, \quad d_k \geq 0, \quad \forall k \in [r]$$

where $\tilde{p}$ denotes an unnormalized probability density function (p.d.f.). It can be checked that under this probabilistic model, the negative log posterior p.d.f. with respect to $(\vec{d}, U, V)$ is exactly proportional to P3.

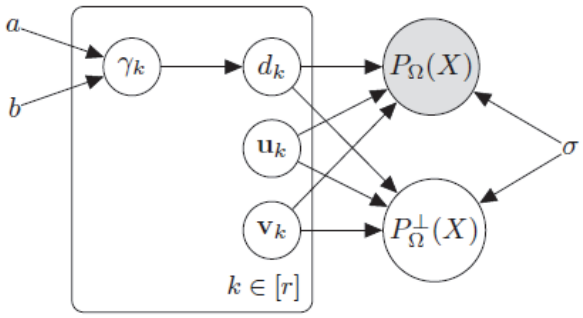

Figure 1: The graphical model for adaptive relaxed spectral regularization, where $P_\Omega(A) = \{a_{ij} \mid (i,j) \in \Omega\}$ and $P_\Omega^\perp(A) = \{a_{ij} \mid (i,j) \notin \Omega\}$.

Now we precede to treat regularization coefficients $\vec{\gamma} := \{\gamma_k : k \in [r]\}$ as random variables and assume gamma priors on them, i.e., $p(\gamma_k) \propto \gamma_k^{a-1} e^{-b\gamma_k}$, $\gamma_k \geq 0$, $\forall k \in [r]$. This has two obvious advantages. First, it includes $\vec{\gamma}$ into the Bayesian framework so that values of these coefficients can be inferred automatically without being tuned as hyperparameters. Second, the prior on d_k after marginalizing out γ_k's becomes $p(d_k) = \int_0^\infty p(d_k \mid \gamma_k) p(\gamma_k) \mathrm{d}\gamma_k = \frac{ab^a}{(d_k+b)^{a+1}}$, which is effectively a Pareto distribution. This distribution has a heavier tail compared to the exponential distribution (Todeschini, Caron, and Chavent 2013), and is therefore expected to be better at sparsity-inducing (Bach et al. 2012).

The graphical model is shown in Figure 1, where we explicitly separate the observed entries of X, i.e., $P_\Omega(X)$, and the unobserved ones, i.e., $P_\Omega^\perp(X)$. Due to the conditional independency structure, we can simply marginalize out $P_\Omega^\perp(X)$ and get the joint distribution

$$p(\vec{d}, U, V, \vec{\gamma}, P_\Omega(X) \mid a, b, \sigma)$$
$$\propto \left(\frac{1}{2\sigma^2}\right)^{|\Omega|/2} \exp\left[-\frac{1}{2\sigma^2} \left\| P_\Omega(X - \sum_{k=1}^{r} d_k \mathbf{u}_k \mathbf{v}_k^\intercal) \right\|_F^2 \right]$$
$$\cdot \prod_{k=1}^{r} \frac{b^a}{\Gamma(a)} \gamma_k^a e^{-(b+d_k)\gamma_k}, \quad (2)$$

with all the variables implicitly constrained to their corresponding valid domains.

Inference

We now present the GASR (Gibbs sampler for Adaptive Relaxed Spectral Regularization) algorithm to infer the posterior, make predictions, and estimate the hyperparameters via Monte Carlo EM (Casella 2001).

Posterior Inference Let $\mathcal{N}(\mu, \sigma^2; a, b)$ denote the normal distribution $\mathcal{N}(\mu, \sigma^2)$ truncated in $[a, b]$. We infer the posterior distribution $p(\vec{\gamma}, \vec{d}, U, V \mid a, b, \sigma, P_\Omega(X))$ via a Gibbs sampler as explained below:

Sample $\vec{\gamma}$: The conditional distributions for regularization coefficients $\vec{\gamma}$ are gamma distributions. We sample $\vec{\gamma}$ by the formula $\gamma_k \sim \Gamma(a+1; b+d_k), \quad k \in [r]$.

Sample $\vec{d}$: Conditioned on $(\vec{\gamma}, U, V)$, the distribution for each d_α ($\alpha \in [r]$) is a truncated Gaussian, $d_\alpha \sim \mathcal{N}(-\frac{B}{A}, \frac{\sigma^2}{A}; 0, \infty)$, where $A = \sum_{(i,j)\in\Omega} \left(u_{\alpha i} v_{\alpha j}\right)^2$ and $B = \sum_{(i,j)\in\Omega} \left(\sum_{k\neq\alpha} d_k u_{\alpha i} u_{ki} v_{\alpha j} v_{kj} - X_{ij} u_{\alpha i} v_{\alpha j}\right) + \sigma^2 \gamma_\alpha$.

Sample U and V: Given the other variables, the distribution for each element in $\mathbf{u}_\alpha$'s (or $\mathbf{v}_\alpha$'s) is a truncated Gaussian, $u_{\alpha\beta} \sim \mathcal{N}\left(-\frac{D}{C}, \frac{\sigma^2}{C}; -\rho, \rho\right)$, $\alpha \in [m]$, $\beta \in [r]$, where $C = \sum_{(\beta,j)\in\Omega} d_\alpha^2 v_{\alpha j}^2$, $D = \sum_{(\beta,j)\in\Omega} \left(\sum_{k\neq\alpha} d_\alpha d_k u_{k\beta} v_{\alpha j} v_{kj} - d_\alpha X_{\beta j} v_{\alpha j}\right)$ and $\rho = \sqrt{1 - \sum_{k\neq\beta} u_{\alpha k}^2}$. A similar procedure can be derived to sample $v_{\alpha\beta}$ and is therefore omitted here.

The time complexity of this Gibbs sampler is $O(|\Omega| r^2)$ per iteration. Although there is a unified scheme on sampling truncated distributions by cumulative distribution function (c.d.f.) inversion, we did not use it due to numerical instabilities found in experiments. In contrast, simple rejection sampling methods prove to work well.

Prediction With the posterior distribution, we can complete the missing elements using the posterior mean:

$$\begin{aligned}
&\mathbb{E}\left[P_\Omega^\perp(X) \mid P_\Omega(X), a, b, \sigma\right] \\
&= \quad \cdots \quad \mathbb{E}\overset{\backprime}{} \; P_\Omega^\perp(X) \mid P_\Omega(X), a, b, \sigma, \vec{\gamma}, U, V, \vec{d} \\
&\cdot p(\vec{\gamma}, \vec{d}, U, V \mid a, b, \sigma, P_\Omega(X)) \mathrm{d}\vec{\gamma} \mathrm{d}\vec{d} \mathrm{d}U \mathrm{d}V.
\end{aligned}$$

This integral is intractable. But we can use samples to approximate the integral and complete the matrix. Since we use the Gaussian likelihood, we have

$$\mathbb{E}\overset{\backprime}{} \; P_\Omega^\perp(X) \mid P_\Omega(X), a, b, \sigma, \vec{\gamma}, U, V, \vec{d} = \sum_{k=1}^{r} d_k \mathbf{u}_k \mathbf{v}_k^\intercal.$$

As a result, we can represent missing elements as $x_{ij} = \langle \sum_{k=1}^{r} d_k u_{ki} v_{kj} \rangle, (i,j) \in \Omega^\perp$, which is the posterior sample mean of $P_\Omega^\perp(X)$. Here we denote the sample mean for $f(x)$ as $\langle f(x) \rangle := \frac{1}{n} \sum_{i=1}^{n} f(x_i)$, with x_i being individual samples and n being the number of samples.

Hyperparameter Estimation We choose the hyperparameters (a, b, σ) by maximizing model evidence $p(P_\Omega(X) \mid a, b, \lambda)$. Since direct optimization is intractable, we adopt an EM algorithm, with latent variable $L := (\vec{d}, U, V, \vec{\gamma})$. In order to compute the joint expectation with respect to $P_\Omega(X)$ and L, we use Monte Carlo EM (Casella 2001), which can fully exploit the samples obtained in the Gibbs sampler.

The expectation of $P_\Omega(X)$ and L with respect to $p(L \mid P_\Omega(X))$ can be written as

$$\begin{aligned}
&\mathbb{E}_{p(L|P_\Omega(X))}\left[\ln p(P_\Omega(X), L)\right] \\
&= \mathbb{E}\overset{\backprime}{} \; \ln p(\vec{d}, U, V, \vec{\gamma}, P_\Omega(X) \mid a, b, \sigma) \\
&\quad -|\Omega| \ln \sigma - \frac{1}{2} \left\| P_\Omega\left(X - \sum_{k=1}^{r} d_k \mathbf{u}_k \mathbf{v}_k^\intercal\right)\right\|_F^2 \\
&\quad + \sum_{i=1}^{r} [a \ln b - \ln \Gamma(a) + a\langle \ln \gamma_i \rangle - b\langle \gamma_i \rangle] + C, \qquad (3)
\end{aligned}$$

where C is a constant. Eq. (3) can be maximized with respect to a, b, σ using Newton–Raphson method. The fixpoint equations are

$$\begin{aligned}
a_{t+1} &= a_t - \frac{\Psi(a_t) - \ln\left(r a_t / \sum_i \langle \gamma_i \rangle\right) - \sum_i \langle \ln \gamma_i \rangle / r}{\Psi'(a_t) - 1/a_t} \\
b_{t+1} &= \frac{r a_{t+1}}{\sum_{i=1}^{r} \langle \gamma_i \rangle} \\
\sigma^2 &= \frac{1}{|\Omega|} \left\| P_\Omega\left(X - \sum_{k=1}^{r} d_k \mathbf{u}_k \mathbf{v}_k^\intercal\right)\right\|_F^2,
\end{aligned}$$

where $\Psi(x)$ and $\Psi'(x)$ are digamma and trigamma functions respectively. In our experiments, we found the results not very sensitive to the number of samples used in $\langle \cdot \rangle$, so we fixed it to 5.

Experiments

We now present experimental results on both synthetic and real datasets to demonstrate the effectiveness on rank recovery and matrix completion.

Experiments on synthetic data

We have two experiments on synthetic data, one is for rank recovery, the other is for examining how well the algorithms perform in situations that matrices are very sparse.

In both experiments, we generate standard normal random matrices $A_{m\times q}$ and $B_{n\times q}$ and produce a rank-q matrix $Z = AB^\intercal$. Then we corrupt Z with standard Gaussian noise to get the observations X, using a signal to noise ratio of 1.

Rank recovery In this experiment, we set $m = 10q$ and $n = 10q$. The algorithm was tested with q ranging from 5 to 45. We set the rank truncation r to 100, which is sufficiently large for all data. For each matrix Z, the iteration number was fixed to 1000 and the result was averaged from last 200 samples (with first 800 discarded as burn-in). We simply initialize our sampler with uniformly distributed U and V with norms fixed to 0.9 and all $\vec{d}$ fixed to zero. We run our Gibbs sampler on all the entries of X to recover Z.

In the spectral regularization formulation, we can get the number of latent factors by simply counting the number of nonzero d_k's. However, since our method uses MCMC sampling, it is diffcult to find some d_k to vanish exactly. Instead of counting nonzero elements of $(d_1, d_2, \cdots, d_r)$ directly, we sort the tuple in an ascending order and try to locate $w = \arg\max_{k\geq 2} d_k / d_{k-1}$ and then discard the set

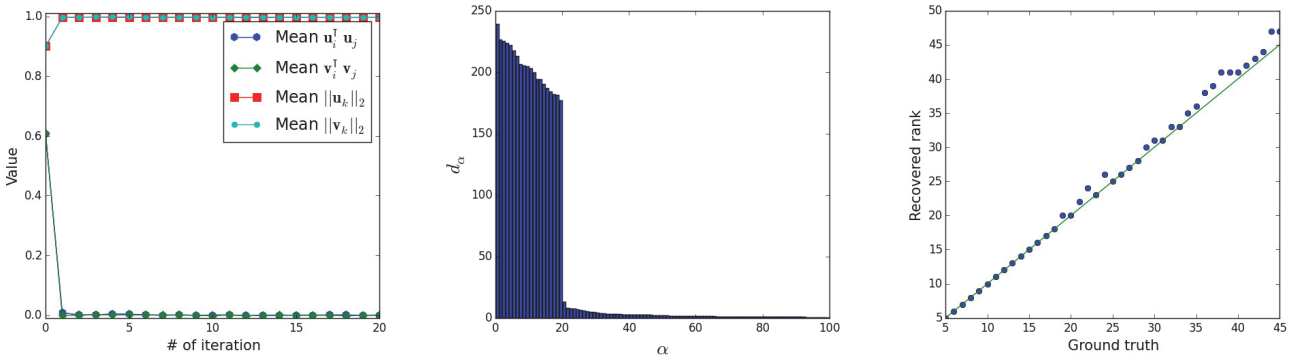

Figure 2: (a) Experiment on 200×200 synthetic matrix shows that vectors tend to get orthonormalized. We measure the mean values of inner products and norms of all column vectors in U and V for each single iteration. The fluctuations at the beginning are due to initialization. (b) The recovered rank of a 200×200 synthetic matrix with rank 20. Following the principle presented in main text, we conclude that the number of latent factors is 20, matching the ground truth. (c) Rank recovery results on synthetic data, with solid line representing the ground truth.

$\{d_k : d_k < d_w\}$. As a result, the recovered rank is $r - w + 1$. The middle panel of Figure 2 provides an example about how to determine the number of latent factors.

The results of this experiment are summarized in Figure 2(c), showing that the recovered ranks are fairly aligned with the ground truth. Our algorithm performs perfectly well when the true rank is relatively small and slightly worse when the rank gets higher. This may be due to that larger rank requires more iterations for convergence.

We also illustrate how vectors get orthonormalized automatically on one of our synthetic matrices in Figure 2(a). The orthonormality of vectors are measured by the average values of their ℓ_2 norms and inner products with each other. Figure 2(a) shows that U and V get nearly orthonormalized after only one iteration. This phenomenon indicates that the vectors still tend to get orthonormalized even in the hierarchial Bayesian model.

Different missing rates We generate matrices of different sizes and different missing rates to test the performance of our method, in comparison with BPMF, as it is the only one that can compete with GASR on real datasets, as illustrated in detail in the next section.

The *Root Mean Squared Error* (RMSE) results are listed in Table 1. The deviations and some additional settings are reported in supplementary material. We can see that GASR is considerably better than BPMF when the observed elements of a matrix are of a small number, demonstrating the robust estimates of GASR via sparsity-inducing priors.

Collaborative filtering on real datasets

We test our algorithm on the MovieLens 1M[2] and EachMovie datasets, and compare results with various strong competitors, including max-margin matrix factorization (M^{3}F) (Rennie and Srebro 2005), infinite probabilistic max-margin matrix factorization (iPM^{3}F) (Xu, Zhu, and Zhang 2012), softImpute (Mazumder, Hastie, and Tibshirani 2010), softImpute-ALS ("ALS" for "alternating least squares") (Hastie et al. 2014), hierarchical adaptive soft impute (HASI) (Todeschini, Caron, and Chavent 2013) and Bayesian probabilistic matrix factorization (BPMF) (Salakhutdinov and Mnih 2008).

Table 1: Results on different missing rates

Setting	$m = 500, n = 500, r = 30, q = 5$			
Missing-Rates	90%	80%	50%	0%
BPMF	1.6842	0.3210	0.1304	0.0933
GASR	0.1992	0.1321	0.0841	0.0724
Setting	$m = 1000, n = 1000, r = 50, q = 10$			
Missing-Rates	90%	80%	50%	0%
BPMF	0.9422	0.2396	0.1105	0.0859
GASR	0.2513	0.1688	0.1270	0.1115

The MovieLens 1M dataset contains 1,000,209 ratings provided by 6,040 users on 3,952 movies. The ratings are integers from $\{1, 2, 3, 4, 5\}$ and each user has at least 20 ratings. The EachMovie dataset consists of 2,811,983 ratings provided by 74,424 users on 1,648 movies. As in (Marlin 2004), we removed duplicates and discarded users with less than 20 ratings. This left us with 36,656 users. There are 6 possible ratings from 0 to 1 and we mapped them to $\{1, 2, \cdots, 6\}$.

Protocol: We randomly split the dataset into 80% training and 20% test. We further split 20% training data for validation for M^{3}F, iPM^{3}F, SoftImpute, SoftImpute-ALS and HASI to tune their hyperparameters. BPMF and GASR can infer hyperparameters from training data and thus do not need validation. We measure the performance using both RMSE and *normalized mean absolute error* (NMAE),

[2]MovieLens datasets can be downloaded from http://grouplens.org/datasets/movielens/.

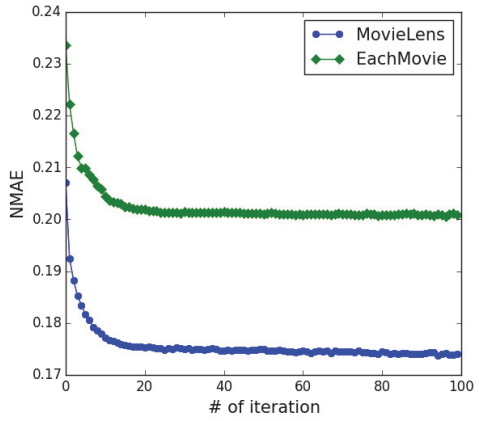

Figure 3: Convergence of NMAE on real datasets. We only show results on our first data partitions.

where NMAE (Goldberg et al. 2001) is defined as

$$\text{NMAE} = \frac{1}{|\Omega_{\text{test}}|} \frac{\sum_{(i,j)\in\Omega_{\text{test}}} |X_{ij} - Z_{ij}|}{\max(X) - \min(X)}, \qquad (4)$$

and Ω_{test} is the index set of entries for testing.

Implementation details: The number of iterations of our sampler is fixed to 100 and the point estimates of $\mathbb{E}\left[P_\Omega^\perp(X)\right]$ are taken from the average of all 100 samples. We initialize our algorithm by generating uniformly distributed items for U and V and set all d_k to 0. We also scale the norms of $\mathbf{u}_k$ and $\mathbf{v}_k$ to 0.9 for initialization. Figure 3 shows that our sampler converges after a few iterations with this fairly naïve initialization.

We use the R package `softImpute` for SoftImpute and SoftImpute-ALS, and use the code provided by the corresponding authors for M^{3}F, iPM^{3}F, HASI and BPMF. The hyperparameters for M^{3}F, iPM^{3}F, SoftImpute, SoftImpute-ALS and HASI are selected via grid search on the validation set. We randomly initialize all methods except HASI for which the initialization is the result of SoftImpute, as suggested in (Todeschini, Caron, and Chavent 2013). The results of BPMF are the averages over 100 samples, the same as ours.

For all the algorithms, we set the maximum number of iterations to 100. The rank truncation r for MovieLens 1M and EachMovie is set to 30, where we follow the setting in (Todeschini, Caron, and Chavent 2013) and find in experiments that larger r does not lead to significant improvements.

Results: Table 2 presents the averaged NMAE and RMSE over 5 replications and their standard deviations.[3] Overall, we can see that our GASR achieves superior performance than most of the baselines. More specifically, we have the following observations:

[3]The NMAE of HASI on MovieLens is slightly different from that in (Todeschini, Caron, and Chavent 2013), which was 0.172, still inferior to ours. This may be due to differences in parameter selecting methods.

First, GASR is comparable to BPMF, the state-of-the-art Bayesian method for low-rank matrix completion, on the MovieLens dataset; while it outperforms BPMF on the EachMovie dataset, with the observation that EachMovie dataset (97.8% missing) is sparser than MovieLens (95.8% missing). On both datasets, GASR also obtains much lower RMSE than iPM^{3}F, a state-of-the-art nonparametric Bayesian method based on IBP (Griffiths and Ghahramani 2011) for matrix completion. Such results demonstrate the promise of performing Bayesian matrix completion via spectral regularization. Furthermore, GASR produces sparser solutions due to its sparsity-inducing priors on $\vec{d}$. The ranks it infers on MovieLens and EachMovie are both 10 on average, but the numbers of latent factors inferred by iPM^{3}F are both 30, which is the rank truncation level. It is reported in (Xu, Zhu, and Zhang 2013) with a similar setting that the optimal latent dimensions inferred by Gibbs iPM^{3}F, a Gibbs sampling version for iPM^{3}F model without rank truncations, are around 450 for MovieLens and 200 for EachMovie, which are much larger than ours.

Second, compared to HASI, a non-Bayesian method that adopts similar adaptive spectral regularization, and the other non-Bayesian methods based on squared-error losses (i.e., SoftImpute and SoftImpute-ALS), we achieve much better results on both datasets, demonstrating the advantages of Bayesian inference. Moreover, the better performance of HASI over SoftImpute demonstrates the benefit of adaptivity.

Finally, the max-margin based methods (i.e., M^{3}F and iPM^{3}F) have slightly better performance on NMAE but much worse results on RMSE than our GASR. One possible reason is that these methods are based on the maximum-margin criterion, which naturally minimizes absolute errors, while our method (and the others) is based on minimizing a squared-error loss. Another reason, which may be the most important one, is that both M^{3}F and iPM^{3}F predict integer values while our method (and the others) gives real value predictions. We found that simply rounding these real value predictions to integers can greatly improve the performance on NMAE. For example, our GASR gives NMAE value 0.1569 ± 0.0006 and 0.1877 ± 0.0003 respectively on MovieLens and EachMovie datasets after rounding predictions to nearest integers.

Conclusions and Discussions

We present a novel Bayesian matrix completion method with adaptive relaxed spectral regularization. Our method exhibits the benefits of hierarchical Bayesian methods on inferring the parameters associated with adaptive relaxed spectral regularization, thereby avoiding parameter tuning. We estimate hyperparameters using Monte Carlo EM. Our Gibbs sampler exhibits good performance both in rank inference on synthetic data and collaborative filtering on real datasets.

Our method is based on a new formulation in Theorem 1. These results can be further generalized to other noise potentials with minor effort. For the Gibbs sampler, we can also extend to non-Gaussian potentials as long as this potential has a regular p.d.f. that enables efficient sampling.

Table 2: Experimental results of various methods on the MovieLens and EachMovie datasets.

	MovieLens		EachMovie	
Algorithm	**NMAE**	**RMSE**	**NMAE**	**RMSE**
M^3F	0.1652 ± 0.0004	0.9644 ± 0.0012	0.1932 ± 0.0003	1.4598 ± 0.0019
iPM^3F	0.1604 ± 0.0004	0.9386 ± 0.0024	0.1819 ± 0.0006	1.3760 ± 0.0034
SoftImpute	0.1829 ± 0.0002	0.9469 ± 0.0009	0.2039 ± 0.0002	1.2948 ± 0.0037
SoftImpute-ALS	0.1783 ± 0.0001	0.9196 ± 0.0013	0.2018 ± 0.0004	1.2757 ± 0.0008
HASI	0.1813 ± 0.0002	0.9444 ± 0.0011	0.1992 ± 0.0003	1.2612 ± 0.0016
BPMF	0.1663 ± 0.0002	0.8460 ± 0.0006	0.2012 ± 0.0001	1.2363 ± 0.0007
GASR	0.1673 ± 0.0005	0.8528 ± 0.0025	0.1930 ± 0.0009	1.2015 ± 0.0044

Finally, though we stick to Gibbs sampling in this paper, it would be interesting to investigate other Bayesian inference methods based on Theorem 1 since it gets rid of many difficulties related to Stiefel manifolds. Such an investigation may lead to more scalable algorithms with better convergence property. Furthermore, better initialization methods other than uniformly generated random numbers may lead to much faster convergence, e.g., results from several iterations of HASI can usually provide a good starting point.

Acknowledgments

The work was supported by the National Basic Research Program (973 Program) of China (Nos. 2013CB329403, 2012CB316301), National NSF of China (Nos. 61322308, 61332007), Tsinghua National Laboratory for Information Science and Technology Big Data Initiative, and Tsinghua Initiative Scientific Research Program (No. 20141080934). We thank the Department of Physics at Tsinghua University for covering part of the travel costs.

References

Bach, F.; Jenatton, R.; Mairal, J.; and Obozinski, G. 2012. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning* 4(1):1–106.

Byrne, S., and Girolami, M. 2013. Geodesic Monte Carlo on embedded manifolds. *Scandinavian Journal of Statistics* 40(4):825–845.

Casella, G. 2001. Empirical Bayes Gibbs sampling. *Biostatistics* 2(4):485–500.

Dobigeon, N., and Tourneret, J.-Y. 2010. Bayesian orthogonal component analysis for sparse representation. *IEEE Transactions on Signal Processing* 58(5):2675–2685.

Fazel, M. 2002. *Matrix rank minimization with applications*. Ph.D. Dissertation, Stanford University.

Goldberg, K.; Roeder, T.; Gupta, D.; and Perkins, C. 2001. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval* 4(2):133–151.

Griffiths, T. L., and Ghahramani, Z. 2011. The Indian buffet process: An introduction and review. *Journal of Machine Learning Research* 12:1185–1224.

Hastie, T.; Mazumder, R.; Lee, J.; and Zadeh, R. 2014. Matrix completion and low-rank SVD via fast alternating least squares. *arXiv preprint arXiv:1410.2596*.

Hoff, P. D. 2009. Simulation of the matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics* 18(2).

James, I. M. 1976. *The topology of Stiefel manifolds*, volume 24. Cambridge University Press.

Lim, Y. J., and Teh, Y. W. 2007. Variational Bayesian approach to movie rating prediction. In *Proceedings of KDD Cup and Workshop*, volume 7, 15–21. Citeseer.

Marlin, B. 2004. *Collaborative filtering: A machine learning perspective*. Ph.D. Dissertation, University of Toronto.

Mazumder, R.; Hastie, T.; and Tibshirani, R. 2010. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research* 11:2287–2322.

Rennie, J. D., and Srebro, N. 2005. Fast maximum margin matrix factorization for collaborative prediction. In *International Conference on Machine Learning*, 713–719.

Salakhutdinov, R., and Mnih, A. 2008. Bayesian probabilistic matrix factorization using MCMC. In *International Conference on Machine Learning*.

Srebro, N.; Rennie, J.; and Jaakkola, T. S. 2004. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems*, 1329–1336.

Stiefel, E. 1935. Richtungsfelder und fernparallelismus in n-dimensionalen mannigfaltigkeiten. *Commentarii Mathematici Helvetici* 8(1):305–353.

Todeschini, A.; Caron, F.; and Chavent, M. 2013. Probabilistic low-rank matrix completion with adaptive spectral regularization algorithms. In *Advances in Neural Information Processing Systems*, 845–853.

Xu, M.; Zhu, J.; and Zhang, B. 2012. Nonparametric max-margin matrix factorization for collaborative prediction. In *Advances in Neural Information Processing Systems*, 64–72.

Xu, M.; Zhu, J.; and Zhang, B. 2013. Fast max-margin matrix factorization with data augmentation. In *International Conference on Machine Learning*, 978–986.