

A Bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution

The Tien Mai

School of Mathematical Sciences, University College Dublin

Belfield, Dublin 4 – Ireland;

Insight Centre for Data Analytics, Ireland

e-mail: mai.thetien@insight-centre.org

url: <http://sites.google.com/site/thetienmai/>

and

Pierre Alquier

ENSAE-CREST-GENES

3 avenue Pierre Larousse

92245 Malakoff CEDEX – France

e-mail: pierre.alquier@ensae.fr

url: <http://alquier.ensae.net/>

Abstract: Bayesian methods for low-rank matrix completion with noise have been shown to be very efficient computationally [3, 18, 19, 24, 28]. While the behaviour of penalized minimization methods is well understood both from the theoretical and computational points of view (see [7, 9, 16, 23] among others) in this problem, the theoretical optimality of Bayesian estimators have not been explored yet. In this paper, we propose a Bayesian estimator for matrix completion under general sampling distribution. We also provide an oracle inequality for this estimator. This inequality proves that, whatever the rank of the matrix to be estimated, our estimator reaches the minimax-optimal rate of convergence (up to a logarithmic factor). We end the paper with a short simulation study.

MSC 2010 subject classifications: Primary 62H12; secondary 62J12, 60B20, 68T05.

Keywords and phrases: Matrix completion, Bayesian Analysis, PAC-Bayesian bounds, oracle inequality, low-rank matrix, Gibbs sampler.

Received August 2014.

Contents

1	Introduction	824
2	Main result	825
3	Experiments and comparison with conjugate priors for simulated datasets	828
3.1	A Gibbs algorithm for \widehat{M}_λ	828

3.2 Experiments	829
4 Conclusion	830
Acknowledgements	830
Appendix: Proof of Theorem 1	831
References	839

1. Introduction

The “Netflix Prize” [5] generated a significant interest in the *matrix completion* problem. The Netflix data can be represented as a sparse matrix made up of ratings given by users (rows) to movies (columns). To infer the missing entries is thus very helpful to propose sensible advertisement and improve the sales. However, it is totally impossible to recover an uncomplete matrix without any assumption. A suitable condition, popular in practice for this problem, is that the matrix has low-rank or approximately low-rank [1, 3, 7, 8, 9, 15, 16]. For the Netflix problem, this assumption is sensible as it means that many movies (or users) have similar profiles.

Let $M_{m \times p}^0$ be an unknown matrix (expected to be low-rank) and $(X_1, Y_1), \dots, (X_n, Y_n)$ be i.i.d random variables drawn from a joint distribution \mathbf{P} . We assume that

$$Y_i = M_{X_i}^0 + \mathcal{E}_i, \quad i = 1, \dots, n, \quad (1)$$

the noise variables \mathcal{E}_i are independent from X_i and $\mathbb{E}(\mathcal{E}_i) = 0$. We let Π denote the marginal distribution of X when $(X, Y) \sim \mathbf{P}$. Remark that Π is a distribution on the set $\mathfrak{X} = \{1, \dots, m\} \times \{1, \dots, p\}$. Then, the problem of estimating M^0 with $n < mp$ is called the noisy matrix completion problem under general sampling distribution.

A special instance of this problem is that the sampling distribution Π is uniform, this assumption is done for example in [3, 7, 8, 9, 16]. Clearly, in practice, the observed entries are not always uniformly distributed: for example, some movies are more famous than others, and thus receive much more ratings. More importantly, the sampling distribution is not known in practice. More general sampling schemes than uniform distribution had been already studied, see e.g. [14, 15, 22], but there are still some assumptions on Π in these papers. Here, we do not impose any restriction on Π . From now, $\Pi_{ij} = \mathbb{P}(X = \{i, j\})$ will denote the probability to observe the (i, j) -th entry.

For any matrix $A_{m \times p}$, let $\|A\|_F$ denote the Frobenius norm, i.e, $\|A\|_F^2 = \text{Tr}(A^T A)$. We define a “generalized Frobenius norm” as follows

$$\|A\|_{F, \Pi}^2 = \sum_{ij} (A_{ij})^2 \Pi_{ij}.$$

Note that when the sampling distribution Π is uniform, then $\|A\|_{F, \Pi}^2 = \|A\|_F^2 / mp$. For any matrix $M_{m \times p} \in \mathbb{R}^{mp}$, we define the empirical risk as

$$r(M) = \frac{1}{n} \sum_{i=1}^n (Y_i - M_{X_i})^2$$

and the prediction risk

$$R(M) = \mathbb{E}_{(X,Y) \sim \mathbf{P}} \left[(Y - M_X)^2 \right].$$

In this paper, the prediction problem is considered, i.e, the objective is to define an estimator \widehat{M} such that $R(\widehat{M}) - R(M^0)$ is as small as possible. Remark that $R(M) - R(M^0) = \|M - M^0\|_{F,\Pi}^2$ for any M (using Pythagorean Theorem).

When handling with this problem, most of the recent methods are often based on minimizing a criterion of the fit to the observations, such as $r(M)$, penalized by the nuclear-norm or the rank of the matrix. A first result can be found in by Candès and Recht [8], Candès and Tao [9] for exact matrix completion (noiseless case, i.e. $\mathcal{E}_i = 0$). These results were then developed in the noisy case [7, 16]. Some efficient algorithms had also been proposed, for example see [23].

Recently, some authors have studied a more general problem, the so-called *Trace regression* problem: [15, 16]. This problem includes matrix completion, together with other well-known problems (linear regression, reduced rank regression and multitask learning) as special cases. They proposed nuclear-norm penalized estimators and provided reconstruction errors for their methods. They also proved that these errors are minimax-optimal (up to a logarithmic factor). Note that the average quadratic error on the entries of a rank- r matrix size $m \times p$ from n -observations can not be better than: $r \max(m, p)/n$ [16].

On the other hand, Bayesian methods have been also considered [3, 18, 19, 24, 28]. Most Bayesian estimators are based on conjugate priors which allow to use Gibbs sampling [3, 24] or Variational Bayes methods [19]. These priors are discussed in details in [3]. These algorithms are fast enough to deal with large datasets like Netflix or MovieLens¹, and are actually tested on these datasets in those papers. However, the theoretical understanding of Bayesian algorithms is not satisfying. Up to our knowledge, the minimax-optimality - and even the consistency - of the Bayesian estimator under conjugate prior is an open question.

In this paper, we design a new prior and prove an minimax-optimal oracle bound for the corresponding Bayesian estimator. This is presented in Section 2. In Section 3, we discuss the implementation of our Bayesian estimator. Some experiments comparing our estimator to the one based on conjugate priors are done on simulated datasets. The proof of the main result is provided in the appendix.

2. Main result

Before we introduce our estimator, let us formulate some assumptions.

Assumption 1. *There is a known constant L such that*

$$\|M^0\|_\infty = \sup_{i,j} |M_{ij}^0| \leq L < +\infty.$$

¹<http://grouplens.org/datasets/movielens/>

This is a mild assumption. In the Netflix and MovieLens datasets, the ratings belong to the set $\{1, 2, 3, 4, 5\}$, so we can take $L = 5$.

Assumption 2. *The noise variables $\mathcal{E}_1, \dots, \mathcal{E}_n$ are independent and independent of X_1, \dots, X_n . There exist two known constants $\sigma > 0$ and $\xi > 0$ such that*

$$\begin{aligned} \mathbb{E}(\mathcal{E}_i^2) &\leq \sigma^2 \\ \forall k \geq 3, \quad \mathbb{E}(|\mathcal{E}_i|^k) &\leq \sigma^2 k! \xi^{k-2}. \end{aligned}$$

Assumption 2 states that the noise is sub-exponential, it includes the cases where the noise is bounded or sub-Gaussian (and of course Gaussian), see e.g. Chapter 2 in [6].

We now describe a prior π on matrices $M_{m \times p}$ as follows. Let $K = \min(m, p)$ and Γ be a random variables taking value in the set $\{\Gamma_1, \dots, \Gamma_K\}$ with $\mathbb{P}(\Gamma = \Gamma_k) = \tau^{k-1} \left(\frac{1-\tau}{1-\tau^K} \right)$ where $\Gamma_k = (\overbrace{1, \dots, 1}^{k \text{ times}}, \overbrace{0, \dots, 0}^{K-k \text{ times}})$ for some constant $\tau \in (0, 1)$ and $k \in \{1, \dots, K\}$. Now, assuming that $\Gamma = \Gamma_k$ and a matrix $M_{m \times p}$ is drawn as $M = U_{m \times K} (V_{p \times K})^T$ where

$$U_{i,\ell}; V_{j,\ell} \stackrel{\text{i.i.d.}}{\sim} \begin{cases} \mathcal{U}([-\delta, \delta]) & \text{when } \Gamma_{k,\ell} = 1, \\ \mathcal{U}([-\kappa, \kappa]) & \text{when } \Gamma_{k,\ell} = 0, \end{cases} \quad \ell = 1, \dots, K$$

with $\delta = \sqrt{2L/K}$ and $0 \leq \kappa \leq (1/n)\sqrt{L/(10K)}$. Note that, in this case, the entries of M satisfy: $\sup_{i,j} |M_{ij}| \leq 2L$. Moreover, when a matrix M is drawn from this prior, as κ is small, most columns of U and V are almost null. So the matrix $M = UV^T$ is very close to a rank- k matrix. Actually, the choice $\kappa = 0$ leads to $\text{rank}(M) \leq k$.

We are now ready to define our estimator. For any $\lambda > 0$, we consider the conditional probability measure $\hat{\rho}_\lambda$ given by its density w.r.t. the probability measure π :

$$\frac{d\hat{\rho}_\lambda}{d\pi}(M) = \frac{e^{-\lambda r(M)}}{\int e^{-\lambda r} d\pi}. \quad (2)$$

The aggregate \widehat{M}_λ is defined as follows

$$\widehat{M}_\lambda = \int M \hat{\rho}_\lambda(dM). \quad (3)$$

Note that, for $\lambda = n/(2\sigma^2)$, this corresponds exactly to the Bayesian estimator that would be obtained for a Gaussian noise $\mathcal{E}_i \sim \mathcal{N}(0, \sigma^2)$. However, a slightly different choice for λ , denoted by λ^* below, will allow to obtain the optimality of the estimator under a wider class of noises. For any $x > 0$, define

$$\mathcal{M}(x) = \left\{ M = UV^T, \text{ with } |U_{i\ell}| \leq \sqrt{\frac{x}{K}}, |V_{j\ell}| \leq \sqrt{\frac{x}{K}} \right\}.$$

and $\mathcal{C} = [12L(2\xi + 3L)] \vee [8\sigma^2 + 2(3L)^2]$. Hereafter, the main result is presented. We provide an oracle bound for our estimator \widehat{M}_{λ^*} .

Theorem 1. *Let Assumption 1 and 2 be satisfied and take $\lambda = \lambda^* := \frac{n}{2\mathcal{C}}$. Then, for any $\epsilon \in (0, 1)$, with probability at least $1 - \epsilon$ and as soon as $n \geq \max(m, p)$, one has*

$$\begin{aligned} \|\widehat{M}_{\lambda^*} - M^0\|_{F, \Pi}^2 &\leq \inf_{M \in \mathcal{M}(L)} \left\{ 3\|M - M^0\|_{F, \Pi}^2 \right. \\ &\quad \left. + \mathcal{C}_{L, \xi, \sigma, \tau} \frac{(m+p)\text{rank}(M)\log(K)}{n} + \frac{8\mathcal{C}\log\left(\frac{2}{\epsilon}\right)}{n} \right\}, \end{aligned}$$

where $\mathcal{C}_{L, \xi, \sigma, \tau}$ is a (known) numerical constant depending on L, ξ, σ and τ only.

The proof of this theorem is given in the [appendix](#). It follows an argument called ‘‘PAC-Bayesian inequality’’. PAC-Bayesian inequalities were introduced in [25, 21] in order to provide empirical bounds on the prevision risk of Bayesian-type estimators. However, our proof is closer to Catoni’s works [10, 11, 12], where it is shown how to derive powerful oracle inequalities from PAC-Bayesian bounds. This approach has been used many times since then to prove oracle inequalities in many dimension-reduction problems like sparse regression estimation [13, 4, 2] or reduced-rank regression [1].

The choice $\lambda = \lambda^*$ comes from the proof of this theorem when optimizing an upper bound on the risk R , see (15) page 838. However, in practice, this choice may not be the best one. For example, in the experiments done in Section 3 with Gaussian noise $\mathcal{E}_i \sim \mathcal{N}(0, \sigma^2)$, we take $\lambda = \frac{n}{4\sigma^2}$ that was shown in [13] to behave very well in regression problems. Also, in practice, to take K smaller than $\min(m, p)$ improves significantly the speed of the algorithm with little consequence on the performance of the estimator [3].

Remark 1. When $M^0 \in \mathcal{M}(L)$, we can take $M = M^0$, one gets

$$\|\widehat{M}_{\lambda^*} - M^0\|_{F, \Pi}^2 \leq \mathcal{C}_{L, \xi, \sigma, \tau} \frac{(m+p)\text{rank}(M^0)\log(K)}{n} + \frac{8\mathcal{C}\log\left(\frac{2}{\epsilon}\right)}{n}.$$

The rate $(m+p)\text{rank}(M^0)\log(K)/n$ is minimax-optimal, or at least almost minimax-optimal: a lower bound in this problem is provided by Theorems 5 and 7 in [16], it is $(m+p)\text{rank}(M^0)/n$. The optimality of the log term is, to our knowledge, an open question. Note however that the upper bound in [16] is $(m+p)\text{rank}(M^0)\log(m+p)/n$. So, our bound represents a slight improvement in the case $\min(m, p) \ll \max(m, p)$.

Remark 2. When the sampling distribution Π is uniform in Theorem 1, we obtain the following oracle bound for the Frobenius norm

$$\begin{aligned} \frac{1}{mp} \|\widehat{M}_{\lambda^*} - M^0\|_F^2 &\leq \inf_{M \in \mathcal{M}(L)} \left\{ \frac{3}{mp} \|M - M^0\|_F^2 \right. \\ &\quad \left. + \mathcal{C}'_{L, \xi, \sigma, \tau} \frac{(m+p)\text{rank}(M)\log(K)}{n} + \frac{8\mathcal{C}\log\left(\frac{2}{\epsilon}\right)}{n} \right\}. \end{aligned}$$

Finally, we want to mention that the rate of [16] is also reached, in a work parallel to ours, by Suzuki [26], in a Bayesian framework. The main difference is that, while [26] provides a rate of convergence in a more general low-rank tensor estimation problem, his works do not bring an oracle inequality like Theorem 1 that can be used when M^0 is not exactly low-rank, but can be well approximated by a low-rank matrix. Moreover, our result holds under any sampling distribution Π .

3. Experiments and comparison with conjugate priors for simulated datasets

3.1. A Gibbs algorithm for \widehat{M}_λ

As it has been shown in Section 2, our estimator \widehat{M}_λ^* satisfies a powerful oracle inequality. However, as mentioned in the introduction, the Bayesian estimator using conjugate priors is popular in practice as it leads to a fast algorithm. The reason is that there is an explicit form for the conditional posterior distribution of the i -th row of U , $U_{i,\cdot}$, given the other rows of U , $U_{-i,\cdot}$, and given V (it is a multivariate normal distribution which parameters are known). This allows to use a Gibbs sampler, with very good convergence properties. This is described for example in [3] and the references therein.

Here, straightforward but tedious computations lead to

$$\begin{aligned} \hat{\rho}_\lambda(U_{i,\cdot}|k, U_{-i,\ell}, V, \Gamma = \Gamma_k) \\ \propto \varphi \left[U_{i,\cdot}; \frac{2\lambda}{n} \Sigma_i \sum_{k:I_k=i} Y_k V_{J_k,\cdot}, \Sigma_i \right] \prod_{\ell=1}^k \mathbf{1}_{\{|U_{i,\ell}| \leq \delta\}} \prod_{\ell=k+1}^K \mathbf{1}_{\{|U_{i,\ell}| \leq \kappa\}} \end{aligned}$$

where we use the notation $X_1 = (I_1, J_1), \dots, X_n = (I_n, J_n)$,

$$(\Sigma_i)^{-1} = \frac{2\lambda}{n} \sum_{k:I_k=i} V_{J_k,\cdot}^T V_{J_k,\cdot}$$

and $\varphi(\cdot; m, V)$ is the density of the multivariate normal distribution with mean vector m and variance-covariance matrix V . So, the conditional posterior distribution of $U_{i,\cdot}$ is a truncated multivariate normal. To sample from such a distribution is known as a very hard problem in general, see for example [17]. However, using the R package `tmvtnorm` [27], it is possible to sample from a truncated multivariate normal fast enough to compute our estimator on reasonably large datasets. Finally, instead of including the hyperparameter $k \in \{1, \dots, K\}$ in the simulations, we simulated K chains simultaneously, one for every $k \in \{1, \dots, K\}$, and selected the realization of one of the chains at each round using the probabilities given by (2).

Also, note that the truncation procedure proposed by Suzuki in [26] cannot be implemented, to our understanding, using this procedure, as the truncation is done directly on the product UV^T rather than on U and V individually.

3.2. Experiments

We use the notation \widehat{M}_λ for our estimator, let us denote $\widehat{M}^{\text{conjugate}}$ the estimator based on the Gaussian prior for U and V with inverse Gamma variance, described in [3] and in the aforementioned references. In order to compare both estimators, a series of experiments were done with simulated data:

- In the first series of simulations, the data are simulated as in [7, 3]. More precisely, a rank-2 matrix $M_{m \times m}^0$ (so $m = p$) has been created as the product of two rank-2 matrices, $M^0 = U_{m \times 2}^0 (V_{m \times 2}^0)^T$, where the entries of U^0 and V^0 are i.i.d $\mathcal{N}(0, 20/\sqrt{m})$. Only 20% entries of the matrix M^0 are observed (using a uniform sampling). This sampled set is then corrupted by noise as in (1), where the \mathcal{E}_i are i.i.d $\mathcal{N}(0, 1)$. We consider the cases $m = 100$, $m = 200$, $m = 500$ and $m = 1000$.
- The second series of simulations is similar to the first one, except that the matrix M^0 is no longer rank 2, but it can be well approximated by a rank 2 matrix:

$$M^0 = U_{m \times 2}^0 (V_{m \times 2}^0)^T + \frac{1}{100} (Z_{m \times 50}^0) (W_{m \times 50}^0)^T$$

where the entries of Z^0 and W^0 are i.i.d $\mathcal{N}(0, 20/\sqrt{m})$.

- The third series of experiments is similar to the first one, but the noise variables \mathcal{E}_i are now i.i.d from a uniform distribution on $[-1, 1]$. Note that, from a purely Bayesian point of view, this corresponds to a misspecified model. However, the bound in Theorem 1 is still valid in this case.
- Finally, the fourth series of experiments is similar to the first one, noise variables \mathcal{E}_i are now i.i.d from a heavy-tailed distribution (Student, with parameter 5). This is another misspecified model, but in this case, Theorem 1 cannot be used.

The behavior of our estimator \widehat{M}_λ is computed through the root-mean-squared error (RMSE) per entry,

$$\text{RMSE} = [(1/mp) \|\widehat{M}_\lambda - M^0\|_F^2]^{1/2} = (1/m) \|\widehat{M}_\lambda - M^0\|_F.$$

The parameters are given as follows: for both \widehat{M}_λ and $\widehat{M}^{\text{conjugate}}$, the parameter λ is set to $n/4$, following [13]. Following [3] we use for the parameters of the inverse Gamma prior in $\widehat{M}^{\text{conjugate}}$ the values $a = 1$, $b = 1/100$. Finally, for \widehat{M}_λ , we used $\kappa = 0$, $K = 5$, $L = 50$ and $\tau = 1/2$ on all the simulations apart from the heavy-tailed noise case, where we used $\tau = 1/4$. Note that a proper optimization with respect to the parameters τ and λ could lead to better results, for example through cross-validation.

The first conclusion is that the results of both methods are very close. In many situations, however, the variance of the estimator with uniform prior is larger than the variance of the estimator with Gaussian prior. The evidence is that this is due to the fact that the MCMC algorithm used to compute the estimator with Gaussian prior, $\widehat{M}^{\text{conjugate}}$, converges faster than the algorithm used to

TABLE 1
RMSEs in the first series of experiments (low-rank matrix, Gaussian noise)

prior	$m = 100$	$m = 200$	$m = 500$	$m = 1000$
Uniform	0.535 (± 0.003)	0.348 (± 0.003)	0.207 (± 0.0001)	0.141 (± 0.0006)
Gaussian	0.538 (± 0.001)	0.345 (± 0.001)	0.210 (± 0.0001)	0.146 (± 0.001)

TABLE 2
RMSEs in the second series of experiments (approx. low-rank, Gaussian noise)

prior	$m = 100$	$m = 200$	$m = 500$	$m = 1000$
Uniform	0.640 (± 0.008)	0.387 (± 0.001)	0.214 (± 0.0008)	0.145 (± 0.0002)
Gaussian	0.620 (± 0.003)	0.385 (± 0.001)	0.216 (± 0.0003)	0.145 (± 0.001)

TABLE 3
RMSEs in the third series of experiments (low-rank matrix, uniform noise)

prior	$m = 100$	$m = 200$	$m = 500$	$m = 1000$
Uniform	0.328 (± 0.002)	0.205 (± 0.001)	0.120 (± 0.001)	0.084 (± 0.002)
Gaussian	0.334 (± 0.003)	0.208 (± 0.001)	0.126 (± 0.003)	0.086 (± 0.001)

TABLE 4
RMSEs in the fourth series of experiments (low-rank matrix, heavy-tailed noise)

prior	$m = 100$	$m = 200$	$m = 500$	$m = 1000$
Uniform	0.745 (± 0.039)	0.567 (± 0.005)	0.340 (± 0.004)	0.237 (± 0.003)
Gaussian	0.659 (± 0.003)	0.439 (± 0.001)	0.268 (± 0.002)	0.186 (± 0.002)

compute the estimator with uniform prior, \widehat{M}_λ . This is supported by Figure 1 page 831. However, it seems that this difference is less and less significant when the dimension m grows.

According to our main oracle inequality, our estimator is robust to misspecification in the low-rank assumption, see Table 2, and in the noise, at least in the sub-Gaussian case, see Table 3. More importantly: despite the fact that the theoretical properties of $\widehat{M}^{\text{conjugate}}$ are not known, this estimator is more robust than ours to heavy-tailed noise, as shown in Table 4.

4. Conclusion

This paper proposes a Bayesian estimator for the noisy matrix completion problem under general sampling distribution. This estimator satisfies an optimal oracle inequality under any sampling scheme. Based on simulations, it is also clear that this estimator performs well in practice, however, a faster algorithm for very large datasets is still an open issue. Another important open question is the minimax-optimality of the estimator based on Gaussian priors.

Acknowledgements

We would like to thank the anonymous referees for their constructive comments and Professor Taiji Suzuki for enlightening discussions.

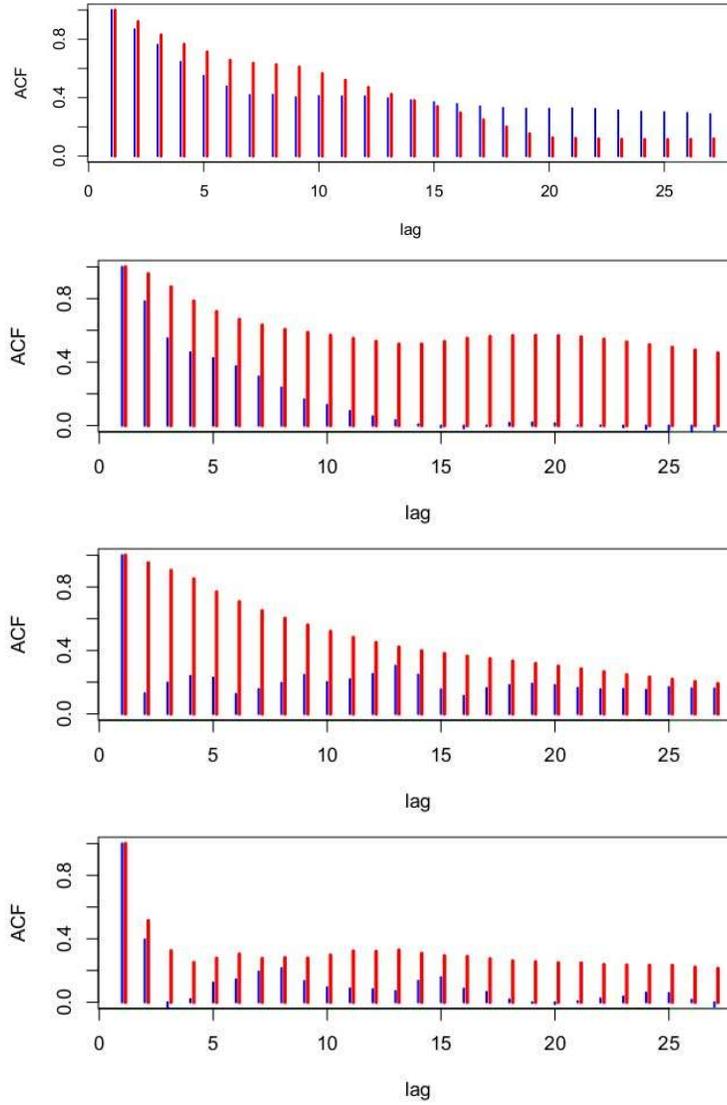


FIG 1. ACF of four randomly selected entries during a simulation. These are taken from the first series of experiments. The ACF of the Gibbs sampler for the Bayesian estimator with uniform priors, \widehat{M}_λ , is in red while the ACF of the Gibbs sampler for the Bayesian estimator with Gaussian priors, $\widehat{M}^{\text{conjugate}}$, is in blue.

Appendix: Proof of Theorem 1

First, we state a version of Bernstein’s inequality useful in the proof of Theorem 1. This version is taken from [20] (Inequality 2.21 in the proof of Proposition 2.9 page 24).

Lemma 2. Let T_1, \dots, T_n be independent real valued random variables. Let us assume that there are two constants v and w such that

$$\sum_{i=1}^n \mathbb{E}[T_i^2] \leq v$$

and for all integers $k \geq 3$,

$$\sum_{i=1}^n \mathbb{E}[(T_i)^k] \leq v \frac{k!w^{k-2}}{2}.$$

Then, for any $\zeta \in (0, 1/w)$,

$$\mathbb{E} \exp \left[\zeta \sum_{i=1}^n [T_i - \mathbb{E}(T_i)] \right] \leq \exp \left(\frac{v\zeta^2}{2(1-w\zeta)} \right).$$

Now, we are ready to present the proof of Theorem 1.

Proof of Theorem 1. the proof is divided in two steps. In the first step, we establish a general PAC-Bayesian inequality for matrix completion, in the style of [11, 13]. In the second step, we derive the oracle inequality from the first step.

Step 1:

Let's define, for any matrix $M \in \mathcal{M}(2L)$, the following random variables

$$T_i = (Y_i - M_{X_i}^0)^2 - (Y_i - M_{X_i})^2.$$

Note that these variables are independent. We first check that the variables T_i satisfy the assumptions of Lemma 2, in order to apply this lemma. We have

$$\begin{aligned} \sum_{i=1}^n \mathbb{E}[T_i^2] &= \sum_{i=1}^n \mathbb{E} \left[(2Y_i - M_{X_i}^0 - M_{X_i})^2 (M_{X_i}^0 - M_{X_i})^2 \right] \\ &= \sum_{i=1}^n \mathbb{E} \left[(2\mathcal{E}_i + M_{X_i}^0 - M_{X_i})^2 (M_{X_i}^0 - M_{X_i})^2 \right] \\ &\leq \sum_{i=1}^n \mathbb{E} \left[[8\mathcal{E}_i^2 + 2(L+2L)^2] [M_{X_i}^0 - M_{X_i}]^2 \right] \\ &= \sum_{i=1}^n \mathbb{E} [8\mathcal{E}_i^2 + 2(3L)^2] \mathbb{E} [M_{X_i}^0 - M_{X_i}]^2 \\ &\leq n [8\sigma^2 + 2(3L)^2] [R(M) - R(M^0)] =: v(M, M^0) = v. \end{aligned}$$

Next we have, for any integer $k \geq 3$, that

$$\sum_{i=1}^n \mathbb{E} [(T_i)^k] \leq \sum_{i=1}^n \mathbb{E} \left[|2Y_i - M_{X_i}^0 - M_{X_i}|^k |M_{X_i}^0 - M_{X_i}|^k \right]$$

$$\begin{aligned}
 &\leq \sum_{i=1}^n \mathbb{E} \left[2^{2k-1} [|\mathcal{E}_i|^k + (L/2 + L)^k] |M_{X_i}^0 - M_{X_i}|^k \right] \\
 &\leq \sum_{i=1}^n \mathbb{E} \left[2^{2k-1} \left(|\mathcal{E}_i|^k + \left(\frac{3}{2}L\right)^k \right) (3L)^{k-2} |M_{X_i}^0 - M_{X_i}|^2 \right] \\
 &\leq 2^{2k-1} \left[\sigma^2 k! \xi^{k-2} + \left(\frac{3}{2}L\right)^k \right] (3L)^{k-2} \sum_{i=1}^n \mathbb{E} |M_{X_i}^0 - M_{X_i}|^2 \\
 &\leq \frac{[\sigma^2 k! \xi^{k-2} + (\frac{3}{2}L)^k] [4(3L)]^{k-2}}{\sigma^2 + (\frac{3}{2}L)^2} v \\
 &\leq \left[k! \xi^{k-2} + \left(\frac{3}{2}L\right)^{k-2} \right] [4(3L)]^{k-2} v \\
 &\leq k! \left(\xi + \frac{3}{2}L \right)^{k-2} (12L)^{k-2} v \leq v \frac{k! w^{k-2}}{2},
 \end{aligned}$$

with $w := 12L(2\xi + 3L)$.

Next, for any $\lambda \in (0, n/w)$, applying Lemma 2 with $\zeta = \lambda/n$ gives

$$\mathbb{E} \exp \left[\lambda \left(R(M) - R(M^0) - r(M) + r(M^0) \right) \right] \leq \exp \left[\frac{v\lambda^2}{2n^2(1 - \frac{w\lambda}{n})} \right].$$

Set $\mathcal{C}_{\sigma,L} = 2 [4\sigma^2 + (3L)^2]$. For the sake of simplicity let us put

$$\alpha = \left(\lambda - \frac{\lambda^2 \mathcal{C}_{\sigma,L}}{2n(1 - \frac{w\lambda}{n})} \right). \tag{4}$$

In order to understand what follows, keep in mind that w is a constant and that our optimal estimator comes with $\lambda = \lambda^* = \frac{n}{2\mathcal{C}}$, so α is of order n .

For any $\varepsilon > 0$, the last display yields

$$\mathbb{E} \exp \left[\alpha \left(R(M) - R(M^0) \right) + \lambda \left(-r(M) + r(M^0) \right) - \log \frac{2}{\varepsilon} \right] \leq \frac{\varepsilon}{2}.$$

Integrating w.r.t. the probability distribution $\pi(\cdot)$, we get

$$\int \mathbb{E} \exp \left[\alpha \left(R(M) - R(M^0) \right) + \lambda \left(-r(M) + r(M^0) \right) - \log \frac{2}{\varepsilon} \right] \pi(dM) \leq \frac{\varepsilon}{2}.$$

Next, Fubini's theorem gives

$$\begin{aligned}
 &\mathbb{E} \int \exp \left[\alpha \left(R(M) - R(M^0) \right) + \lambda \left(-r(M) + r(M^0) \right) - \log \frac{2}{\varepsilon} \right] \pi(dM) \\
 &= \mathbb{E} \int \exp \left\{ \alpha \left(R(M) - R(M^0) \right) + \lambda \left(-r(M) + r(M^0) \right) \right. \\
 &\quad \left. - \log \left[\frac{d\hat{\rho}_\lambda}{d\pi}(M) \right] - \log \frac{2}{\varepsilon} \right\} \hat{\rho}_\lambda(dM) \leq \frac{\varepsilon}{2}.
 \end{aligned}$$

Jensen's inequality yields

$$\mathbb{E} \exp \left[\alpha \left(\int R d\hat{\rho}_\lambda - R(M^0) \right) + \lambda \left(- \int r d\hat{\rho}_\lambda + r(M^0) \right) - \mathcal{K}(\hat{\rho}_\lambda, \pi) - \log \frac{2}{\varepsilon} \right] \leq \frac{\varepsilon}{2},$$

where $\mathcal{K}(p, q)$ is the Kullback–Leibler divergence of p from q . Now, using the basic inequality $\exp(x) \geq \mathbf{1}_{\mathbb{R}_+}(x)$, we get

$$\mathbb{P} \left\{ \left[\alpha \left(\int R d\hat{\rho}_\lambda - R(M^0) \right) + \lambda \left(- \int r d\hat{\rho}_\lambda + r(M^0) \right) - \mathcal{K}(\hat{\rho}_\lambda, \pi) - \log \frac{2}{\varepsilon} \right] \geq 0 \right\} \leq \frac{\varepsilon}{2}.$$

Using Jensen's inequality again gives

$$\int R d\hat{\rho}_\lambda \geq R \left(\int M \hat{\rho}_\lambda(dM) \right) = R(\widehat{M}_\lambda).$$

Combining the last two displays we obtain

$$\mathbb{P} \left\{ R(\widehat{M}_\lambda) - R(M^0) \leq \frac{\int r d\hat{\rho}_\lambda - r(M^0) + \frac{1}{\lambda} [\mathcal{K}(\hat{\rho}_\lambda, \pi) + \log \frac{2}{\varepsilon}]}{\frac{\alpha}{\lambda}} \right\} \geq 1 - \frac{\varepsilon}{2}.$$

Using Donsker and Varadhan's variational inequality (Lemma 1.1.3 in Catoni [12]), we get

$$\mathbb{P} \left\{ R(\widehat{M}_\lambda) - R(M^0) \leq \inf_{\rho \in \mathfrak{M}_+^1(M)} \frac{\int r d\rho - r(M^0) + \frac{1}{\lambda} [\mathcal{K}(\rho, \pi) + \log \frac{2}{\varepsilon}]}{\frac{\alpha}{\lambda}} \right\} \geq 1 - \frac{\varepsilon}{2}, \quad (5)$$

where $\mathfrak{M}_+^1(M)$ is the set of all positive probability measures over the set of $m \times p$ matrices equipped with the Borel σ -algebra.

We now want to bound from above $r(M) - r(M^0)$ by $R(M) - R(M^0)$. We can use Lemma 2 again, to $\tilde{T}_i(\theta) = -T_i(\theta)$ and similar computations yield successively

$$\mathbb{E} \exp \left[\lambda \left(R(M^0) - R(M) + r(M) - r(M^0) \right) \right] \leq \exp \left[\frac{v\lambda^2}{2n^2(1 - \frac{w\lambda}{n})} \right],$$

and so for any (data-dependent) ρ ,

$$\mathbb{E} \exp \left[\beta \left(- \int R d\rho + R(M^0) \right) + \lambda \left(\int r d\rho - r(M^0) \right) - \mathcal{K}(\rho, \pi) - \log \frac{2}{\varepsilon} \right] \leq \frac{\varepsilon}{2},$$

where

$$\beta = \left(\lambda + \frac{\lambda^2 \mathcal{C}_{\sigma, L}}{2n(1 - \frac{w\lambda}{n})} \right). \quad (6)$$

Here again, with the same spirit with α in (4), β is of order n also. So:

$$\mathbb{P}\left\{\int rd\rho - r(M^0) \leq \frac{\beta}{\lambda} \left[\int Rd\rho - R(M^0)\right] + \frac{1}{\lambda} \left[\mathcal{K}(\rho, \pi) + \log \frac{2}{\varepsilon}\right]\right\} \geq 1 - \frac{\varepsilon}{2}. \tag{7}$$

Combining (7) and (5) with a union bound argument gives the general PAC-Bayesian bound

$$\mathbb{P}\left\{R(\widehat{M}_\lambda) - R(M^0) \leq \inf_{\rho \in \mathfrak{M}_+^1(M)} \frac{\beta[\int Rd\rho - R(M^0)] + 2[\mathcal{K}(\rho, \pi) + \log \frac{2}{\varepsilon}]}{\alpha}\right\} \geq 1 - \varepsilon. \tag{8}$$

Step 2:

In the second step, we derive an explicit form for the upper bound in (8). The idea is that, if we restrict the infimum in the upper bound in (8) to a small set of measures ρ , we are able to provide an explicit bound for this infimum. This trick was introduced in [11].

Let $M \in \mathcal{M}(L)$, it means that $M = UV^T$ with $|U_{i\ell}| \leq \sqrt{L/K}, |V_{j\ell}| \leq \sqrt{L/K}$. Let us take, for any c such that $\kappa \leq c < (\sqrt{2}-1)\sqrt{L/K}$, the probability distribution

$$\rho_{U,V,c}(d\mu, d\nu) \propto \mathbf{1}(\|\mu - U\|_\infty \leq c, \|\nu - V\|_\infty \leq c) \pi(d\mu, d\nu).$$

Note that, as $c < (\sqrt{2}-1)\sqrt{L/K}$, we have $\text{supp}(\rho_{U,V,c}) \subset \text{supp}(\pi)$ and so $\mathcal{K}(\rho_{U,V,c}, \pi) < \infty$.

Thus, (8) becomes

$$\mathbb{P}\left\{R(\widehat{M}_\lambda) - R(M^0) \leq \inf_{U,V,c} \frac{\beta[\int Rd\rho_{U,V,c} - R(M^0)] + 2[\mathcal{K}(\rho_{U,V,c}, \pi) + \log \frac{2}{\varepsilon}]}{\alpha}\right\} \geq 1 - \varepsilon. \tag{9}$$

Let us fix c, U, V . The end the proof consists in calculations to derive an upper bound for the two terms in (9). Firstly

$$\begin{aligned} & \int Rd\rho_{U,V,c} - R(M^0) \\ &= \int \|\mu\nu^T - M^0\|_{F,\Pi}^2 \rho_{U,V,c}(d\mu, d\nu) \\ &= \int \|\mu\nu^T - U\nu^T + U\nu^T - UV^T + UV^T - M^0\|_{F,\Pi}^2 \rho_{U,V,c}(d\mu, d\nu) \\ &= \int \left(\|\mu\nu^T - U\nu^T\|_{F,\Pi}^2 + \|U\nu^T - UV^T\|_{F,\Pi}^2 \right) \end{aligned}$$

$$\begin{aligned}
& + \|UV^T - M^0\|_{F,\Pi}^2 + 2 \langle \mu\nu^T - U\nu^T, U\nu^T - UV^T \rangle_{F,\Pi} \\
& + 2 \langle \mu\nu^T - U\nu^T, UV^T - M^0 \rangle_{F,\Pi} \\
& + 2 \langle U\nu^T - UV^T, UV^T - M^0 \rangle_{F,\Pi} \Big) \rho_{U,V,c}(d\mu, d\nu).
\end{aligned}$$

(note that we use the notation $\langle A, B \rangle_{F,\Pi} = \sum_{i,j} A_{ij} B_{ij} \Pi_{ij}$). As $\int \mu \rho_{U,V,c}(d\mu) = U$ and $\int \nu \rho_{U,V,c}(d\nu) = V$, it can be seen that integral of the three scalar products in the previous equation vanish. Moreover,

$$\begin{aligned}
\|(\mu - U)\nu^T\|_{F,\Pi}^2 &= \sum_{ij} [(\mu - U)\nu^T]_{ij}^2 \Pi_{ij} \leq \left(\sup_{ij} [(\mu - U)\nu^T]_{ij} \right)^2 \sum_{ij} \Pi_{ij} \\
&\leq \left(\sup_{ij} \sum_{\ell=1}^K |\mu - U|_{i\ell} |\nu|_{j\ell} \right)^2 \leq \left(K \sup_{i\ell} |\mu - U|_{i\ell} \sup_{j\ell} |\nu|_{j\ell} \right)^2 \\
&\leq \left[Kc \left(c + \sqrt{\frac{L}{K}} \right) \right]^2 = Kc^2 (\sqrt{K}c + \sqrt{L})^2,
\end{aligned}$$

similarly $\|U\nu^T - UV^T\|_{F,\Pi}^2 \leq Kc^2$. Therefore, from (9), we have

$$\begin{aligned}
& \int \| \mu\nu^T - M^0 \|_{F,\Pi}^2 \rho_{U,V,c}(d\mu, d\nu) \\
& \leq Kc^2 \left[(\sqrt{K}c + \sqrt{L})^2 + L \right] + \|UV^T - M^0\|_{F,\Pi}^2. \quad (10)
\end{aligned}$$

So, we have an upper bound for the first term in (9). We now deal with the Kullback-Leibler term:

$$\begin{aligned}
\mathcal{K}(\rho_{U,V,c}, \pi) &= \log \frac{1}{\pi(\{\mu, \nu : \|\mu - U\|_\infty \leq c, \|\nu - V\|_\infty \leq c\})} \\
&= \log \frac{1}{\pi(\{\mu : \|\mu - U\|_\infty \leq c\})} + \log \frac{1}{\pi(\{\nu : \|\nu - V\|_\infty \leq c\})} \\
&= \log \frac{1}{\int \pi(\{\|\mu - U\|_\infty \leq c\} | \Gamma) \pi(\Gamma) d\Gamma} \\
& \quad + \log \frac{1}{\int \pi(\{\|\nu - V\|_\infty \leq c\} | \Gamma) \pi(\Gamma) d\Gamma}. \quad (11)
\end{aligned}$$

Note that, up to a reordering of the columns of U and V , we can assume that $U = (U_1 | \dots | U_{k_0} | 0 | \dots | 0)$ and $V = (V_1 | \dots | V_{k_0} | 0 | \dots | 0)$, where $k_0 = \text{rank}(UV^T) \leq K$. Then

$$\int \pi(\{\|\mu - U\|_\infty \leq c\} | \Gamma) \pi(\Gamma) d\Gamma = \tau^{k_0-1} \left(\frac{1-\tau}{1-\tau^K} \right) \pi(\{\|\mu - U\|_\infty \leq c\} | \Gamma = \Gamma_{k_0})$$

and, as $\kappa \leq c$,

$$\begin{aligned} & \pi(\{\|\mu - U\|_\infty \leq c\} | \Gamma = \Gamma_{k_0}) \\ & \geq \prod_{i=1}^m \prod_{\ell=1}^{k_0} \pi(\{|\mu_{i\ell} - U_{i\ell}| \leq c\} | \Gamma = \Gamma_{k_0}) \prod_{\ell=k_0+1}^K \pi(\{|\mu_{i\ell}| \leq c\} | \Gamma = \Gamma_{k_0}) \\ & \geq \left(c\sqrt{\frac{K}{2L}}\right)^{mk_0}. \end{aligned}$$

So,

$$\begin{aligned} & \log \frac{1}{\int \pi(\{\|\mu - U\|_\infty \leq c\} | \Gamma) \pi(\Gamma) d\Gamma} \\ & \leq (k_0 - 1) \log(1/\tau) + \log\left(\frac{1 - \tau^K}{1 - \tau}\right) + mk_0 \log\left(\frac{1}{c}\sqrt{\frac{2L}{K}}\right) \\ & \leq (k_0 - 1) \log(1/\tau) + \log\left(\frac{1}{1 - \tau}\right) + mk_0 \log\left(\frac{1}{c}\sqrt{\frac{2L}{K}}\right). \end{aligned} \quad (12)$$

By symmetry,

$$\begin{aligned} \log \frac{1}{\int \pi(\{\|\nu - V\|_\infty \leq c\} | \Gamma) \pi(\Gamma) d\Gamma} & \leq (k_0 - 1) \log(1/\tau) + \log\left(\frac{1}{1 - \tau}\right) \\ & \quad + pk_0 \log\left(\frac{1}{c}\sqrt{\frac{2L}{K}}\right). \end{aligned} \quad (13)$$

Plugging (12) and (13) into (11), we obtain finally our upper bound for the Kullback-Leibler term:

$$\begin{aligned} \mathcal{K}(\rho_{U,V,c}, \pi) & \leq 2(k_0 - 1) \log(1/\tau) + 2 \log\left(\frac{1}{1 - \tau}\right) + (m + p)k_0 \log\left(\frac{1}{c}\sqrt{\frac{2L}{K}}\right) \\ & \leq 2k_0 \log(1/\tau) + 2 \log\left(\frac{\tau}{1 - \tau}\right) + (m + p)k_0 \log\left(\frac{1}{c}\sqrt{\frac{2L}{K}}\right). \end{aligned} \quad (14)$$

Finally, substituting (10) and (14) into (9),

$$\begin{aligned} \mathbb{P}\left\{ R(\widehat{M}) - R(M^0) \leq \inf_{\substack{U, V, c \\ U_j, V_j = 0 \text{ when } j > k_0}} \frac{1}{\alpha} \left[\beta \left(Kc^2 [(\sqrt{K}c + \sqrt{L})^2 + L] \right. \right. \right. \\ \left. \left. \left. + \|UV^T - M^0\|_{F, \Pi}^2 \right) + 2(m + p)k_0 \log\left(\frac{1}{c}\sqrt{\frac{2L}{K}}\right) \right. \right. \\ \left. \left. \left. + 4k_0 \log(1/\tau) + 4 \log\left(\frac{\tau}{1 - \tau}\right) + 2 \log\frac{2}{\varepsilon} \right] \right\} \geq 1 - \varepsilon. \end{aligned}$$

Let us put $c = \sqrt{(m+p)L/(18nK)}$. Note that as $n \geq \max(m, p)$ then $\sqrt{(m+p)/(3n)} < 1$ and thus the condition $c < (\sqrt{2} - 1)\sqrt{L/K}$ is satisfied. So we have the following inequality with probability at least $1 - \varepsilon$:

$$\begin{aligned} & R(\widehat{M}_\lambda) - R(M^0) \\ & \leq \inf_{\substack{U, V \\ U_j, V_j = 0 \text{ when } j > k_0}} \frac{1}{1 - \frac{\lambda \mathcal{C}_{\sigma, L}}{2(n-w\lambda)}} \left\{ \left(1 + \frac{\lambda \mathcal{C}_{\sigma, L}}{2(n-w\lambda)} \right) \left[\|UV^T - M^0\|_{F, \Pi}^2 \right. \right. \\ & \quad \left. \left. + L \frac{m+p}{18n} \left(2L \frac{m+p}{18n} + 3L \right) \right] + \frac{2}{\lambda} \left[(m+p)k_0 \log \left(\sqrt{\frac{36n}{m+p}} \right) \right. \right. \\ & \quad \left. \left. + 2k_0 \log(1/\tau) + 2 \log \left(\frac{\tau}{1-\tau} \right) + \log \frac{2}{\varepsilon} \right] \right\}, \end{aligned}$$

where α and β have been replaced by their definitions, see (4) and (6). Taking now $\lambda = \lambda^* = n/(2\mathcal{C})$ with $\mathcal{C} = \mathcal{C}_{\sigma, L} \vee w$ in the last above display, gives

$$\begin{aligned} \mathbb{P} \left\{ R(\widehat{M}_{\lambda^*}) - R(M^0) \leq \inf_{M \in \mathcal{M}(L)} \left\{ 3 \left[L^2 \frac{m+p}{18n} \left(\frac{m+p}{9n} + 3 \right) + \|M - M^0\|_{F, \Pi}^2 \right] \right. \right. \\ \left. \left. + \frac{8\mathcal{C}}{n} \left[\frac{1}{2} (m+p) \text{rank}(M) \log \left(\frac{36n}{m+p} \right) + \log \frac{2}{\varepsilon} \right] \right. \right. \\ \left. \left. + 2 \text{rank}(M) \log(1/\tau) + 2 \log \left(\frac{\tau}{1-\tau} \right) \right] \right\} \geq 1 - \varepsilon, \end{aligned} \tag{15}$$

where we have used that $1 - \frac{\lambda \mathcal{C}_{\sigma, L}}{2(n-w\lambda)} \geq 1/2$ and $1 + \frac{\lambda \mathcal{C}_{\sigma, L}}{2(n-w\lambda)} \leq 3/2$. As

$$\log \left(\frac{36n}{m+p} \right) \leq \log \left(\frac{36mp}{\max(m, p)} \right) = \log \left(\frac{36 \min(m, p) \max(m, p)}{\max(m, p)} \right) = \log(36K),$$

we have

$$\begin{aligned} \mathbb{P} \left\{ R(\widehat{M}_{\lambda^*}) - R(M^0) \leq \inf_{M \in \mathcal{M}(L)} \left\{ 3 \left[L^2 \frac{m+p}{18n} \left(\frac{m+p}{9n} + 3 \right) + \|M - M^0\|_{F, \Pi}^2 \right] \right. \right. \\ \left. \left. + \frac{8\mathcal{C}}{n} \left[\frac{1}{2} (m+p) \text{rank}(M) \log(36K) + \log \frac{2}{\varepsilon} \right] \right. \right. \\ \left. \left. + 2 \text{rank}(M) \log(1/\tau) + 2 \log \left(\frac{1}{1-\tau} \right) \right] \right\} \geq 1 - \varepsilon. \end{aligned} \tag{16}$$

Moreover,

$$L^2 \frac{m+p}{6n} \left(\frac{m+p}{9n} + 3 \right) \leq \mathcal{C}(L) \frac{(m+p) \text{rank}(M) \log(K)}{n},$$

for some constant $\mathcal{C}(L) > 0$ depending on L only. Remind that τ is a constant in $(0, 1)$, we have

$$2\text{rank}(M) \log(1/\tau) + 2 \log \left(\frac{\tau}{1-\tau} \right) \leq \mathcal{C}(\tau) \frac{(m+p)\text{rank}(M) \log(K)}{n},$$

for some constant $\mathcal{C}(\tau) > 0$ depending on τ only. Finally, from (16), we obtain

$$\mathbb{P} \left\{ R(\widehat{M}_{\lambda^*}) - R(M^0) \leq \inf_{M \in \mathcal{M}(L)} \left[3\|M - M^0\|_{F, \Pi}^2 + \mathcal{C}(L, \mathcal{C}, \tau) \frac{(m+p)\text{rank}(M) \log(K)}{n} + \frac{8\mathcal{C} \log \left(\frac{2}{\varepsilon} \right)}{n} \right] \right\} \geq 1 - \varepsilon,$$

for some constant $\mathcal{C}(L, \mathcal{C}, \tau) > 0$ depending only on L, τ and \mathcal{C} . However, as the constant \mathcal{C} also depends on L, ξ, σ then $\mathcal{C}(L, \mathcal{C}, \tau)$ can be rewritten as $\mathcal{C}_{L, \xi, \sigma, \tau}$ as in the statement of the theorem. \square

References

- [1] ALQUIER, P., Bayesian methods for low-rank matrix estimation: Short survey and theoretical study. In *Algorithmic Learning Theory 2013*, pages 309–323. Springer, 2013. [MR3133074](#)
- [2] ALQUIER, P. and BIAU, G., Sparse single-index model. *The Journal of Machine Learning Research*, 14(1):243–280, 2013. [MR3033331](#)
- [3] ALQUIER, P., COTTET, V., CHOPIN, N., and ROUSSEAU, J., Bayesian matrix completion: Prior specification. *arXiv:1406.1440*, 2014.
- [4] ALQUIER, P. and LOUNICI, K., Pac-Bayesian bounds for sparse regression estimation with exponential weights. *Electronic Journal of Statistics*, 5:127–145, 2011. [MR2786484](#)
- [5] BENNETT, J. and LANNING, S., The netflix prize. In *Proceedings of KDD Cup and Workshop*, volume 2007, page 35, 2007.
- [6] BOUCHERON, S., LUGOSI, G., and MASSART, P., *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. [MR3185193](#)
- [7] CANDÈS, E. J. and PLAN, Y., Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.
- [8] CANDÈS, E. J. and RECHT, B., Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009. [MR2565240](#)
- [9] CANDÈS, E. J. and TAO, T., The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2010. [MR2723472](#)
- [10] CATONI, O., *A PAC-Bayesian Approach to Adaptive Classification*. Preprint Laboratoire de Probabilités et Modèles Aléatoires PMA-840, 2003.
- [11] CATONI, O., *Statistical Learning Theory and Stochastic Optimization*. Saint-Flour Summer School on Probability Theory 2001 (Jean Picard ed.), Lecture Notes in Mathematics. Springer, 2004. [MR2163920](#)

- [12] CATONI, O., *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 56. Institute of Mathematical Statistics, Beachwood, OH, 2007. [MR2483528](#)
- [13] DALALYAN, A. and TSYBAKOV, A. B., Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Machine Learning*, 72(1–2):39–61, 2008.
- [14] FOYGEL, R., SHAMIR, O., SREBRO, N., and SALAKHUTDINOV, R., Learning with the weighted trace-norm under arbitrary sampling distributions. In *Advances in Neural Information Processing Systems*, pages 2133–2141, 2011.
- [15] KLOPP, O., Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303, 2014. [MR3160583](#)
- [16] KOLTCHINSKII, V., LOUNICI, K., and TSYBAKOV, A. B., Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011. [MR2906869](#)
- [17] KOTECHA, J. H. and DJURIC, P. M., Gibbs sampling approach for generation of truncated multivariate Gaussian random variables. *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, 3:1757–1760, 1999.
- [18] LAWRENCE, N. D. and URTASUN, R., Non-linear matrix factorization with gaussian processes. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 601–608. ACM, 2009.
- [19] LIM, Y. J. and TEH, Y. W., Variational bayesian approach to movie rating prediction. In *Proceedings of KDD Cup and Workshop*, volume 7, pages 15–21, 2007.
- [20] MASSART, P., *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, Edited by Jean Picard. [MR2319879](#)
- [21] MCALLESTER, D., Some PAC-Bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 230–234, New York, 1998. ACM. [MR1811587](#)
- [22] NEGAHBAN, S. and WAINWRIGHT, M. J., Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012. [MR2930649](#)
- [23] RECHT, B. and RÉ, C., Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013. [MR3069879](#)
- [24] SALAKHUTDINOV, R. and MNIH, A., Bayesian probabilistic matrix factorization using Markov Chain Monte Carlo. In *Proceedings of the 25th International Conference on Machine Learning*, pages 880–887. ACM, 2008.
- [25] SHAWE-TAYLOR, J. and WILLIAMSON, R., A PAC analysis of a Bayes estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, pages 2–9, New York, 1997. ACM.

- [26] SUZUKI, T., Convergence rate of bayesian tensor estimation: optimal rate without restricted strong convexity. arXiv:1408.3092.
- [27] WILHELM, S., Package “tmvtnorm”, <http://cran.r-project.org/web/packages/tmvtnorm/>.
- [28] ZHOU, M., WANG, C., CHEN, M., PAISLEY, J., DUNSON, D., and CARIN, L., Nonparametric bayesian matrix completion. *Proc. IEEE SAM*, 2010.