Bayesian matrix completion: prior specification and consistency

Pierre Alquier(1,2), Vincent Cottet(1,3), Nicolas Chopin(3), Judith Rousseau(3,4)

(1) UCD School of Mathematical Sciences
(2) Insight: The National Centre for Big Data Analytics
(3) CREST-LS ENSAE
(4) CEREMADE, Université Paris Dauphine

Abstract

Low-rank matrix estimation from incomplete measurements recently received increased attention due to the emergence of several challenging applications, such as recommender systems; see in particular the famous Netflix challenge. While the behaviour of algorithms based on nuclear norm minimization is now well understood [SRJ05, SS05, CP09, CT09, CR09, Gro11, RT11, Klo11], an as yet unexplored avenue of research is the behaviour of Bayesian algorithms in this context. In this paper, we briefly review the priors used in the Bayesian literature for matrix completion. A standard approach is to assign an inverse gamma prior to the singular values of a certain singular value decomposition of the matrix of interest; this prior is conjugate. However, we show that two other types of priors (again for the singular values) may be conjugate for this model: a gamma prior, and a discrete prior. Conjugacy is very convenient, as it makes it possible to implement either Gibbs sampling or Variational Bayes. Interestingly enough, the maximum a posteriori for these different priors is related to the nuclear norm minimization problems. Our main contribution is to prove the consistency of the posterior expectation when the discrete prior is used. We also compare all these priors on simulated datasets, and on the classical MovieLens and Netflix datasets.

1 Introduction

We cite the introductory paper [BL07]: “In Oct. 2006 Netflix released a dataset containing 10^9 anonymous movie ratings and challenged the data mining, machine learning and computer science communities to develop systems that could beat the accuracy of its recommendation system.” This challenge (among others) generated a lot of excitement in the statistical community, and an increasing interest in the matrix completion problem. Seeing users as rows and movies as columns, the problem reduces to recovering a full matrix based on only a few of its entries. While, in general, this task is impossible, it becomes feasible when the matrix has low rank. In the Netflix problem, this amounts to assume (reasonably) the existence of a small number of typical patterns among users, eg,
those who like a particular type of movie. Note however that this recommendation system problem was studied since the 90s through collaborative filtering algorithms. An example is given by [HKBR99] on the open dataset MovieLens (available online http://grouplens.org/datasets/movielens/). The first attempt to perform recommendation through low-rank matrix completion algorithms is due to [SRJ05, SS05].

The methods used in this model usually rely on minimization of a measure of the fit to the observations penalized by the rank or the nuclear norm of the matrix (the nuclear norm is actually to be preferred as it leads to computationally feasible methods). A ground breaking result came from Candès and Tao [CT09] and Candès and Recht [CR09] when they exhibited conditions ensuring that the recovery of the matrix from a few experiments can be perfect. This result was extended to the context of noisy observations (in this case, the recovery of the matrix is not exact) in [CP09, Groll11] and efficient algorithms are proposed for example in [BR13].

A recent series of paper study a more general problem called trace regression, including matrix completion, as well as other popular models (linear regression, reduced rank regression and multi-task learning) as special cases [RT11, Klo11, KLT11]. These papers propose nuclear-norm penalized estimators, derive the reconstruction error of this method and also prove that this error is minimax-optimal: basically, the average quadratic error on the entries of an $m_1 \times m_2$ matrix with rank $r$ from the observation of $n$ entries cannot be better than $(m_1 \vee m_2)r/n$, where we use the notation $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$ for any real numbers $a, b$.

Bayesian estimation is also possible in this context. Based on priors defined for reduced rank regression [Gew96] and multi-task learning [YTS05], several authors proposed various Bayesian estimators for the matrix completion problem [LT07, SM08, LU09, YeZG09, PC10, ZWC+10, BLMK11]. The computational implementation of these estimators relies on either Gibbs sampling [SM08] or Variational Bayes (VB) methods [LT07], and these algorithms are fast enough to deal with such large datasets as Netflix or MovieLens. However, and contrary to penalized minimization methods, there has been little research on the theoretical properties of these Bayesian estimators; in particular on their consistency. In fact, our simulations suggest that the consistency will depend on the tuning of the hyperparameters in the prior.

Our contribution in this paper is threefold:

1. We study various families of prior distributions on the parameters which lead to tractable posterior distributions and thus to feasible algorithms, be them based on Gibbs sampling or VB approaches.

2. We compare these different priors on simulated and real datasets.

3. We prove consistency of the posterior distribution for a specific type of prior. The proof relies on PAC-Bayesian techniques [Cat03, Cat04, Cat07, DT12].

The extension of this convergence result to other priors is, to our knowledge, an open question.

The paper is organized as follows. In Section 2 we introduce the notations, review the conjugate priors in the literature and introduce the gamma and
discrete priors. Consistency for the discrete prior is stated in Theorem 3.1 in Section 3. The link between the maximum a posteriori (MAP) and the penalized minimization problems of [SRJ05, SS05, CP09, CT09, CR09, Gro11, RT11, Klo11, KLT11] is discussed in Section 4. A simulation study is provided in Section 5 in order to illustrate the strengths and weaknesses of each prior. The estimator that performs the best on the simulated datasets is tested on the MovieLens and Netflix dataset in Section 6. Finally, the proofs are postponed to Section 8.

2 Notations and priors

2.1 Notations

Given a matrix $M$, $M_i, \cdot$ will denote the $i$-th row of $M$ and $M, \cdot_h$ will denote the $h$-th column of $M$. Given a vector $x = (x_1, \ldots, x_d)$, $\text{diag}(x)$ will denote the $d \times d$ matrix $\text{diag}(x) = \begin{pmatrix} x_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & x_d \end{pmatrix}$.

Let $m_1, m_2$ and $n$ denote respectively the dimensions of the matrix $\theta$ and the number of observed entries, and define $\Theta$ as the set of $m_1 \times m_2$ matrices with real coefficients $\theta = (\theta_{i,j})_{1 \leq i \leq m_1, 1 \leq j \leq m_2}$. We fix an integer $K \leq m_1 \wedge m_2$ and we assume than the unknown matrix $\theta_0$ to be estimated may be written as $\theta_0 = M_0 (N_0)^T$ where $M_0$ is $m_1 \times K$ and $N_0$ is $m_2 \times K$ and, for some integer $r \in \{1, \ldots, K\}$, $M_0 = (M_0^1 \cdots M_0^r | 0 \cdots 0)$ and $N_0 = (N_0^1 \cdots N_0^r | 0 \cdots 0)$ (thus $M_0^h$ and $N_0^h$ are null when $h > r$). Note that when $K = m_1 \wedge m_2$, any matrix can be decomposed in such a way, with $r = \text{rank}(\theta_0)$. Here $r$ is unknown together with $M_0$, $N_0$ and $\theta_0$.

The observations are supposed to be distributed according to the following model: first, $n$ pairs $(i_k, j_k)$, $1 \leq k \leq n$, are drawn uniformly in $\{1, \ldots, m_1\} \times \{1, \ldots, m_2\}$, and observed with noise:

$$Y_k = \theta^0_{i_k, j_k} + \varepsilon_k$$  \hspace{1cm} (1)

for any $k \in \{1, \ldots, n\}$, where the $\varepsilon_k$ are centered independent random variables drawn from a common distribution, for which the only assumption is that it is sub-Gaussian with a known parameter $\sigma^2$. By this, we mean that $\log \mathbb{E} \exp(\varepsilon_k) \leq \ell^2 \sigma^2 / 2$. We denote by $\mathbb{P}$ the distribution induced by (1). We summarize the observation as $Y = (i_k, j_k, Y_k)_{k \in \{1, \ldots, n\}}$.

Note that the sub-Gaussian assumption is rather general: it encompasses the classical Gaussian noise $\varepsilon_k \sim \mathcal{N}(0, s^2)$ with $s \leq \sigma$ as well as bounded noise. For a dataset like Netflix, bounded noise makes more sense as we know that the observed ratings are between 1 and 5.

We end this subsection with two remarks. First, it must be kept in mind that in most applications, it does not make sense to assume that $n \to \infty$ with $m_1$ and
for \( n \) large enough, we observe all the entries in the matrix, so this is no longer a matrix completion problem. In the Netflix prize, \( m_1 = 480,189 \) users, \( m_2 = 17,770 \) movies and \( n = 100,480,507 \) ratings over \( m_1 m_2 = 8,532,958,530 \) entries in the matrix, which means that less than 1.2% of the entries of the matrix are observed. So, it might be more sensible to assume that \( m_1, m_2 \to \infty \) with \( n \) (and even \( r \)). Actually, we derive nonasymptotic bounds in this paper, see Theorem 3.1 in Section 3. However, one of the consequences of our result will be that, if \( m_1, m_2 \to \infty \) with \( n \) in such a way that we observe a constant fraction of the matrix, and if \( r \) is constant, then the Bayesian estimation of the matrix \( \theta^0 \) is consistent.

To define a Bayesian procedure we must specify a likelihood, i.e. the distribution of the noise \( \varepsilon_k \). In practice, this distribution is often unknown. Following the PAC-Bayesian approach, we use a Gaussian likelihood as a proxy for the true likelihood. We would like to emphasize that, according to our convergence result (Theorem 3.1 page 8), this leads to a consistent estimation for a large class of distributions for the noise \( \varepsilon_k \). Thus, for a given prior \( \pi(d\theta) \), we consider the following posterior distribution:

\[
\rho_\lambda(d\theta) \propto \exp \left[ -\frac{\lambda}{n} \sum_{k=1}^{n} (Y_k - \theta_{i_k,j_k})^2 \right] \pi(d\theta).
\]  

Taking \( \lambda = n / 2s^2 \) leads to the usual posterior when the noise is \( \mathcal{N}(0,s^2) \). However, according to the PAC-Bayesian theory \cite{Cat03, Cat04, Cat07, Alq08, Suz12, GA13, AB13}, and more generally to works on aggregation with exponential weights \cite{DT12, DS12, RT12}, one might want to consider a smaller \( \lambda \) in order to obtain theoretical guarantees for the larger class of noise introduced above.

Our estimator of \( \theta \) will be the posterior mean:

\[
\hat{\theta}_\lambda = \int_{\Theta} \theta \rho_\lambda(d\theta).
\]

### 2.2 Priors

All the aforementioned papers on Bayesian matrix completion assign conditional Gaussian priors to \( M \in \mathbb{R}^{m_1 \times K} \) and \( N \in \mathbb{R}^{m_2 \times K} \), with

\[
\theta = MN^T = \sum_{h=1}^{K} M_{\cdot,h} N_{\cdot,h}^T.
\]

More precisely, given some vector \( \gamma = (\gamma_1, \ldots, \gamma_K) \) with positive entries the columns \( M_{\cdot,h} \sim \mathcal{N}(0, \gamma_h I_{m_1}) \) and \( N_{\cdot,h} \sim \mathcal{N}(0, \gamma_h I_{m_2}) \) independently, where \( I_m \) denotes the \( m \times m \) identity matrix.

Note that we usually expect that \( \text{rank}(\theta) = K \). However, if the probability distribution on \( M \) and \( N \) ensures that most of the \( M_{\cdot,h} \) and \( N_{\cdot,h} \) are close to 0, then \( \theta \) will be close to a lower rank matrix.

On the other hand, Bayesian matrix completion methods usually differ in the choice of the prior distribution on \( \gamma \) and on \( r \).

1. As summarized in the survey paper \cite{Gew96}, in econometrics, for the estimation of simultaneous equation models, and then later for the reduced rank regression model, it is reasonable to assume that the rank
$r = \text{rank}(\theta^0)$ is known. In this case, it makes sense to fix $K = r$ and to consider a constant $\gamma$, e.g. $\gamma = (1, \ldots, 1)$. We denote by $\delta_1$ this prior (the Dirac mass at 1).

2. For large scale matrix completion problems, it does not make sense to assume that the rank $r$ is known. In this case, a reasonable approach is to set $K$ to a “large” value, so that $K \geq r$, and to consider $\gamma = (\gamma_1, \ldots, \gamma_K)$ itself as random. Using a prior that would enforce many of the $\gamma_h$ to be close to 0 would lead the prior to give more weight on (approximately) low-rank matrices. This is the approach followed in \cite{LT07, SM08, LU09, YLZG09, PC10, ZWC10, BLMK11}. Up to minor variants, all these authors propose to consider the $\gamma_h$ as iid inverse gamma with parameters $a$ and $b$, denoted by $\Gamma_{-1}(a,b)$ as it leads to conjugate marginal distributions for $\gamma_h$. (Among the possible variants, some authors consider $M_h \sim N(0, \gamma_h I_m)$ and $N_h \sim N(0, \gamma_h' I_m)$ for different vectors $\gamma_h$ and $\gamma_h'$ but this does not seem to give a better estimation of the unknown rank of the matrix; simulations tend to confirm this intuition).

In order to ensure that $K \geq r$ holds, it seems natural to take $K$ as large as possible, $K = m_1 \land m_2$, but this may computationally prohibitive if $K$ is large.

3. Based on a similar idea in the case of the Bayesian LASSO \cite{PC08}, one may assign instead to the $\gamma_h$’s independent gamma priors $\Gamma((m_1 + m_2 + 1)/2, \beta^2/2)$ for some $\beta > 0$. This leads to simple close-form expressions for the conditional distribution of $\gamma | M, N$ as discussed in the following Section.

4. Finally, the following prior with finite support has not yet been considered in the literature on matrix completion

$$\gamma_h \sim (1 - p)\delta_\epsilon + p\delta_C,$$

where $\epsilon$ is a small positive constant, $p \in (0, 1)$ and $C >> \epsilon$. This prior is similar in spirit to the spike and slab prior for variable selection \cite{MB88, GM93}.

2.3 Conjugacy, Gibbs sampling

The four types of prior distributions discussed in the previous section lead to the same conditional posterior distributions for the rows $M_i$, and $N_j$, (conditional on $\gamma = (\gamma_1, \ldots, \gamma_K)$ and on the data), which we now describe.

For $1 \leq i \leq m_1$, let $V_{i,N,\gamma}$ be the $K \times K$ matrix given by

$$V_{i,N,\gamma}^{-1} = \text{diag}(\gamma)^{-1} + \frac{2\lambda}{n} \sum_{k: i_k = i} N_{j_k,N} N_{j_k,N}^T,$$

and $m_{i,N,\gamma}$ the $1 \times K$ vector given by

$$m_{i,N,\gamma}^T = \frac{2\lambda}{n} V_{i,N,\gamma} \sum_{k: j_k = i} Y_{i_k,j_k} N_{j_k,N}^T.$$
Table 1: Conditional posterior distribution of $\gamma$ given $M$ and $N$ for different priors.

| Prior on $\gamma_h$ | Conditional posterior |
|----------------------|-----------------------|
| $\delta_1$          | $\delta_1$           |
| $\Gamma^{-1}(a,b)$  | $\Gamma^{-1}(\hat{a}_h, b_h)$ |
| $\Gamma((m_1 + m_2 + 1)/2, \beta^2/2)$ | $\mathcal{IG}(\hat{\mu}_h, \hat{\lambda}_h)$ |
| $(1-p)\delta_x + p\delta_C$ | $(1 - \hat{p}_h)\delta_x + \hat{p}_h\delta_C$ |

Then, given $N, \gamma$ and $Y$, the rows of $M$ are independent and

$$M_i^T | N, \gamma, Y \sim N(m_i^T_{N,\gamma}, V_{i,N,\gamma}). \quad (3)$$

Similarly, for $1 \leq j \leq m_2$, let $W_{j,M,\gamma}$ be the $K \times K$ matrix given by

$$W_{j,M,\gamma}^{-1} = \text{diag}(\gamma)^{-1} + \frac{2\lambda}{n} \sum_{k: j_k = j} M_{ik}M_{ik}^T,$$

and $n_{j,M,\gamma}$ the $1 \times K$ vector given by

$$n_{j,M,\gamma}^T = \frac{2\lambda}{n} W_{j,M,\gamma} \sum_{k: j_k = j} Y_{ik,j_k}M_{ik}^T.$$

Then, given $M, \gamma$ and $Y$, the rows of $N$ are independent and

$$N_j^T | M, \gamma, Y \sim N(n_{j,M,\gamma}^T, W_{j,M,\gamma}). \quad (4)$$

On the other hand, the priors discussed in the previous section generate different conditional posterior distributions for $\gamma$ given $M, N$ and the data, which are summarized in Table 1 where $\mathcal{IG}(\mu, \lambda)$ denotes the inverse Gaussian distribution with parameters $(\mu, \lambda)$ and

$$\hat{a}_h = a + \frac{m_1 + m_2}{2}, \quad \hat{b}_h = b + \frac{||M_{h}.||^2 + ||N_{h}.||^2}{2},$$

$$\hat{p}_h = \frac{\beta}{\sqrt{||M_{h}.||^2 + ||N_{h}.||^2}}, \quad \hat{\lambda}_h = \beta^2,$$

and finally,

$$\hat{p}_h = \frac{\pi_h}{\pi_h + \pi_h}$$

with

$$\pi_h = \frac{p}{C^{(m_1 + m_2)/2}} \exp\left(\frac{-||M_{h}.||^2 + ||N_{h}.||^2}{2C}\right),$$

$$\pi_h' = \frac{1 - p}{\epsilon^{(m_1 + m_2)/2}} \exp\left(\frac{||M_{h}.||^2 + ||N_{h}.||^2}{2\epsilon}\right).$$

We skip the calculations that lead to these expressions, as they are a bit tedious and follow from first principles. The surprising result in this array is
the simple expression obtained for the less common gamma prior (third row). We shall see in our simulations that this gamma prior actually leads to better performance than the more standard inverse gamma prior that has been used in most papers on Bayesian matrix completion.

Of course, the main motivation for deriving these conditional posterior distributions is to be able to implement Gibbs sampling to simulate from the joint posterior of \( M, N, \) and \( \gamma \) (and therefore \( \theta \)). The corresponding Gibbs sampler may be summarised as:

1. Simulate each row \( M_i, \) from (3).
2. Simulate each row \( N_j, \) from (4).
3. Simulate \( \gamma \) from the appropriate distribution from Table 1.

Given the typical size of matrix completion problems, it is essential to be able to implement a Gibbs sampler that updates jointly large blocks of random variables, as any other type of MCMC sampler (such as Metropolis-Hastings) would be likely to show very poor performance on such high-dimensional problems.

2.4 Variational Bayes

Using (conditionally) conjugate priors makes it also possible to quickly obtain a VB (Variational Bayes) approximation of the posterior, which is convenient when Gibbs sampling is too expensive, either because of bad mixing, or a high cost per iteration (large datasets), or both.

VB amounts to compute iteratively the optimal approximation of \( \rho_\lambda(M, N, \gamma) \) among a certain class of distributions; in our case, the class of factorized distributions \( q(M, N, \gamma) = q(M_i) q(N_j) q(\gamma_k) \). The optimality criterion is the Kullback-Leibler divergence between \( q(M, N, \gamma) \) and \( \rho_\lambda(M, N, \gamma) \), \( K(q, \rho_\lambda) \). The algorithm works iteratively, by updating each factor \( q(M), q(N), \) and \( q(\gamma) \), in turn.

Here again, we skip the tedious but elementary calculations. Note that, in the simulation section, we use this algorithm only with the inverse-gamma prior, so we only describe this version of the algorithm. First, it appears that the optimal factors \( q(M), q(N), \) and \( q(\gamma) \) necessarily factorise as:

\[
q(M) = \prod_{i=1}^{m_1} q(M_i), \quad q(N) = \prod_{j=1}^{m_2} q(N_j), \quad q(\gamma) = \prod_{k=1}^{K} q(\gamma_k)
\]

where \( q(M_i) \) is \( \mathcal{N}(m_i^T, V_i) \), \( q(N_j) \) is \( \mathcal{N}(n_j^T, W_j) \) and \( q(\gamma_k) \) is \( IG(a + (m_1 + m_2)/2, b_k) \) for some \( m_1 \times K \) matrix \( m \) whose rows are denoted by \( m_i \), and some \( m_2 \times K \) matrix \( n \) whose rows are denoted by \( n_j \), and some vector \( b = (b_1, \ldots, b_K). \) The parameters are updated iteratively through the formulas

1. moments of \( M \):

\[
m_{i}^T := \frac{2\lambda}{n} V_i \sum_{k:j_k=i} Y_{k,j_k} n_{j_k}^T,
\]

\[
V_i^{-1} := \frac{2\lambda}{n} \sum_{k:j_k=i} [W_{j_k} + n_{j_k} n_{j_k}^T] + \left( a + \frac{m_1 + m_2}{2} \right) \text{diag}(b)^{-1}
\]
2. moments of $N$:
\[
\mathbf{n}_j^T := \frac{2\lambda}{n} \mathbf{W}_j \sum_{k:j_k=j} \mathbf{Y}_{i} \mathbf{m}_{k}^T,
\]

\[
\mathbf{W}_j^{-1} := \frac{2\lambda}{n} \sum_{k:j_k=j} \left[ \mathbf{V}_i + \mathbf{m}_{k} \mathbf{m}_{k}^T \right] + \left( a + \frac{m_1 + m_2}{2} \right) \text{diag}(b)^{-1}
\]

3. moments of $\gamma$:
\[
b_k := \frac{1}{2} \left[ \sum_{i=1}^{m_1} \mathbf{m}_{i,k}^2 + (\mathbf{V}_i)_{k,k} \right] + \sum_{j=1}^{m_2} \left( \mathbf{n}_{j,k}^2 + (\mathbf{V}_j)_{k,k} \right)
\]

(where $(\mathbf{V}_i)_{k,k}$ denotes the $(k,k)$-th entry of the matrix $\mathbf{V}_i$ and $(\mathbf{W}_j)_{k,k}$ denotes the $(k,k)$-th entry of the matrix $\mathbf{W}_j$).

3 Consistency of the posterior distribution

We use, as a distance on matrices, the Frobenius norm, $\| \cdot \|_F$ given by, for any $\theta \in \Theta$,
\[
\| \theta \|_F^2 = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} \theta_{i,j}^2.
\]

Consider the posterior (2) associated to the prior defined in Section (2.2), with $\gamma$ following the discrete prior:
\[
\gamma_h \sim p\delta_{\varepsilon} + (1-p)\delta_C,
\]

for each $h$, independently. We then have the following posterior consistency result:

**Theorem 3.1** Assume that $\theta^0 = (M^0)(N^0)^T$ satisfies: there exists $c > 0$ such that for all $i \leq m_1$, $j \leq m_2$ and $\ell \leq K$, $|M^0_{i,\ell}|, |N^0_{j,\ell}| \leq \sqrt{c/K}$. If $\varepsilon = 1/(2nm_1m_2), C = 1/(288K), p \in (0,1), K = m_1 \wedge m_2$ (the dimension of $\gamma$) and $\lambda = \sqrt{Kn}$ where $\lambda$ is defined in [2], then, as soon as $n \geq (m_1 \wedge m_2)$, for any $\eta \in (0,1)$,

\[
\frac{1}{m_1m_2} \| \hat{\theta}_\lambda - \theta^0 \|_F^2 \leq \frac{r(m_1 \vee m_2)}{\sqrt{(m_1 \wedge m_2)n}} \left[ 576 + 2 \log n \left( r(m_1 \vee m_2) + 2 \right) \right] + \frac{4}{\sqrt{n}} \left[ 19 + 37c + 36\sigma^2 + \log \left( \frac{16\pi p^2 m_1 m_2}{\eta(1-p)} \right) \right]
\]

with probability greater than $1 - \eta$ under $\mathbb{P}$.

The proof relies on PAC-Bayesian bounds [Cat03, Cat07] and is provided in Section [8].

**Remark 3.1** The assumption $n \geq (m_1 \wedge m_2)$ is natural: otherwise, we can’t even have one observation per row and per column.
Remark 3.2 A similar assumption to \(|M_{i,ℓ}|, |N_{j,ℓ}| ≤ \sqrt{\frac{c}{K}}\) is used in order to prove results on non Bayesian estimators, see e.g. \([CP09]\). Here, this assumption implies that \(|θ_{i,j}| = \sum_{ℓ=1}^{K} M_{i,ℓ} N_{j,ℓ}| ≤ c\), which is a sensible assumption in the Netflix challenge as the ratings \(θ_{i,j}\) are all in the finite set \(\{1, 2, 3, 4, 5\}\).

Remark 3.3 In the model described in Section 2, the minimax-optimal rate of convergence for the square Frobenius distance is given by \(\frac{r(m_1 ∨ m_2)}{\sqrt{n}}\), see \([KLT11, RT11]\). So, our rate is not minimax-optimal. For example, in the case of a square matrix, \(m_1 = m_2 = m\), the minimax rate is \(\frac{cm}{n}\); hence, we only have \(r\sqrt{\frac{m}{n}} \log(nm)\). As soon as \(r^2 mn^2 = o(n)\), the posterior distribution is however consistent. To our knowledge, this is the first convergence result for a Bayesian estimator in this context. When the rate of observed entries is constant, that is \(\frac{n}{m^2} = α\), then for a fixed rank, the reconstruction error is in \(r\sqrt{\frac{1}{m}} \log(m)\) and so, when \(m\) grows, we reconstruct perfectly the matrix.

[Alq13] proved that the Bayesian estimator using the inverse gamma prior is minimax-optimal in a different but related model, the reduced rank regression. The proof may be easily adapted to the discrete prior.

The questions:

- is the Bayes estimator based on the inverse gamma prior consistent in the matrix completion problem?
- is the Bayes estimator based on one of the priors introduced in Section 2 minimax-optimal in the matrix completion problem?

are, to our knowledge still open problems, and will be the objects of future work.

4 Link with minimization problems

In this section, we highlight some connections between Bayesian estimation based on a certain prior, as discussed in Section 2 and penalized estimators based on penalty terms that are popular for matrix completion (or other problems). More precisely, we show that, for a given prior, the MAP (maximum a posteriori), that is the mode of the posterior density, may be recovered as a certain penalized estimator. The motivation is to provide additional insight into the choice of the prior distribution. In particular, we shall see that the gamma prior corresponds to a certain penalty function, which is popular and easy to interpret, but that may not be easy to implement directly.

4.1 Prior \(δ_1\)

When the prior is \(δ_1\), the MAP is

\[
\arg \min_{θ=MN^T} \left\{ \frac{λ}{n} \sum_{k=1}^{n} (Y_k - θ_{i_k,j_k})^2 + \frac{\|M\|_F^2 + \|N\|_F^2}{2} \right\} = \arg \min_{θ=MN^T} \left\{ \frac{λ}{n} \sum_{h=1}^{n} (Y_h - θ_{i_h,j_h})^2 + \sum_{ℓ=1}^{K} \frac{\|M_{i,ℓ}\|^2 + \|N_{j,ℓ}\|^2}{2} \right\}.
\]
The penalization is very similar to the ridge penalty used for regression problems. It is a classical result that when $K = m_1 \wedge m_2$, 

$$\|\theta\|_* = \inf_{M,N} \frac{\|M\|_F^2 + \|N\|_F^2}{2}$$

where $\|\theta\|_*$ is the nuclear norm of $\theta$ (see e.g. Equation 2 page 203 in [RR13] or Lemma 1 in [SRJ05]). So, this MAP can be rewritten as 

$$\arg\min_\theta \left\{ \sum_{k=1}^n (Y_k - \theta^*_{i_k,j_k})^2 + \|\theta\|_* \right\}$$

and linked with the penalization problems studied in [RR13, RT11, Klo11, KLT11].

### 4.2 Inverse-gamma prior

When we use the prior $\Gamma^{-1}(a,b)$, the MAP is 

$$\arg\min_{\theta=MNT,\gamma} \left\{ \sum_{k=1}^n \frac{\lambda}{n} (Y_k - \theta^*_{i_k,j_k})^2 \right. $$

$$\left. + \frac{1}{2} \sum_{\ell=1}^K \left[ \frac{\|M_{\ell}\|_2^2 + \|N_{\ell}\|_2^2 + b}{\gamma_{\ell}} - (a - 1) \log(\gamma_{\ell}) + b \gamma_{\ell} \right] \right\}.$$ 

When the $\gamma_h$ are fixed, we can interpret this as a weighted ridge regression. For small $\gamma_h$, $M_{\ell}$ and $N_{\ell}$ are close to zero. So, the essential rank of $\theta$ is the number of $\gamma_h$ that are not too small. On the other hand, the penalization $(a + 1) \log(\gamma_{\ell})$ will cause many $\gamma_h$ to be small.

### 4.3 Gamma prior

When we use a prior $\Gamma(a,b)$, the MAP is 

$$\arg\min_{\theta=MNT,\gamma} \left\{ \sum_{k=1}^n \frac{\lambda}{n} (Y_k - \theta^*_{i_k,j_k})^2 \right. $$

$$\left. + \frac{1}{2} \sum_{\ell=1}^K \left[ \frac{\|M_{\ell}\|_2^2 + \|N_{\ell}\|_2^2}{\gamma_{\ell}} - (a - 1) \log(\gamma_{\ell}) + b \gamma_{\ell} \right] \right\}.$$ 

The interpretation is similar to the one in the inverse gamma case. Note that, we used this prior with parameters $(a,b) = (m_1 + m_2 + 1)/2, \beta^2/2)$. In this case, there is an interesting phenomenon. If we do not consider the MAP with respect to $M$, $N$, and $\gamma$, but instead integrate with respect to $\gamma$ and only consider the MAP with respect to $M$ and $N$, the estimator is actually similar to Yuan and Lin’s group-LASSO estimator [YLD06].

**Proposition 4.1** The MAP of the marginal posterior distribution of $\theta$ under the Gamma prior is given by 

$$\arg\min_{\theta=MNT} \left\{ \sum_{k=1}^n \frac{\lambda}{n} (Y_k - \theta^*_{i_k,j_k})^2 + \beta \sum_{\ell=1}^K \sqrt{\|M_{\ell}\|_2^2 + \|N_{\ell}\|_2^2} \right\}.$$ 

10
The proof is given in Section 5. Note that, contrary to the group-LASSO, this optimization problem has no reason to be convex, and might not lead to feasible algorithms for large scale problems.

On the other hand, it gives a nice extra motivation for the gamma prior: this prior tends to set some columns of $M$ and $N$ to 0 in the same way than the group LASSO set some group of coefficients to 0 simultaneously.

5 Simulations

We first compare all the Bayesian estimators corresponding to the different priors on a toy example. Following [CP09], we generate a square matrix $\theta^0$ (ie $m_1 = m_2 = m$) with rank $r = 2$ in the following way: $\theta^0 = M^0(N^0)^T$ where the entries of the $m \times 2$ matrices $M^0$ and $N^0$ are iid $\mathcal{N}(0, 20/\sqrt{m})$.

We observe 20% of the entries matrix, $n = 0.2m^2$, corrupted by a $\mathcal{N}(0, 1)$ noise. The performance of each estimator $\hat{\theta}$ is measured through its RMSE

$$\text{RMSE} = \sqrt{\frac{1}{m_1m_2} \| \hat{\theta} - \theta^0 \|_F^2} = \frac{1}{m} \| \hat{\theta} - \theta^0 \|_F.$$

In a first set of experiments, we fix $K = 5$ and study the convergence of the estimators when $m$ grows, $m \in \{100, 200, 500, 1000\}$. In a second set of experiments, we fix $m = 500$ and study the effect of $K$ on the performance of the different estimators, $K \in \{2, 5, 10, 20\}$.

Note that in this simulation study, we always use Gibbs sampling to simulate from the posterior. The convergence of the chain seems very quick, as illustrated by Figure 1 which is taken from one of the simulations for $m = 200$, $K = 5$ with the Bernoulli prior with $(C, p, \varepsilon) = (1, 0.05, 0.05)$. The autocorrelations for each entry of the matrix $\theta$ are not large and vanishes after 2 or 3 lags, as shown by Figure 2. In any case, we let the Gibbs sampler run 1000 iterations and remove the first 100 iterations as a burn-in period. The thinning parameter is set to 10.

Figure 1: RMSE by iterations, $s$ for $m = 200$, $K = 5$ with the Bernoulli prior with $(C, p, \varepsilon) = (1, 0.05, 0.05)$. 
Figure 2: ACF for some entries of $\theta$, s for $m = 200, K = 5$ with the Bernoulli prior with $(C, p, \varepsilon) = (1, 0.05, 0.05)$.

Table 2: Experiments with fixed $K$: RMSE for different priors and different values of $m$.

| $m$ | prior distribution | 100  | 200  | 500  | 1000 |
|-----|--------------------|------|------|------|------|
|     | Fixed              | .75  | .47  | .27  | .18  |
|     | Gamma              | .60  | .37  | .23  | .16  |
|     | Inverse Gamma      | .59  | .39  | .25  | .18  |
|     | Discrete           | .60  | .36  | .22  | .16  |

5.1 Convergence when $m$ grows

The results of the experiments with $K$ fixed are reported in Table 2. For each prior, we tried many hyperparameters and only report the best results. The corresponding hyperparameters are reported in Table 3. First, it appears that the results from the four different priors are very close. Note that the choice $a = 1$ and $b \ll 1$ in the inverse Gamma distribution was done according to the theoretical results in [Alq13] in reduced rank regression. The non-adaptive prior always performs worse than the adaptive priors though, as expected in this case where $r = 2 < K = 5$.

The results improve with the size of the data in line with Theorem 3.1.

5.2 $m = 500$

We focus on the case $m = 500$ to explore the behaviour of the results when the size of the parameters varies. The rationale beyond this is that, in real-life applications, we don’t know the rank $r$ of $\theta^0$, so we usually set $K$ “too large”
Table 3: Experiments with fixed $K$: choice of hyperparameters for the different considered priors.

| prior distribution | 100  | 200  | 500  | 1000 |
|--------------------|------|------|------|------|
| Fixed              | $\gamma = 0.2$ | $\gamma = 1$ | $\gamma = 7$ | $\gamma = 10$ |
| Gamma              | $\beta^2 = 500$ | $\beta^2 = 2000$ | $\beta^2 = 10000$ | $\beta^2 = 40000$ |
| Inverse Gamma ($a = 1$) | $b = 0.015$ | $b = 0.012$ | $b = 0.005$ | $b = 1, 0.007$ |
| Discrete ($((C, p) = (1, 0.05))$) | $\varepsilon = 0.11$ | $\varepsilon = 0.08$ | $\varepsilon = 0.05$ | $\varepsilon = 0.03$ |

Table 4: Experiments with fixed $m$: RMSE for different priors and different choices for $K$.

| prior distribution | Fixed | Gamma | Inv. Gamma | Discrete |
|--------------------|-------|-------|------------|----------|
| $K = 2$            | .22   | .22   | .22        | .22      |
| $K = 5$            | .27   | .23   | .25        | .22      |
| $K = 10$           | .31   | .23   | .26        | .22      |
| $K = 20$           | .37   | .22   | .27        | .22      |

and hope for adaptation as explained in Section 2. The results are reported in Table 4. The corresponding hyperparameters are reported in Table 5. As expected, the non-adaptive $\delta_1$ prior does not lead to stable results and get worse when $K$ is too large. The lowest RMSE is achieved when the size is equal to the true rank. The three “adaptive” estimators perform better. However, it is to be noted that the best performance is reached by the gamma and the discrete distributions, while the inverse gamma is the most popular in the literature. As expected, these priors adapt automatically to the true rank of the matrix $\theta^0$, $r = 2$, and taking a large $K$ do not deteriorate the performances of the estimators.

Note that the (rather) poor performance of the inverse gamma distribution seems to be caused by a slower convergence of the MCMC algorithm in this context.

Table 5: Experiments with fixed $m$: choice of hyperparameters for the different considered priors.

| prior distribution | Fixed | Gamma | Inv. Gamma, $a = 1$ | Discrete, $((C, p) = (1, 0.05))$ |
|--------------------|-------|-------|---------------------|----------------------------------|
| $K = 2$            | $\gamma = 1$ | $\beta^2 = 5000$ | $b = 0.001$ | $\varepsilon = 0.05$ |
| $K = 5$            | $\gamma = 7$ | $\beta^2 = 10000$ | $b = 0.005$ | $\varepsilon = 0.05$ |
| $K = 10$           | $\gamma = 6$ | $\beta^2 = 12500$ | $b = 0.006$ | $\varepsilon = 0.03$ |
| $K = 20$           | $\gamma = 6$ | $\beta^2 = 13000$ | $b = 0.003$ | $\varepsilon = 0.02$ |
Figure 3: Posterior distribution approximated by Variational Bayes (VB) and the Gibbs sampler (GS) on a few randomly selected entries of the matrix. The prior is the inverse gamma prior.

6 Test on MovieLens

We now test the Bayesian estimators with the discrete prior and with the inverse gamma prior on the MovieLens dataset, available online: [http://grouplens.org/datasets/movielens/](http://grouplens.org/datasets/movielens/)

There are actually three different datasets with respectively about 100K, 1M and 10M ratings. Note that this is a challenging situation, because the size of the matrix $\theta^0$ makes the Gibbs iteration very slow, preventing from doing 1000 iterations, at least in the 1M and 10M cases. In this case, it is tempting to use the Variational Bayes algorithm (VB) [Bis06] instead of the Gibbs sampler. VB has been used on the Netflix challenge in [LT07] with a similar model.

The dataset is split into two parts, the training set (80 %) and the test set (20 %). The model is fitted on the training set and we measure the RMSE on the other part.

First, we compared VB and Gibbs on the 100K dataset in Figure [6]. It appears that the distribution of the matrix is quite different at least for a few
Table 6: Tests on the various MovieLens datasets.

| Dataset | Algorithm | prior        | hyperparameters | RMSE |
|---------|-----------|--------------|-----------------|------|
| 100K    | GS        | Discrete     | \((C, p, \varepsilon) = (1, 0.05, 0.07)\) | .92  |
| 100K    | GS        | Inverse Gamma| \((a, b) = (1, 0.1)\)          | .92  |
| 100K    | VB        | Inverse Gamma| \((a, b) = (1, 0.1)\)          | .92  |
| 1M      | VB        | Inverse Gamma| \((a, b) = (1, 0.1)\)          | .84  |
| 10M     | VB        | Inverse Gamma| \((a, b) = (1, 0.1)\)          | .79  |

Figure 4: Convergence of the VB algorithm on the three MovieLens dataset.

entries, but in the end, the performance of the approximation in terms of RMSE are comparable, as shown by Table 6. Also, in this case, the estimator based on the inverse gamma prior performs as well as the one base on the discrete prior. So, we only used the inverse gamma prior with the VB algorithm for the more time-consuming tests on the 1M and 10M dataset. As shown by figure 4 even in the 10M dataset, the VB algorithm converges in less than 20 iterations.

7 Conclusion

We reviewed the popular priors in Bayesian matrix completion (non-adaptive and inverse gamma priors) and proposed two new priors (gamma and discrete priors). We proved that the Bayesian estimator based on the discrete prior is consistent, and demonstrated the efficiency of all these estimators on simulated and real-life datasets. Future work should include the study of the optimality of these estimators. Extension to tensors in the spirit of [GRY11, TS13], would also be of interest.
8 Proofs

8.1 Proof of Theorem 3.1

**Definition 8.1** We define, for any $\theta \in \Theta$,

$$T_k(\theta) = Y_k - \theta_{i_k,j_k}, \quad r(\theta) = \frac{1}{n} \sum_{k=1}^{n} T_k(\theta)^2, \quad R(\theta) = \mathbb{E}[r(\theta)].$$

Note that

$$R(\theta) = \frac{1}{m_1m_2} \|\theta - \theta^0\|^2_{\mathcal{F}} + \mathbb{E}(\varepsilon_2^2).$$

**Lemma 8.1** The variable $T_k(\theta) - \mathbb{E}[T_k(\theta)]$ is centered and sub-Gaussian with constant $4C(\theta)$ where $C(\theta) = 36 \left( \sigma^2 + \frac{c}{K} + \sup_{i,j} |\theta_{i,j}| \right)$.

**Proof of Lemma 8.1** Fix $q \geq 1$, using the sub-Gaussian assumption on $\varepsilon_k$ and Theorem 2.1 page 26 in [BLM13],

\[
T_k(\theta) - \mathbb{E}[T_k(\theta)] = \varepsilon_k + (\theta_{i_k,j_k}^0 - \theta_{i_k,j_k}) - \frac{1}{m_1m_2} \sum_{i,j} (\theta_{i,j}^0 - \theta_{i,j})
\]

\[
|T_k(\theta) - \mathbb{E}[T_k(\theta)]| \leq |\varepsilon_k| + 2\sup_{i,j} |\theta_{i,j}^0| + 2\sup_{i,j} |\theta_{i,j}|
\]

\[
|T_k(\theta) - \mathbb{E}[T_k(\theta)]|^{2q} \leq 9^q [\varepsilon_k^{2q} + 4^q \sup_{i,j} |\theta_{i,j}^0|^{2q} + 4^q \sup_{i,j} |\theta_{i,j}|^{2q}]
\]

\[
\mathbb{E} \left[ |T_k(\theta) - \mathbb{E}[T_k(\theta)]|^{2q} \right] \leq 9^q [q! (2\sigma^2)^q + 4^q \sup_{i,j} |\theta_{i,j}^0|^{2q} + 4^q \sup_{i,j} |\theta_{i,j}|^{2q}]\]

\[
\leq q! \left[ 36 \left( \sigma^2 + \frac{c}{K} + \sup_{i,j} |\theta_{i,j}| \right) \right] = q! C(\theta)^q
\]

so, here again according to Theorem 2.1 in [BLM13], $T_k(\theta) - \mathbb{E}[T_k(\theta)]$ is sub-Gaussian with parameter $4C(\theta)$. □

**Lemma 8.2** For any $\lambda > 0$,

$$\mathbb{E} \exp \left\{ \lambda [r(\theta) - R(\theta)] \right\} \leq \exp \left[ \frac{2\lambda^2 C(\theta)}{n} \right],$$

and

$$\mathbb{E} \exp \left\{ \lambda [R(\theta) - r(\theta)] \right\} \leq \exp \left[ \frac{2\lambda^2 C(\theta)}{n} \right].$$

**Proof of Lemma 8.2** we have

\[
\mathbb{E} \exp \left\{ \lambda [r(\theta) - R(\theta)] \right\} = \prod_{k=1}^{n} \mathbb{E} \exp \left\{ \frac{\lambda}{n} (T_k(\theta) - \mathbb{E}[T_k(\theta)]) \right\}
\]

\[
\leq \prod_{k=1}^{n} \exp \left( \frac{2\lambda^2 C(\theta)}{n^2} \right) \quad \text{through Lemma 8.1}
\]

\[
= \exp \left[ \frac{2\lambda^2 C(\theta)}{n} \right].
\]

For the proof of the reverse inequality, just replace $\lambda$ by $-\lambda$. □
Lemma 8.3 As soon as $72\lambda^2/n < \frac{1}{2C}$, we have

\[
\int \mathbb{E} \exp \left\{ \lambda \left[ R(\theta) - r(\theta) \right] \right\} \pi(d\theta) 
\leq \exp \left[ \frac{72\lambda^2 C}{2} + 72\lambda^2 \sigma^2 \right] + \log(m_1 m_2) + K \log \left( \frac{1}{1 - 144C\lambda^2/n} \right) .
\]

Proof of Lemma 8.3: from Lemma 8.2,

\[
\int \mathbb{E} \exp \left\{ \lambda \left[ R(\theta) - r(\theta) \right] \right\} \pi(d\theta) 
\leq \int \exp \left[ \frac{2\lambda^2 C(\theta)}{n} \right] \pi(d\theta)
\]

so we have to upper bound, for $\zeta = 72\lambda^2/n$,

\[
\int \exp \left[ \zeta \sup_{i,j} |\theta_{i,j}|^2 \right] \pi(d\theta) = \int \sup_{i,j} \exp \left[ \zeta |\theta_{i,j}|^2 \right] \pi(d\theta)
\]

\[
\leq m_1 m_2 \int \exp \left[ \zeta |\theta_{1,1}|^2 \right] \pi(d\theta)
\]

\[
= m_1 m_2 \mathbb{E}_{(M,N,\gamma) \sim \pi} \exp \left\{ \zeta \left[ \sum_{\ell=1}^K M_{1,\ell} N_{1,\ell} \right]^2 \right\}
\]

\[
\leq m_1 m_2 \left[ \mathbb{E}_{(M,N,\gamma) \sim \pi} \exp \left( \zeta \sum_{\ell=1}^K M_{1,\ell}^2 \right) \right]
\times \left[ \mathbb{E}_{(M,N,\gamma) \sim \pi} \exp \left( \zeta \sum_{\ell=1}^K N_{1,\ell}^2 \right) \right]
\]

\[
= m_1 m_2 \left[ \mathbb{E}_{(M,N,\gamma) \sim \pi} \exp \left( \zeta M_{1,1}^2 \right) \right]^{2K}
\leq m_1 m_2 \left[ \mathbb{E}_{U \sim N(0,C)} \exp \left( \zeta U^2 \right) \right]^{2K}
\]

\[
= m_1 m_2 \left[ \frac{1}{\sqrt{2\pi C}} \int_{-\infty}^{\infty} \exp \left( -u^2/2C + \zeta u^2 \right) du \right]^{2K}
\]

\[
= \left( \frac{m_1 m_2}{(1 - 2\zeta C)^K} \right)
\]

as soon as $\zeta \leq \frac{1}{2C}$. The proof of the second inequality is similar. □

Lemma 8.4 As soon as $72\lambda^2/n < \frac{1}{2C}$, we have, for any $\eta' > 0$,

\[
P \left\{ \exists \rho, \int R(\theta) \rho(d\theta) \geq \int r(\theta) \rho(d\theta) + \frac{72\lambda (\frac{\zeta}{C} + \sigma^2)}{n} \right. 
\]

\[
+ K(\rho, \pi) + \log \left( \frac{m_1 m_2}{\eta'} \right) + K \log \left( \frac{1}{1 - 144C\lambda^2/n} \right) \right\} \leq \eta'
\]

and

17
\[ \mathbb{P} \left\{ \exists \rho, \int r(\theta) \rho(d\theta) \geq \int R(\theta) \rho(d\theta) + \frac{72\lambda \left( \frac{c}{R} + \sigma^2 \right)}{n} + \frac{\mathcal{K}(\rho, \pi)}{\lambda} + \log \left( \frac{m_1 m_2}{\eta^2} \right) + K \log \left( \frac{1}{1 - 144C\lambda^2/n} \right) \right\} \leq \eta'. \]

**Proof of Lemma 8.4:** this proof follows the standard argument for PAC-Bayesian bounds \cite{Cat03, Cat04, Cat07}. From the first inequality in Lemma 8.3,

\[ \int \mathbb{E} \exp \left\{ \lambda [R(\theta) - r(\theta)] - \frac{72\lambda^2 c}{n} + \frac{72\lambda^2 \sigma^2}{n} - \log \left( \frac{m_1 m_2}{\eta'} \right) - K \log \left( \frac{1}{1 - 144\lambda^2/n} \right) \right\} \pi(d\theta) \leq \eta'. \]

Fubini leads to:

\[ \mathbb{E} \int \exp \left\{ \lambda [R(\theta) - r(\theta)] - \frac{72\lambda^2 c}{n} + \frac{72\lambda^2 \sigma^2}{n} - \log \left( \frac{m_1 m_2}{\eta'} \right) - K \log \left( \frac{1}{1 - 144\lambda^2/n} \right) \right\} \pi(d\theta) \leq \eta'. \]

Using 5.2.1 page 159 in \cite{Cat04} we obtain

\[ \mathbb{E} \sup_\rho \left\{ \int \lambda [R(\theta) - r(\theta)] - \frac{72\lambda^2 c}{n} + \frac{72\lambda^2 \sigma^2}{n} - \log \left( \frac{m_1 m_2}{\eta'} \right) - \mathcal{K}(\rho, \pi) - K \log \left( \frac{1}{1 - 144\lambda^2/n} \right) \right\} \rho(d\theta) \leq \eta'. \]

Finally, the inequality \( 1_{[0, +\infty)}(x) \leq \exp(x) \) leads to

\[ \mathbb{P} \left\{ \exists \rho, \int R(\theta) \rho(d\theta) \geq \int r(\theta) \rho(d\theta) + \frac{72\lambda \left( \frac{c}{R} + \sigma^2 \right)}{n} + \frac{\mathcal{K}(\rho, \pi)}{\lambda} + \log \left( \frac{m_1 m_2}{\eta^2} \right) + K \log \left( \frac{1}{1 - 144C\lambda^2/n} \right) \right\} \leq \eta'. \]

The proof of the second inequality follows the same scheme, but we start from the second inequality in Lemma 8.3. \( \square \)

**Lemma 8.5** As soon as \( 72\lambda^2/n < \frac{1}{2C} \), we have

\[ \mathbb{P} \left\{ \int R(\theta) \rho_\lambda(d\theta) \leq \inf_\rho \left\{ \int R(\theta) \rho(d\theta) + \frac{144\lambda \left( \frac{c}{R} + \sigma^2 \right)}{n} + \frac{2\mathcal{K}(\rho, \pi)}{\lambda} + \log \left( \frac{2m_1 m_2}{\eta} \right) + K \log \left( \frac{1}{1 - 144C\lambda^2/n} \right) \right\} \geq 1 - \eta. \]
Proof of Lemma 8.5: here again, this proof follows [Cat03, Cat04, Cat07]. From the first inequality in Lemma 8.4, with \( \eta' = \frac{\eta}{2} \),

\[
P\left\{ \int R(\theta)\rho_\lambda(d\theta) \geq \int r(\theta)\rho_\lambda(d\theta) + \frac{72\lambda(\frac{c}{8} + \sigma^2)}{n} + \mathcal{K}(\rho_\lambda, \pi) + \log \left( \frac{2m_1m_2}{\eta} \right) + K \log \left( \frac{1}{1-144C\lambda^2/n} \right) \right\} \leq \frac{\eta}{2}
\]

and from 5.2.1 page 159 in [Cat04], this leads to

\[
P\left\{ \int R(\theta)\rho_\lambda(d\theta) \geq \inf_{\rho} \left[ \int r(\theta)\rho(d\theta) + \frac{72\lambda(\frac{c}{8} + \sigma^2)}{n} + \mathcal{K}(\rho, \pi) + \log \left( \frac{2m_1m_2}{\eta} \right) + K \log \left( \frac{1}{1-144C\lambda^2/n} \right) \right] \right\} \leq \frac{\eta}{2}
\]

A union bound with the second inequality in Lemma 8.4 leads to

\[
P\left\{ \int R(\theta)\rho_\lambda(d\theta) \geq \inf_{\rho} \left[ \int R(\theta)\rho(d\theta) + 2\frac{72\lambda(\frac{c}{8} + \sigma^2)}{n} + \mathcal{K}(\rho, \pi) + \log \left( \frac{2m_1m_2}{\eta} \right) + K \log \left( \frac{1}{1-144C\lambda^2/n} \right) \right] \right\} \leq \eta.
\]

Lemma 8.5 is the main tool to prove Theorem 3.1. Actually, we are going to fix a special \( \rho \), and upper bound explicitly the terms in the right-hand side.

Definition 8.2 We define \( \rho^*_\delta(d\theta) \propto \pi(d\theta) \mathbf{1}(\|M - M^0\|_F \leq \delta, \|N - N^0\|_F \leq \delta) \) for any fixed \( \delta > 0 \).

Lemma 8.6 We have

\[
\int R(\theta)\rho^*_\delta(d\theta) \leq R(\theta^0) + 4\delta^2 \left( \frac{c}{K} + \delta^2 \right).
\]

Proof of Lemma 8.6: this roughly follows the calculations in [Alq13]:

\[
\int R(\theta)\rho^*_\delta(d\theta) = \int \left( E(\varepsilon^2) + \frac{1}{m_1m_2} \|M^0(N^0)^T - MN^T\|_F^2 \right) \rho^*_\delta(dM, dN)
\]

\[
= R(\theta^0) + \int \left( \frac{2}{m_1m_2} \|M^0(N^0)^T - MN^T\|_F^2 \right) \rho^*_\delta(dM, dN)
\]

\[
\leq R(\theta^0) + \int \left( \frac{2}{m_1m_2} \|M^0(N^0 - N)^T\|_F^2 \right) \rho^*_\delta(dM, dN)
\]

\[
\leq R(\theta^0) + \int \left( \frac{2}{m_1m_2} \|M^0\|_F^2 \|N^0 - N\|_F^2 \right) \rho^*_\delta(dM, dN)
\]

19
We control the first term (the upper bound for the second term is exactly similar):
\[ + \frac{2}{m_1m_2} ||M^0 - M||_F^2 ||N||_F^2 \rho_\delta^a(dM, dN) \]
\[ \leq R(\delta^a) + 2\delta^2 \int \frac{1}{m_1m_2} (||M^0||_F^2 + ||N||_F^2) \rho_\delta^a(dM, dN) \]
\[ \leq R(\delta^a) + 2\delta^2 \int \frac{1}{m_1m_2} (||M^0||_F^2 + ||N^0||_F^2 + \delta^2) \rho_\delta^a(dM, dN) \]
\[ \leq R(\delta^a) + 2\delta^2 \int \frac{1}{m_1m_2} (c(m_1 + m_2) + \delta^2) \rho_\delta^a(dM, dN) \]
\[ \leq R(\delta^a) + 4\delta^2 \left( \frac{c}{K} + \delta^2 \right). \]

**Lemma 8.7** As soon as \(2\varepsilon(m_1 \lor m_2)K \leq \delta^2\) we have
\[ K(\rho_\delta^a, \pi) \leq 2(r \log(p/(1 - p)) + K \log p) + \frac{cr(m_1 + m_2)}{2KC} + \frac{\delta^2}{2C} \]
\[ + \frac{r(m_1 + m_2)}{2} \log(C(r(m_1 \lor m_2) + 2)/\delta^2) + 2 \log(4\pi) \]

**Proof of Lemma 8.7** we have
\[ K(\rho_\delta^a, \pi) = \log \frac{1}{\pi(\|M - M^0\|_F \leq \delta, ||N - N^0||_F \leq \delta)} \]
\[ = \log \frac{1}{\pi(\|M - M^0\|_F \leq \delta)} + \log \frac{1}{\pi(||N - N^0||_F \leq \delta)}. \]

We control the first term (the upper bound for the second term is exactly similar):
\[ \pi(\|M - M^0\|_F \leq \delta) = \sum_{\gamma \in \{\varepsilon, C\}^K} \pi(\|M - M^0\|_F \leq \delta | \gamma = g) \pi(\gamma = g) \]
\[ \geq \pi(\|M - M^0\|_F \leq \delta | \gamma = G)p^K-r(1-p)^r \]
where \(G = (C, \ldots, C, \varepsilon, \ldots, \varepsilon)\) (respectively \(r\) times and \(K - r\) times). Then, we use the small ball probabilities inequality, see for instance Lemma 5.2 of [vdVvZ08]:
\[ \pi(\|M - M^0\|_F \leq \delta | \gamma = G) \geq \exp \left\{ -\frac{\|M^0\|_F^2}{2C} \right\} \pi \left( \|M\|_F^2 \leq \delta^2 | \gamma = G \right) \]
\[ \geq \exp \left\{ -\frac{\|M^0\|_F^2}{2C} \right\} \mathbb{P} \left( CA^2(m_1r) \leq \delta^2/2 \right) \]
\[ \geq \exp \left\{ -\frac{\|M^0\|_F^2}{2C} \right\} \mathbb{P} \left( \varepsilon A^2(m_1(K - r)) \leq \delta^2/2 \right) \]
\[ \geq \frac{\exp \left\{ -\frac{\|M^0\|_F^2}{2C} - \frac{\delta^2}{2C} - r m_1 \log 2 - \log(\delta^2/2C) \right\}}{\Gamma(mr_1/2 + 1)} \]
\[ \mathbb{P} \left( \varepsilon A^2(m_1(K - r)) \leq \delta^2/2 \right) \]
where \(A^2(d)\) denotes a chi-square random variable with \(d\) degrees of freedom. If \(2\varepsilon m_1K \leq \delta^2\)
\[ \mathbb{P} \left( \varepsilon A^2(m_1(K - r)) \leq \delta^2/2 \right) \geq \frac{1}{2} \]
and

\[
\pi(\|M - M^0\|_F \leq \delta \mid \gamma = G) \\
\geq \exp\left\{ -\frac{\|M^0\|^2}{2C} - \frac{\delta^2}{4C} - \frac{\rho_m \log(2 - \log(\delta^2/2C))}{2}\right\} \\
\geq \exp\left\{ -\frac{\|M^0\|^2}{2C} - \frac{\delta^2}{4C} - \frac{\rho_m \log(2 - \log(\delta^2/2C) - \log(\delta_{1}/2+1)+1)}{2}\right\}
\]

We finally obtain:

\[
\pi(\|M - M^0\|_F \leq \delta) \\
\geq p^{K-r} (1-p)^r \frac{\exp\left\{ -\frac{\|M^0\|^2}{2C} - \frac{\delta^2}{4C} - \frac{\rho_m \log(C(rm_1 + 2)/\delta^2)}{2}\right\}}{4\pi}
\]

This leads to

\[
K(\rho_5, \pi) \leq 2 (r \log(p/(1-p)) + K \log p) + \frac{cr(m_1 + m_2)}{2KC} + \frac{\delta^2}{2C} \\
+ \frac{r(m_1 + m_2)}{2} \log(C(r(m_1 \lor m_2) + 2)/\delta^2) + 2 \log(4\pi).
\]

We are now ready to prove Theorem 3.1.

**Proof of Theorem 3.1.** We plug the bounds of Lemmas 8.6 and Lemma 8.7 into Lemma 8.5 to get:

\[
P\left\{ \int_R (\theta) \rho_\lambda(d\theta) \leq R(\theta^0) + 4\delta^2 \left( \frac{c}{K} + \frac{\delta^2}{4C^2} + \frac{1}{4KC^2} \right) + \frac{144\lambda \left( \frac{c}{K} + \sigma^2 \right)}{n} \\
+ \frac{r(m_1 + m_2)}{\lambda} \left\{ \frac{1}{KC} + \log \left( C(r(m_1 \lor m_2)) + 2 \right) / 2 \right\} \\
+ \frac{4K}{\lambda} \log \left( \frac{p}{1 - 1444\sigma^2 \lambda n} \right) + \frac{4r}{\lambda} \log \left( \frac{p}{1 - p} \right) + \frac{2}{\lambda} \log \left( \frac{8\pi m_1 m_2}{\eta} \right) \right\} \geq 1 - \eta.
\]

The choice \(\delta = 1/\sqrt{n}\) leads to

\[
P\left\{ \int_R (\theta) \rho_\lambda(d\theta) \leq R(\theta^0) + \frac{4}{n} \left( \frac{c}{K} + \frac{1}{n} + \frac{1}{4KC^2} \right) + \frac{144\lambda \left( \frac{c}{K} + \sigma^2 \right)}{n} \\
+ \frac{r(m_1 + m_2)}{\lambda} \left\{ \frac{1}{KC} + \log [n (C r(m_1 \lor m_2)) + 2] \right\} \\
+ \frac{4K}{\lambda} \log \left( \frac{p}{1 - 1444\sigma^2 \lambda n} \right) + \frac{4r}{\lambda} \log \left( \frac{p}{1 - p} \right) + \frac{2}{\lambda} \log \left( \frac{8\pi m_1 m_2}{\eta} \right) \right\} \geq 1 - \eta.
\]

Now, substitute \(1/(288K)\) to \(C\) to get

\[
\]
Note that (by Jensen’s inequality) is trivially satisfied (with an equality). Also, note that the constraint is trivially satisfied with our choice for $\lambda$.

We can also substitute $\sqrt{Kn}$ to $\lambda$, note that the constraint $72\lambda^2/n \leq 1/(2C)$ is trivially satisfied (with an equality). Also, note that the constraint $2e(m_1 \lor m_2)K \leq \delta^2$ can now be rewritten $\varepsilon \leq 1/[2n(m_1 \land m_2)] = 1/[2nm_1m_2]$ so it is trivially satisfied with our choice for $\varepsilon$. We obtain:

$$P\left\{ \int R(\hat{\theta})d\theta \leq R(\theta^0) + \frac{4}{n} \left( \frac{c}{K} + \frac{1}{n} + 72\sqrt{\frac{K}{n}} \right) + \frac{144\lambda}{n} \left( \frac{\sigma^2}{\lambda} + \sigma^2 \right) + \frac{r(m_1 + m_2)}{\lambda} \left( 288 + \log [n (r(m_1 \lor m_2) + 2)] \right) \right\} \geq 1 - \eta.$$
8.2 Proof of Proposition 4.1

Proof of Proposition 4.1: This idea of the proof comes from a similar argument for the Bayesian LASSO in [PC08], namely, to use the formula

\[
\int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp \left( \frac{z^2}{2s} \right) \frac{a^2}{2} \exp \left( -\frac{a^2 s}{2} \right) ds = \frac{a}{2} \exp (-a|z|)
\]

for any \( a \) and \( z \). We have:

\[
\rho_\lambda(M,N) = \int \rho_\lambda(M,N,d\gamma)
\]

\[
= \int \exp \left[ -\frac{\lambda n}{n} \sum_{k=1}^n (Y_k - (MN^T)_{i_k,j_k})^2 \right] \pi(M,N,d\gamma)
\]

\[
= \exp \left[ -\frac{\lambda}{n} \sum_{k=1}^n (Y_k - (MN^T)_{i_k,j_k})^2 \right] \int \pi(M,N,d\gamma)
\]

and then,

\[
\int \pi(M,N,d\gamma) \propto \prod_{\ell=1}^K \int_0^\infty \gamma_\ell^{-\frac{m_1+m_2}{2}} \exp \left[ -\frac{1}{2\gamma_k} \left( \|M_{.,k}\|^2 + \|N_{.,k}\|^2 \right) \right] \gamma_\ell^{\frac{m_1+m_2+1}{2}} \exp \left( -\frac{\beta^2}{2} \gamma_\ell \right) d\gamma_\ell
\]

\[
\propto \prod_{\ell=1}^K \int_0^\infty \gamma_\ell^{\frac{1}{2}} \exp \left[ -\frac{1}{2\gamma_k} \left( \|M_{.,k}\|^2 + \|N_{.,k}\|^2 \right) - \frac{\beta^2}{2} \gamma_\ell \right] d\gamma_\ell
\]

\[
\propto \exp \left[ -\beta \sum_{\ell=1}^K \sqrt{\|M_{.,k}\|^2 + \|N_{.,k}\|^2} \right]
\]

using (5). This ends the proof. □

Acknowledgements

We would like to thank Prof. Taiji Suzuki for insightful comments.

References

[AB13] P. Alquier and G. Biau. Sparse single-index model. Journal of Machine Learning Research, 14:243–280, 2013.

[Alq08] P. Alquier. PAC-Bayesian bounds for randomized empirical risk minimizers. Mathematical Methods of Statistics, 17(4):279–304, 2008.

[Alq13] P. Alquier. Bayesian estimation of low-rank matrices: short survey and theoretical study. In S. Jain, R. Munos, F. Stephan, and T. Zeugmann, editors, Proceedings of the 24th Conference on Algorithmic Learning Theory (ALT). Springer LNAI, 2013.
[Bis06] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[BL07] J. Bennett and S. Lanning. The netflix prize. In *Proceedings of KDD Cup and Workshop 07*, 2007.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities*. Oxford University Press, 2013.

[BLMK11] S. D. Babacan, M. Luessi, R. Molina, and A. K. Katsaggelos. Low-rank matrix completion by variational sparse bayesian learning. In *IEEE International Conference on Audio, Speech and Signal Processing*, pages 2188–2191. Prague (Czech Republic), 2011.

[Cat03] O. Catoni. A PAC-Bayesian approach to adaptative classification. Preprint Laboratoire de Probabilités et Modèles Aléatoires, 2003.

[Cat04] O. Catoni. *Statistical Learning Theory and Stochastic Optimization*. Springer Lecture Notes in Mathematics, 2004.

[Cat07] O. Catoni. *PAC-Bayesian Supervised Classification (The Thermodynamics of Statistical Learning)*, volume 56 of *Lecture Notes-Monograph Series*. IMS, 2007.

[CP09] E. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):625–636, 2009.

[CR09] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[CT09] E. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2009.

[DS12] A. Dalalyan and J. Salmon. Sharp oracle inequalities for aggregation of affine estimators. *The Annals of Statistics*, 40(4):2327–2355, 2012.

[DT12] A. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and langevin. *J. Comput. System Sci.*, 78(5):1423–1443, 2012.

[GA13] B. Guedj and P. Alquier. Pac-bayesian estimation and prevision in sparse additive models. *Electronic Journal of Statistics*, 7:264–291, 2013.

[Gew96] J. Geweke. Bayesian reduced rank regression in econometrics. *Journal of Econometrics*, 75:121–146, 1996.

[GM93] E.I. George and R.E. McCulloch. Variable selection via Gibbs sampling. 88(423):pp. 881–889, 1993.

[Gro11] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57:1548–1566, 2011.
[SS05] N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. In *Proceedings of COLT-2005*. Springer LNCS, 2005.

[Suz12] T. Suzuki. PAC-Bayesian bound for gaussian process regression and multiple kernel additive model. *JMLR Workshop and Conference Proceedings - COLT2012*, 23:8.1–8.20, 2012.

[TS13] R. Tomioka and T. Suzuki. Convex tensor decomposition via structured schatten norm regularization. *NIPS 2013 Proceedings*, 2013.

[vdVvZ08] A. W. van der Vaart and H. van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. *IMS Collections*, 3:200–222, 2008.

[YL06] M. Yuan and Y. Lin. Model selection and estimation in regression with groupe variables. *Journal of the Royal Statistical Society - series B*, 68(1):49–67, 2006.

[YLZG09] K. Yu, J. Lafferty, S. Zhu, and Y. Gong. Large-scale collaborative prediction using a non-parametric random effects model. In *Proceedings of the 26th annual International Conference on Machine Learning (ICML09)*. ACM, New York, 2009.

[YTS05] K. Yu, V. Tresp, and A. Schwaighofer. Learning Gaussian processes for multiple tasks. In *Proceedings of the 22th annual International Conference on Machine Learning (ICML05)*, 2005.

[ZWC+10] M. Zhou, C. Wang, M. Chen, J. Paisley, D. Dunson, and L. Carin. Nonparametric bayesian matrix completion. In *IEEE Sensor Array and Multichannel Signal Processing Workshop*, 2010.