Bayesian Methods for Low-Rank Matrix Estimation: Short Survey and Theoretical Study

Pierre Alquier

School of Mathematical Sciences - University College Dublin
528 James Joyce Library, Belfield
Dublin 4, Ireland
pierre.alquier@ucd.ie

Abstract. The problem of low-rank matrix estimation recently received a lot of attention due to challenging applications. A lot of work has been done on rank-penalized methods [1] and convex relaxation [2], both on the theoretical and applied sides. However, only a few papers considered Bayesian estimation. In this paper, we review the different type of priors considered on matrices to favour low-rank. We also prove that the obtained Bayesian estimators, under suitable assumptions, enjoys the same optimality properties as the ones based on penalization.

Keywords: Bayesian inference, collaborative filtering, reduced-rank regression, matrix completion, PAC-Bayesian bounds, oracle inequalities.

1 Introduction

The problem of low-rank matrix estimation recently received a lot of attention, due to challenging high-dimensional applications provided by recommender systems, see e.g. the NetFlix challenge [3]. Depending on the application, several different models are studied: matrix completion [2], reduced-rank regression [4], trace regression, e.g. [5], quantum tomogaphy, e.g. [6], etc.

In all the above mentionned papers, the authors considered estimators obtained by minimizing a criterion that is the sum of two terms: a measure of the quality of data fitting, and a penalization term that is added to avoid overfitting. This term is usually the rank of the matrix, as in [1], or, for computational reasons, the nuclear norm of the matrix, as in [2] (the nuclear norm is the sum of the absolute values of the singular values, it can be seen as a matrix equivalent of the vectors ℓ_1 norm). However, it is to be noted that only a few papers considered Bayesian methods: we mention [7] for a first study of reduced-rank regression in Bayesian econometrics, and more recently [8, 9, 10] for matrix completion and reduced-rank regression (a more exhaustive bibliography is given below).

The objective of this paper is twofold: first, in Section 2 we provide a short survey of the priors that have been effectively used in various problems of low-rank estimation. We focus on two models, matrix completion and reduced rank-regression, but all the priors can be used in any model involving low-rank matrix estimation.

S. Jain et al. (Eds.): ALT 2013, LNAI 8139, pp. 309–323, 2013.

Then, in Section 3 we prove a theoretical result on the Bayesian estimator in the context of reduced rank regression. It should be noted that for some appropriate choice of the hyperparameters, the rate of convergence is the same as for penalized methods, up to log terms. The theoretical study in the context of matrix completion will be the object of a future work.

2 Model and Priors

In this section, we briefly introduce two models: reduced rank regression, 2.1, and matrix completion, 2.2. We then review the priors used in these models, 2.3.

2.1 Reduced Rank Regression

In the matrix regression model, we observe two matrices X and Y with

$$Y = XB + \mathcal{E}$$

where X is an $\ell \times p$ deterministic matrix, B is a $p \times m$ deterministic matrix and $\mathcal{E}$ is an $\ell \times m$ random matrix with $\mathbb{E}(\mathcal{E}) = 0$. The objective is to estimate the parameter matrix B. This model is sometimes refered as multivariate linear regression, matrix regression or multitask learning. In many applications, it makes sense to assume that the matrix B has low rank, i.e. $\mathrm{rank}(B) \ll \min(p, m)$. In this case, the model is known as *reduced rank regression*, and was studied as early as [11, 12]. We refer the reader to the monograph [4] for a complete introduction.

Depending on the application the authors have in mind, additional assumptions on the distribution of the noise matrix $\mathcal{E}$ are used:

– the entries $\mathcal{E}_{i,j}$ of $\mathcal{E}$ are i.i.d., and the probability distribution of $\mathcal{E}_{1,1}$ is bounded, sub-Gaussian or Gaussian $\mathcal{N}(0, \sigma^2)$. In this case, note that the likelihood of any matrix β is given by

$$\mathcal{L}(\beta|Y,\sigma) \propto \exp\left\{-\frac{1}{2\sigma^2}\|Y - X\beta\|_F^2\right\}$$

where we let $\|M\|_F$ denote the Frobenius norm, $\|M\|_F^2 = \mathrm{Tr}(M^T M)$.
– as a generalization of the latter case, it is often assumed in econometrics papers that the rows $\mathcal{E}_i$ of $\mathcal{E}$ are i.i.d. $\mathcal{N}_m(0, \Sigma)$ for some $m \times m$ variance-covariance matrix Σ.

In order to estimate B, we have to specify a prior on B and, depending on the assumptions on $\mathcal{E}$, a prior on σ or on Σ. Note however that in most theoretical papers, it is assumed that σ is known, or can be upper bounded, as in [1]. This assumption is clearly a limitation but it makes sense in some applications: see e.g. [6] for quantum tomography (that can bee seen as a special case of reduced rank regression).

In non-Bayesian studies, the estimator considered is usually obtained by minimizing the least-square criterion $\|Y - XB\|_F^2$ penalized by the rank of the matrix [1] or the nuclear norm [13]. In [1], the estimator $\hat{B}$ obtained by this method is shown to satisfy, for some constant $\mathcal{C} > 0$,

$$\mathbb{E}(\|X\hat{B} - XB\|_F^2) \leq \mathcal{C}\sigma^2 \text{rank}(B)(\text{rank}(X) + m)$$

(Corollary 6 p. 1290).

2.2 Matrix Completion

In the problem of matrix completion, one observes entries $Y_{i,j}$ of an $\ell \times m$ matrix $Y = B + \mathcal{E}$ for (i,j) in a given set of indices I. Here again, the noise matrix satisfies $\mathbb{E}(\mathcal{E}) = 0$ and the objective is to recover B under the assumption that $\text{rank}(B) \ll \min(\ell, m)$. Note that under the assumption that the $\mathcal{E}_{i,j}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$, the likelihood is given by

$$\mathcal{L}(\beta|Y,\sigma) \propto \exp\left\{ -\frac{1}{2\sigma^2} \sum_{(i,j)\in I} (Y_{i,j} - \beta_{i,j})^2 \right\}.$$

In [2], this problem is studied without noise (i.e. $\mathcal{E} = 0$), the general case is studied among others in [14, 15, 16].

Note that recently, some authors studied the *trace regression model*, that includes linear regression, reduced-rank regression and matrix completion as special cases: see [17, 18, 5, 19]. Up to our knowledge, this model has not been considered from a Bayesian perspective until now, so we will mainly focus on reduced regression and matrix completion in this paper. However, all the priors defined for reduced-rank regression can also be used for the more general trace regression setting.

2.3 Priors on (Approximately) Low-Rank Matrices

It appears that some econometrics models can actually be seen as special cases of the reduced rank regression. Some of them were studied from a Bayesian perspetive from the seventies, to our knowledge, it was the first Bayesian study of a reduced rank regression:

- incomplete simultaneous equation model: [20, 21, 22, 23],
- cointegration: [24, 25, 26].

The first systematic treatment of the reduced rank model from a Bayesian perspective was carried out in [7]. The idea of this paper is to write the matrix parameter B as $B = MN^T$ for two matrices M and N respectively $p \times k$ and $m \times k$, and to give a prior on M and N rather than on B. Note that the rank of B is in any case smaller than k. So, to choose $k \ll \min(m, p)$ imposes a low rank structure to the matrix B.

The prior in [7] is given by

$$\pi(M, N, \Sigma) = \pi(M, N)\pi(\Sigma)$$

where $\pi(M, N)$ is a Gaussian shrinkage on all the entries of the matrices:

$$\pi(M, N) \propto \exp\left\{-\frac{\tau^2}{2}\left(\|M\|_F^2 + \|N\|_F^2\right)\right\}$$

for some parameter $\tau > 0$. Then, $\pi(\Sigma)$ is an ℓ-dimensional inverse-Wishart distribution with d degrees of freedom and matrix parameter S, $\Sigma^{-1} \sim \mathcal{W}_\ell(d, S)$:

$$\pi(\Sigma) \propto |\Sigma|^{-\frac{m+d+1}{2}} \exp\left(-\frac{1}{2}\mathrm{Tr}(S\Sigma^{-1})\right).$$

Remark that this prior is particularly convenient as it is then possible to give explicit forms for the marginal posteriors. This allows an implementation of the Gibbs algorithm to sample from the posterior. As the formulas are a bit cumbersome, we do not provide them here, however, the interested reader can find them in [7].

The weak point in this approach is that the question of the choice of the reduced rank k is not addressed. It is possible to estimate M and N for any possible k and to use Bayes factors for model selection, as in [26]. Numerical approximation and assessment of convergence for this method are provided by [27].

A more recent approach consists in fixing a large k, as $k = \min(p, m)$, and then in calibrating the prior so that it would naturally favour matrices with rank smaller than k (or, really close to such matrices). To our knowledge, the first attempt in this direction is [8]. Note that this paper was actually about matrix completion rather than reduced rank regression, but once again, all the priors in this subsection can be used in both settings. Here again, we write $B = MN^T$, and

$$\pi(M, N) \propto \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{p}\sum_{j=1}^{k}\frac{M_{i,j}^2}{\sigma_j^2} + \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{N_{i,j}^2}{\rho_j^2}\right)\right\}.$$

In other words, if we write $M = (M_1|\dots|M_k)$ and $N = (N_1|\dots|N_k)$, then the M_j and N_j are independent and respectively $\mathcal{N}_p(0, \sigma_j^2 I_p)$ and $\mathcal{N}_m(0, \rho_j^2 I_m)$ where I_d is the indentity matrix of size d. In order to understand the idea behind this prior, assume for one moment that σ_j^2 and ρ_j^2 are large for $1 \leq j \leq k_0$ and very small for $j > k_0$. Then, for $j > k_0$, M_j and N_j have entries close to 0, and so $M_j N_j^T \simeq 0$. So, the matrix

$$B = MN^T = \sum_{j=1}^{k} M_j N_j^T \simeq \sum_{j=1}^{k_0} M_j N_j^T,$$

a matrix that has a rank at most k_0. In practice, the choice of the σ_j^2's and ρ_j^2's is the main difficulty of this approach. Based on a heuristic, the authors proposed an estimation of these quantities that seems to perform well in practice. Remark that the authors assume that the $\mathcal{E}_{i,j}$ are independent $\mathcal{N}(0, \sigma^2)$ and the parameter σ^2 is not modelled in the prior (but is still estimated on the data). They finally propose a variational Bayes approach to approximate the posterior.

Very similar priors were used by [9] and in the PMF method (Probabilistic Matrix Factorisation) of [10]. However, improved versions were proposed in [28, 29, 30, 31]: the authors proposed a full Bayesian treatment of the problem by putting priors on the hyperparameters. We describe more precisely the prior in [28]: the M_j and N_j are independent and respectively $\mathcal{N}_p(\mu_M, \Sigma_M)$ and $\mathcal{N}_m(\mu_N, \Sigma_N)$, and then: $\mu_M \sim \mathcal{N}_p(\mu_0, \beta_0^{-1}\Sigma_M)$, $\mu_M \sim \mathcal{N}_p(\mu_0, \beta_0^{-1}\Sigma_N)$, and finally $\Sigma_M^{-1}, \Sigma_N^{-1} \sim \mathcal{W}_p(d, S)$. Here again, the hyperparameters β_0, d and S are to be specified. The priors in [29, 30] are quite similar, and we give more details about the one in [30] in Section 3. In [10, 28, 29, 30], the authors simulate from the posterior thanks to the Gibbs sampler (the posterior conditional distribution are explicitely provided e.g. in [28]). Alternatively, [9] uses a stochastic gradient descent to approximate the MAP (maximum a posteriori).

Some papers proposed a kernelized version of the reduced rank regression and matrix completion models. Let M^i denote the i-th row of M and N^h the h-th row of N. Then, $B = MN^T$ leads to $B_{i,h} = M^i(N^h)^T$. We can replace this relation by

$$B_{i,h} = K(M^i, N^h)$$

for some RKHS Kernel K. In [32], the authors propose a Bayesian formulation of this model: B is seen as a Gaussian process on $\{1, \ldots, p\} \times \{1, \ldots, m\}$ with expectation zero and covariance function related to the kernel K. The same idea is refined in [33] and applied successfully to very large datasets, including the NetFlix challenge dataset, thanks to two algorithms: the Gibbs sampler, and the EM algorithm to approximate the MAP.

Finally, we want to mention the nice theoretical work [34, 35]: in these papers, the authors study the asymptotic performance of Bayesian estimators in the reduced rank regression model under a general prior $\pi(M, N)$ that has a compactly supported and infinitely differentiable density. Clearly, the priors aforementioned do not fit the compact support assumption. The question wether algorithmically tractable priors fit this assumption is, to our knowledge, still open. In Section 3, we propose a non-asymptotic analysis of the prior of [30].

3 Theoretical Analysis

In this section, we provide a theoretical analysis of the Bayesian estimators obtained by using the idea of hierarchical priors of [28, 29, 30, 31]. More precisely, we use exactly the prior of [30] and provide a theoretical result on the performance of the estimator in the reduced-rank regression model.

Several approaches are available to study the performance of Bayesian estimators: the asymptotic approach based on Bernstein-von-Mises type theorems, see Chapter 10 in [36], and a non-asmptotic approach based on PAC-Bayesian inequalities. PAC-Bayesian inequalities were introduced for classification by [37, 38] but tighter bounds and extentions to regression estimation can be found in [39, 40, 41, 42, 43]. In all approaches, the variance of the noise is assumed to be known or at least upper-bounded by a given constant, so we use this framework here. To our knowledge, this is the first application of PAC-Bayesian bounds to a matrix estimation problem.

3.1 Theorem

Following [30] we write $B = MN^T$ where M is $p \times k$, N is $m \times k$, $k \leq \min(p, m)$ and then

$$\pi(M, N|\Gamma) \propto \exp\left[-\frac{1}{2}\left(\mathrm{Tr}(M^T \Gamma^{-1} M) + \mathrm{Tr}(N^T \Gamma^{-1} N)\right)\right]$$

for some diagonal matrix

$$\Gamma = \begin{pmatrix} \gamma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \gamma_k \end{pmatrix},$$

the γ_j are i.i.d. and $1/\gamma_j \sim \mathrm{Gamma}(a, b)$:

$$\pi(M, N) = \int \pi(M, N|\Gamma)\pi(\Gamma)\mathrm{d}\Gamma$$

where

$$\pi(\Gamma) = \frac{b^{ka}}{\Gamma(a)^k} \prod_{j=1}^{k} \left\{ \gamma_j^{-a-1} \exp\left(-\frac{b}{\gamma_j}\right) \right\}.$$

We will make one of the following assumptions on the noise:

- **Assumption (A1):** the entries $\mathcal{E}_{i,j}$ of $\mathcal{E}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$, and we know an upper bound s^2 for σ^2.
- **Assumption (A2):** the entries of $\mathcal{E}$ are iid according to any distribution supported by the compact interval $[-\zeta, \zeta]$ with a density f w.r.t. the Lebesgue measure and $f(x) \geq f_{\min} > 0$, and we know an upper bound $s^2 \geq \mathbb{E}(|\mathcal{E}_{1,1}|)/(2 f_{\min})$.

Note that **(A1)** and **(A2)** are special case of the one in [41], the interested reader can replace these assumptions by the more technical condition given in [41]. We define

$$\hat{B}_\lambda = \int MN^T \hat{\rho}_\lambda(\mathrm{d}(M, N))$$

where $\hat{\rho}_\lambda$ is the probability distribution given by

$$\hat{\rho}_\lambda(\mathrm{d}(M, N)) \propto \exp\left(-\lambda \|Y - XMN^T\|_F^2\right) \pi(\mathrm{d}(M, N)).$$

Note that in the case where the entries of $\mathcal{E}$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ then this is the Bayesian posterior, $\hat{\rho}_\lambda(\mathrm{d}(M, N)) = \pi(\mathrm{d}(M, N)|Y)$, when $\lambda = 1/(2\sigma^2)$, and so $\hat{B}_\lambda$ is the expectation under the posterior. However, for theoretical reasons, we have to consider slightly smaller λ to prove theoretical results.

Theorem 1. *Assume that either* **(A1)** *or* **(A2)** *is satisfied. Let us put $a = 1$ and $b = \frac{s^2}{2\ell p k^2(m^2+p^2)}$ in the prior $\pi(\Gamma)$. For $\lambda = \frac{1}{4s^2}$,*

$$\mathbb{E}\left(\|X\hat{B}_\lambda - XB\|_F^2\right) \leq \inf_{\substack{J,M,N \\ M_j,N_j=0 \text{ when } j\notin J}} \Bigg\{ \|X(MN^T - B)\|_F^2 + 6s^2(m+p)|J|\log\left(\frac{1.34\ell p}{s^2}\right) + 8s^2 k \log\left(\frac{22.17\ell p k^2(m^2+p^2)}{s^2}\right) + \frac{2s^2\|X\|_F^2}{\ell p}\left\{\|N\|_F^2 + \|M\|_F^2 + \frac{2s^2}{\ell p} + 16s^2\right\} + 8s^2\left(\|N\|_F^2 + \|M\|_F^2 + \log(2)\right)\Bigg\}.$$

Remark 1. *Note that when all the entries of X satisfy $|X_{i,j}| \leq C$ for some $C > 0$, $\|X\|_F^2/(\ell p) \leq C^2$. Moreover, let us assume that* $\operatorname{rank}(B) = k_0$ *and that we can write $B = MN^T$ with $M_{k_0+1} = \cdots = M_k = 0$ and $N_{k_0+1} = \cdots = N_k = 0$ and $|N_{i,j}|, |M_{i,j}| \leq c$. Assume that the noise is Gaussian. We get*

$$\mathbb{E}\left(\|X\hat{B}_\lambda - XB\|_F^2\right) \leq 50s^2(m+p)k_0\left\{\log(\ell(p\vee m)) + \log\left(\frac{1}{s^2}\vee 1\right) + 1 + C^2(1+c^2+s^2)\right\}$$

where we remind that $p \vee m = \max(p,m)$. When $\operatorname{rank}(X) = p$, *we can see that we recover the same upper bound as in [1], up to a* $\log(\ell(p\vee m))$ *term. This rate (without the* $\log$*) is known to be optimal, see [1] remark (ii) p. 1293 and [17]. However, the presence of the terms $\|M\|_F^2$ and $\|N\|_F^2$ can lead to suboptimal rates in less classical asymptotics where $\|B\|_F$ would grow with the sample size ℓ. In the case of linear regression, a way to avoid these terms is to use heavy-tailed priors as in [41, 42], or compactly supported priors as in [44]. However, it is not clear whether this approach would lead to feasible algorithms in matrix estimation problems. This question will be the object of a future work.*

Remark 2. *We do not claim that the choice $b = \frac{s^2}{2\ell p k^2(m^2+p^2)}$ is optimal in practice. However, from the proof it is clear that our technique requires that b decreases with the dimension of B as well as with the sample size to produce a meaningfull bound. Note that in [30], there is no theoretical approach for the choice of b, but their simulation study tends to show that b must be very small for MN^T to be approximately low-rank.*

Remark 3. *In all the above mentionned papers on PAC-Bayesian bounds, it is assumed that the variance of the noise is known, or upper-bounded by a known constant. More recently, [45] managed to prove PAC-Bayesian inequalities for* *regression with unknown variance. However, the approach is rather involved and it is not clear whether it can be used in our context. This question will also be addressed in a future work.*

3.2 Proof

First, we state the following result:

Theorem 2. *Under* **(A1)** *or* **(A2)***, for any* $\lambda \leq 1/(4s^2)$*, we have*

$$\mathbb{E}\left(\|X\hat{B}_\lambda - XB\|_F^2\right) \leq \inf_\rho \left\{ \int \|X\mu\nu^T - XB\|_F^2 \rho(\mathrm{d}(\mu,\nu)) + \frac{\mathcal{K}(\rho,\pi)}{\lambda} \right\}$$

where $\mathcal{K}(\rho,\pi)$ *stands for the Kullback divergence between* ρ *and* π*,* $\mathcal{K}(\rho,\pi) = \int \log(\frac{\mathrm{d}\rho}{\mathrm{d}\pi})\mathrm{d}\rho$ *if* ρ *is absolutely continuous with respect to* π *and* $\mathcal{K}(\rho,\pi) = \infty$ *otherwise.*

Proof of Theorem 2. Follow the proof of Theorem 1 in [41] and check that every step is valid when B is a matrix instead of a vector. □

We are now ready to prove our main result.

Proof of Theorem 1. Let us introduce, for any $c > 0$, the probability distribution $\rho_{M,N,c}(\mathrm{d}\mu,\mathrm{d}\nu) \propto \mathbf{1}(\|\mu - M\|_F \leq c, \|\nu - N\|_F \leq c)\pi(\mathrm{d}\mu,\mathrm{d}\nu)$. According to Theorem 2 we have

$$\begin{aligned} &\mathbb{E}\left(\|X\hat{B}_\lambda - XB\|_F^2\right) \\ &\qquad \leq \inf_{M,N,c} \left\{ \int \|X\mu\nu^T - XB\|_F^2 \rho_{M,N,c}(\mathrm{d}\mu,\mathrm{d}\nu) + \frac{\mathcal{K}(\rho_{M,N,c},\pi)}{\lambda} \right\}. \end{aligned} \tag{1}$$

Let us fix c, M and N. The remaining steps of the proof are to upper-bound the two terms in the r.h.s. Both upper bounds will depend on c, we will optimize on c after these steps to end the proof. We have

$$\begin{aligned} &\int \|X\mu\nu^T - XB\|_F^2 \rho_{M,N,c}(\mathrm{d}\mu,\mathrm{d}\nu) \\ &\quad = \int \|X\mu\nu^T - XM\nu^T + XM\nu^T - XMN^T \\ &\qquad\qquad + XMN^T - XB\|_F^2 \rho_{M,N,c}(\mathrm{d}\mu,\mathrm{d}\nu) \\ &\quad = \int \Big(\|X\mu\nu^T - XM\nu^T\|_F^2 + \|XM\nu^T - XMN^T\|_F^2 \\ &\qquad\qquad + \|XMN^T - XB\|_F^2 + 2\left\langle X\mu\nu^T - XM\nu^T, XM\nu^T - XMN^T \right\rangle_F \\ &\qquad\qquad + 2\left\langle X\mu\nu^T - XM\nu^T, XMN^T - XB \right\rangle_F \\ &\qquad\qquad + 2\left\langle XM\nu^T - XMN^T, XMN^T - XB \right\rangle_F \Big) \rho_{M,N,c}(\mathrm{d}\mu,\mathrm{d}\nu) \end{aligned}$$

and, as $\int \mu\rho_{M,N,c}(\mathrm{d}\mu) = M$ and $\int \nu\rho_{M,N,c}(\mathrm{d}\nu) = N$, it is easy to see that integral of the three scalar product vanish. So

$$\begin{aligned}
&\int \|X\mu\nu^T - XB\|_F^2 \rho_{M,N,c}(\mathrm{d}\mu,\mathrm{d}\nu) \\
&\quad = \int \left\{\|X\mu\nu^T - XM\nu^T\|_F^2 + \|XM\nu^T - XMN^T\|_F^2\right\} \rho_{M,N,c}(\mathrm{d}\mu,\mathrm{d}\nu) \\
&\qquad\qquad + \|XMN^T - XB\|_F^2 \\
&\quad \le \|X\|_F^2 \int \left\{\|\mu - M\|_F^2\|\nu\|_F^2 + \|M\|_F^2\|\nu - N\|_F^2\right\} \rho_{M,N,c}(\mathrm{d}\mu,\mathrm{d}\nu) \\
&\qquad\qquad + \|X(MN^T - B)\|_F^2 \\
&\quad \le 2c^2\|X\|_F^2 \left\{(\|N\|_F^2 + c^2) + (\|M\|_F^2 + c^2)\right\} + \|X(MN^T - B)\|_F^2. \qquad (2)
\end{aligned}$$

Now, we deal with the second term:

$$\mathcal{K}(\rho_{M,N,c},\pi) = \log \frac{1}{\pi(\{\mu,\nu : \|\mu - M\|_F \le c, \|\nu - N\|_F \le c\})}.$$

We remind that $M = (M_1|\dots|M_k)$ and $N = (N_1|\dots|N_k)$ and let us denote J the subset of $\{1,\dots,k\}$ such that $M_j = N_j = 0$ for $j \notin J$. We let k_0 denote the cardinality of J, $k_0 = |J|$. Note that we have $\mathrm{rank}(MN^T) \le k_0$. For any $\kappa \in (0,1)$ let E_κ be the event

$$\left\{\frac{\kappa}{2} < |\gamma_j| < \kappa \text{ for any } j \notin J \text{ and } |\gamma_j - 1| < \frac{1}{2} \text{ for any } j \in J\right\}.$$

Then

$$\begin{aligned}
\mathcal{K}(\rho_{M,N,c},\pi) &\le \log \frac{1}{\int \pi(\{\mu,\nu : \|\mu - M\|_F \le c, \|\nu - N\|_F \le c\}|\mathit{\Gamma})\pi(\mathit{\Gamma})\mathrm{d}\mathit{\Gamma}} \\
&= \log \frac{1}{\int \pi(\{\|\mu - M\|_F \le c\}|\mathit{\Gamma})\pi(\mathit{\Gamma})\mathrm{d}\mathit{\Gamma}} \\
&\qquad + \log \frac{1}{\int \pi(\{\|\nu - M\|_F \le c\}|\mathit{\Gamma})\pi(\mathit{\Gamma})\mathrm{d}\mathit{\Gamma}} \\
&\le \log \frac{1}{\int_{E_\kappa} \pi(\{\|\mu - M\|_F \le c\}|\mathit{\Gamma})\pi(\mathit{\Gamma})\mathrm{d}\mathit{\Gamma}} \\
&\qquad + \log \frac{1}{\int_{E_\kappa} \pi(\{\|\nu - M\|_F \le c\}|\mathit{\Gamma})\pi(\mathit{\Gamma})\mathrm{d}\mathit{\Gamma}}. \qquad (3)
\end{aligned}$$

By symmetry, we will only bound the first of these two terms. We have

$$\begin{aligned}
&\int_{E_\kappa} \pi(\{\|\mu - M\|_F \le c\}|\mathit{\Gamma})\pi(\mathit{\Gamma})\mathrm{d}\mathit{\Gamma} \\
&\quad = \int_{E_\kappa} \pi\left(\sum_{i=1}^{p}\sum_{j=1}^{k}(\mu_{i,j} - M_{i,j})^2 \le c^2 \,\middle|\, \mathit{\Gamma}\right) \pi(\mathit{\Gamma})\mathrm{d}\mathit{\Gamma}
\end{aligned}$$

$$\geq \int_{E_\kappa} \pi\left(\forall i, \forall j, (\mu_{i,j} - M_{i,j})^2 \leq \frac{c^2}{pk}\middle| \Gamma\right) \pi(\Gamma)\mathrm{d}\Gamma$$
$$= \int_{E_\kappa} \left\{1 - \pi\left(\exists i \in \{1,\dots,p\}, \exists j \notin J, (\mu_{i,j} - M_{i,j})^2 \geq \frac{c^2}{pk}\middle| \Gamma\right)\right\}$$
$$\prod_{i=1}^{p}\prod_{j\in J} \pi\left((\mu_{i,j} - M_{i,j})^2 \leq \frac{c^2}{pk}\middle| \Gamma\right) \pi(\Gamma)\mathrm{d}\Gamma$$
$$\geq \int_{E_\kappa} \left\{1 - \sum_{i=1}^{p}\sum_{j\notin J} \pi\left((\mu_{i,j} - M_{i,j})^2 \geq \frac{c^2}{pk}\middle| \Gamma\right)\right\}$$
$$\prod_{i=1}^{p}\prod_{j\in J} \pi\left((\mu_{i,j} - M_{i,j})^2 \leq \frac{c^2}{pk}\middle| \Gamma\right) \pi(\Gamma)\mathrm{d}\Gamma. \tag{4}$$

We lower-bound the three factors in the integral in (4) separately. First, note that, on E_κ,

$$\pi(\Gamma) = \prod_{j=1}^{k} \frac{b^a}{\Gamma(a)}\gamma_j^{-a-1}\exp\left(-\frac{b}{\gamma_j}\right)$$
$$= \frac{b^{ka}}{\Gamma(a)^k}\left\{\prod_{j\in J}\gamma_j^{-a-1}\exp\left(-\frac{b}{\gamma_j}\right)\right\}\left\{\prod_{j\notin J}\gamma_j^{-a-1}\exp\left(-\frac{b}{\gamma_j}\right)\right\}$$
$$\geq \frac{b^{ka}}{\Gamma(a)^k}\left\{\kappa^{-a-1}\exp\left(-\frac{2b}{\kappa}\right)\right\}^{k-k_0}\left\{\left(\frac{3}{2}\right)^{-a-1}\exp\left(-2b\right)\right\}^{k_0}$$
$$\geq \frac{b^{ka}}{\Gamma(a)^k}\exp\left\{-2b\left(\frac{k-k_0}{\kappa} - k\right)\right\}\left(\frac{3}{2}\right)^{(-a-1)k_0}\kappa^{(-a-1)(k-k_0)}$$
$$\geq \frac{b^{ka}}{\Gamma(a)^k}\left(\frac{2}{3}\right)^{(a+1)k}\exp\left\{\frac{-2bk}{\kappa}\right\}\kappa^{(-a-1)(k-k_0)}. \tag{5}$$

On E_κ, and for $j \notin J$:

$$\pi\left(|\mu_{i,j}| \geq \frac{c}{\sqrt{pk}}\middle| \Gamma\right) = 2\Phi\left(\frac{c}{\sqrt{pk\gamma_j}}\right)$$

where Φ is the c.d.f. of $\mathcal{N}(0,1)$. We use the classical inequality

$$\Phi(x) \leq \frac{\exp\left(-\frac{x^2}{2}\right)}{2}$$

to get:

$$\pi\left(|\mu_{i,j}| \geq \frac{c}{\sqrt{pk}}\middle| \Gamma\right) \leq \exp\left(-\frac{c^2}{2pk\gamma_j}\right) \leq \exp\left(-\frac{c^2}{2pk\kappa}\right)$$

and finally

$$\sum_{i=1}^{p}\sum_{j\notin J}\pi\left((\mu_{i,j}-M_{i,j})^2\geq\frac{c^2}{pk}\middle|\Gamma\right)\leq pk_0\exp\left(-\frac{c^2}{2pk\kappa}\right). \tag{6}$$

Then, on E_κ, and for $j\in J$:

$$\begin{aligned}\pi\left((\mu_{i,j}-M_{i,j})^2\leq\frac{c^2}{pk}\middle|\Gamma\right)&=\pi\left((\mu_{i,j}-M_{i,j})^2\leq\frac{c^2}{pk}\middle|\Gamma\right)\\&=\frac{1}{\sqrt{2\pi\gamma_j}}\int_{M_{i,j}-\frac{c}{\sqrt{pk}}}^{M_{i,j}+\frac{c}{\sqrt{pk}}}\exp\left(-\frac{x^2}{2\gamma_j}\right)\mathrm{d}x\\&\geq c\sqrt{\frac{2}{\pi pk\gamma_j}}\exp\left(-\frac{M_{i,j}^2}{\gamma_j}-\frac{c^2}{pk\gamma_j}\right)\\&\geq c\sqrt{\frac{4}{3\pi pk}}\exp\left(-2M_{i,j}^2-\frac{2c^2}{pk}\right)\end{aligned}$$

and so

$$\begin{aligned}&\prod_{i=1}^{p}\prod_{j\in J}\pi\left((\mu_{i,j}-M_{i,j})^2\leq\frac{c^2}{pk}\middle|\Gamma\right)\\&\qquad\geq\left(c\sqrt{\frac{4}{3\pi pk}}\right)^{pk_0}\exp\left(-2\|M\|_F^2-2c^2\right).\end{aligned} \tag{7}$$

We plug (5), (6) and (7) into (4) and we obtain:

$$\begin{aligned}&\int_{E_\kappa}\pi(\{\|\mu-M\|_F\leq c\}|\Gamma)\pi(\Gamma)\mathrm{d}\Gamma\\&\quad\geq\int_{E_\kappa}\kappa^{(-a-1)(k-k_0)}\frac{b^{ka}}{\Gamma(a)^k}\left(\frac{2}{3}\right)^{(a+1)k}\exp\left\{\frac{-2bk}{\kappa}\right\}\left(c\sqrt{\frac{4}{3\pi pk}}\right)^{pk_0}\\&\qquad\exp\left(-2\|M\|_F^2-2c^2\right)\left(1-pk_0\exp\left(-\frac{c^2}{2pk\kappa}\right)\right)\mathrm{d}\gamma_1\ldots\mathrm{d}\gamma_k\\&\quad=\left(\frac{\kappa}{2}\right)^{k-k_0}\kappa^{(-a-1)(k-k_0)}\frac{b^{ka}}{\Gamma(a)^k}\left(\frac{2}{3}\right)^{(a+1)k}\exp\left\{\frac{-2bk}{\kappa}\right\}\left(c\sqrt{\frac{4}{3\pi pk}}\right)^{pk_0}\\&\qquad\exp\left(-2\|M\|_F^2-2c^2\right)\left(1-pk_0\exp\left(-\frac{c^2}{2pk\kappa}\right)\right).\end{aligned}$$

Now, let us impose the following restrictions: $b=\kappa\leq\frac{c^2}{2pk\log(2pk)}\leq\frac{c^2}{2pk\log(2pk_0)}$ so the last factor is $\geq 1/2$. So we have:

$$\int_{E_\kappa}\pi(\{\|\mu-M\|_F\leq c\}|\Gamma)\pi(\Gamma)\mathrm{d}\Gamma$$

$$\geq \frac{\kappa^{ka}}{\Gamma(a)^k} \frac{2^{ak+1}}{3^{(a+1)^k}} \exp\{-2k\} \left(c\sqrt{\frac{4}{3\pi pk}}\right)^{pk_0} \exp\left(-2\|M\|_F^2 - 2c^2\right).$$

So,

$$\log \frac{1}{\int_{E_\kappa} \pi(\{\|\mu - M\|_F \leq c\}|\Gamma)\pi(\Gamma)\mathrm{d}\Gamma} \leq 2c^2 + 2\|M\|_F^2 + \log(2) + pk_0 \log\left(\frac{1}{c}\sqrt{\frac{3\pi pk}{4}}\right) + k\log\left(\frac{\Gamma(a)3^{a+1}\exp(2)}{\kappa^{a+1}2^a}\right). \quad (8)$$

By symmetry,

$$\log \frac{1}{\int_{E_\kappa} \pi(\{\|\nu - N\|_F \leq c\}|\Gamma)\pi(\Gamma)\mathrm{d}\Gamma} \leq 2c^2 + 2\|N\|_F^2 + \log(2) + mk_0 \log\left(\frac{1}{c}\sqrt{\frac{3\pi pk}{4}}\right) + k\log\left(\frac{\Gamma(a)3^{a+1}\exp(2)}{\kappa^{a+1}2^a}\right), \quad (9)$$

and finally, plugging (8) and (9) into (3)

$$\mathcal{K}(\rho_{M,N,c}, \pi) \leq 4c^2 + 2\|M\|_F^2 + 2\|N\|_F^2 + 2\log(2) + (m+p)k_0 \log\left(\frac{1}{c}\sqrt{\frac{3\pi pk}{4}}\right) + 2k\log\left(\frac{\Gamma(a)3^{a+1}\exp(2)}{\kappa^{a+1}2^a}\right). \quad (10)$$

Finally, we can plug (2) and (10) into (1):

$$\mathbb{E}\left(\|X\hat{B}_\lambda - XB\|_F^2\right) \leq \inf_{\substack{J,M,N,c \\ M_j,N_j=0 \text{ when } j\notin J}} \Bigg\{ 2c^2\|X\|_F^2\left\{\|N\|_F^2 + \|M\|_F^2 + 2c^2\right\} + \|X(MN^T - B)\|_F^2 + \frac{4c^2 + 2\|M\|_F^2 + 2\|N\|_F^2 + 2\log(2)}{\lambda} + \frac{(m+p)|J|\log\left(\frac{1}{c}\sqrt{\frac{3\pi pk}{4}}\right) + 2k\log\left(\frac{\Gamma(a)3^{a+1}\exp(2)}{\kappa^{a+1}2^a}\right)}{\lambda} \Bigg\}.$$

Let us put $c = \sqrt{s^2/\ell p}$ to get:

$$\mathbb{E}\left(\|X\hat{B}_\lambda - XB\|_F^2\right) \leq \inf_{\substack{J,M,N \\ M_j,N_j=0 \text{ when } j\notin J}} \Bigg\{ \|X(MN^T - B)\|_F^2$$

$$+ \frac{(m+p)|J|\log\left(p\sqrt{\frac{\ell k 3\pi}{4s^2}}\right) + 2k\log\left(\frac{\Gamma(a)3^{a+1}\exp(2)}{\kappa^{a+1}2^a}\right)}{\lambda}$$

$$+ \frac{2\|M\|_F^2 + 2\|N\|_F^2 + 2\log(2)}{\lambda} + \frac{2s^2\|X\|_F^2\left\{\|N\|_F^2 + \|M\|_F^2 + \frac{2s^2}{\ell p} + \frac{4}{\lambda}\right\}}{\ell p}\Bigg\}.$$

Finally, remember that the conditions of the theorem impose that $a = 1$, and $b = \frac{s^2}{2\ell p k^2(m^2+p^2)}$. However, we used until now that $b = \kappa$, that $\kappa < 1/2$, that $\kappa \leq c^2/(2pk\log(2pk)) = s^2/(2p^2\ell k\log(2pk))$, and that $\kappa \leq c^2/(2mk\log(2mk)) = s^2/(2mp\ell k\log(2mk))$. Remember that $k \leq \min(p,m)$ so all these equations are compatible. We obtain:

$$\mathbb{E}\left(\|X\hat{B}_\lambda - XB\|_F^2\right) \leq \inf_{\substack{J,M,N \\ M_j,N_j = 0 \text{ when } j\notin J}} \Bigg\{\|X(MN^T - B)\|_F^2$$

$$+ \frac{(m+p)|J|\log\left(p\sqrt{\frac{\ell k 3\pi}{4s^2}}\right) + 2k\log\left(\frac{2\ell p k^2(m^2+p^2)3\exp(2)}{s^2}\right)}{\lambda}$$

$$+ \frac{2\|M\|_F^2 + 2\|N\|_F^2 + 2\log(2)}{\lambda} + \frac{2s^2\|X\|_F^2\left\{\|N\|_F^2 + \|M\|_F^2 + \frac{2g}{\ell p} + \frac{4}{\lambda}\right\}}{\ell p}\Bigg\}.$$

This ends the proof. □

4 Conclusion

We proved that the use of Gaussian priors in reduced-rank regression models leads to nearly optimal rates of convergence. As mentionned in the paper, alternative priors would possibly lead to better bounds but could also result in less computationaly efficient methods (computational efficiency is a major issue when dealing with high-dimensional datasets such as the NetFlix dataset). A complete exploration of this issue will be addressed in future works.

References

[1] Bunea, F., She, Y., Wegkamp, M.H.: Optimal selection of reduced rank estimators of high-dimensional matrices. The Annals of Statistics 39(2), 1282–1309 (2011)

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

[3] Bennett, J., Lanning, S.: The netflix prize. In: Proceedings of KDD Cup and Workshop 2007 (2007)

[4] Reinsel, G.C., Velu, R.P.: Multivariate reduced-rank regression: theory and applications. Springer Lecture Notes in Statistics, vol. 136 (1998)

[5] Koltchinskii, V., Lounici, K., Tsybakov, A.B.: Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. The Annals of Statistics 39(5), 2302–2329 (2011)
[6] Alquier, P., Butucea, C., Hebiri, M., Meziani, K., Morimae, T.: Rank-penalized estimation of a quantum system. Preprint arXiv:1206.1711 (2012)
[7] Geweke, J.: Bayesian reduced rank regression in econometrics. Journal of Econometrics 75, 121–146 (1996)
[8] Lim, Y.J., Teh, Y.W.: Variational Bayesian approach to movie rating prediction. In: Proceedings of KDD Cup and Workshop 2007 (2007)
[9] Lawrence, N.D., Urtasun, R.: Non-linear matrix factorization with Gaussian processes. In: Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, pp. 601–608. ACM, New York (2009)
[10] Salakhutdinov, R., Mnih, A.: Bayesian probabilistic matrix factorization. In: Platt, J.C., Koller, D., Singer, Y., Roweis, S. (eds.) Advances in Neural Information Processing Systems 20, NIPS 2007. MIT Press, Cambridge (2008)
[11] Anderson, T.: Estimating linear restrictions on regression coefficients for multivariate normal distributions. Annals of Mathematical Statistics 22, 327–351 (1951)
[12] Izenman, A.: Reduced rank regression for the multivariate linear model. Journal of Multivariate Analysis 5(2), 248–264 (1975)
[13] Yuan, M., Ekici, A., Lu, Z., Monteiro, R.: Dimension reduction and coefficient estimation in multivariate linear regression. Journal of the Royal Statistical Society - Series B 69, 329–346 (2007)
[14] Candès, E., Plan, Y.: Matrix completion with noise. Proceedings of the IEEE 98(6), 625–636 (2009)
[15] Candès, E., Recht, B.: Exact matrix completion via convex optimization. Foundations of Computational Mathematics 9(6), 717–772 (2009)
[16] Gross, D.: Recovering low-rank matrices from few coefficients in any basis. IEEE Transactions on Information Theory 57, 1548–1566 (2011)
[17] Rohde, A., Tsybakov, A.B.: Estimation of high-dimensional low-rank matrices. The Annals of Statistics 39, 887–930 (2011)
[18] Klopp, O.: Rank-penalized estimators for high-dimensionnal matrices. Electronic Journal of Statistics 5, 1161–1183 (2011)
[19] Koltchinskii, V.: Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. Springer Lecture Notes in Mathematics (2011)
[20] Dreze, J.H.: Bayesian limited information analysis of the simultaneous equation model. Econometrica 44, 1045–1075 (1976)
[21] Dreze, J.H., Richard, J.F.: Bayesian analysis of simultaneous equation models. In: Griliches, Z., Intriligater, J.F. (eds.) Handbook of Econometrics, vol. 1. North-Holland, Amsterdam (1983)
[22] Zellner, A., Min, C., Dallaire, D.: Bayesian analysis of simultaenous equation and related models using the Gibbs sampler and convergence checks. H. G. B. Alexander Research Founsation working paper, University of Chicago (1993)
[23] Kleibergen, F., van Dijk, H.K.: Bayesian simultaneous equation analysis using reduced rank structures. Econometric Theory 14, 699–744 (1998)
[24] Bauwens, L., Lubrano, M.: Identification restriction and posterior densities in cointegrated gaussian var systems. In: Fomby, T.M., Carter Hill, R. (eds.) Advances in Econometrics, vol. 11(B). JAI Press, Greenwich (1993)
[25] Kleibergen, F., van Dijk, H.K.: On the shape of the likelihood-posterior in cointegration models. Econometric Theory 10, 514–551 (1994)
[26] Kleibergen, F., Paap, R.: Priors, posteriors and Bayes factors for a Bayesian analysis of cointegration. Journal of Econometrics 111, 223–249 (2002)

[27] Corander, J., Villani, M.: Bayesian assessment of dimensionality in reduced rank regression. Statistica Neerlandica 58(3), 255–270 (2004)
[28] Salakhutdinov, R., Mnih, A.: Bayesian probabilistic matrix factorization using markov chain monte carlo. In: Proceedings of the 25th Annual International Conference on Machine Learning, ICML 2008. ACM, New York (2008)
[29] Zhou, M., Wang, C., Chen, M., Paisley, J., Dunson, D., Carin, L.: Nonparametric Bayesian matrix completion. In: IEEE Sensor Array and Multichannel Signal Processing Workshop (2010)
[30] Babacan, S.D., Luessi, M., Molina, R., Katsaggelos, A.K.: Low-rank matrix completion by variational sparse Bayesian learning. In: IEEE International Conference on Audio, Speech and Signal Processing, Prague (Czech Republic), pp. 2188–2191 (2011)
[31] Paisley, J., Carin, L.: A nonparametric Bayesian model for kernel matrix completion. In: Proceedings of ICASSP 2010, Dallas, USA (2010)
[32] Yu, K., Tresp, V., Schwaighofer, A.: Learning Gaussian processes for multiple tasks. In: Proceedings of the 22th Annual International Conference on Machine Learning, ICML 2005 (2005)
[33] Yu, K., Lafferty, J., Zhu, S., Gong, Y.: Large-scale collaborative prediction using a non-parametric random effects model. In: Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009. ACM, New York (2009)
[34] Aoyagi, M., Watanabe, S.: The generalization error of reduced rank regression in Bayesian estimation. In: International Symposium on Information Theory and its Applications, ISITA 2004, Parma, Italy (2004)
[35] Aoyagi, M., Watanabe, S.: Stochastic complexities of reduced rank regression in Bayesian estimation. Neural Networks 18, 924–933 (2005)
[36] van der Vaart, A.W.: Asymptotic Statistics. Cambridge University Press (1998)
[37] Shawe-Taylor, J., Williamson, R.: A PAC analysis of a Bayes estimator. In: Proceedings of the Tenth Annual Conference on Computational Learning Theory, pp. 2–9. ACM, New York (1997)
[38] McAllester, D.A.: Some pac-bayesian theorems. In: Proceedings of the Eleventh Annual Conference on Computational Learning Theory, Madison, WI, pp. 230–234. ACM (1998)
[39] Catoni, O.: Statistical Learning Theory and Stochastic Optimization. Springer Lecture Notes in Mathematics (2004)
[40] Catoni, O.: PAC-Bayesian Supervised Classification (The Thermodynamics of Statistical Learning). Lecture Notes-Monograph Series, vol. 56. IMS (2007)
[41] Dalalyan, A.S., Tsybakov, A.B.: Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. Machine Learning 72, 39–61 (2008)
[42] Dalalyan, A.S., Tsybakov, A.B.: Sparse regression learning by aggregation and Langevin Monte-Carlo. J. Comput. System Sci. 78(5), 1423–1443 (2012)
[43] Dalalyan, A.S., Salmon, J.: Sharp oracle inequalities for aggregation of affine estimators. The Annals of Statistics 40(4), 2327–2355 (2012)
[44] Alquier, P., Lounici, K.: PAC-Bayesian bounds for sparse regression estimation with exponential weights. Electronic Journal of Statistics 5, 127–145 (2011)
[45] Audibert, J.Y., Catoni, O.: Robust linear least squares regression. The Annals of Statistics 39, 2766–2794 (2011)