1-Bit matrix completion: PAC-Bayesian analysis of a variational approximation

Vincent Cottet¹ · Pierre Alquier¹

Received: 14 April 2016 / Accepted: 14 August 2017 / Published online: 19 September 2017
© The Author(s) 2017

Abstract We focus on the completion of a (possibly) low-rank matrix with binary entries, the so-called 1-bit matrix completion problem. Our approach relies on tools from machine learning theory: empirical risk minimization and its convex relaxations. We propose an algorithm to compute a variational approximation of the pseudo-posterior. Thanks to the convex relaxation, the corresponding minimization problem is bi-convex, and thus the method works well in practice. We study the performance of this variational approximation through PAC-Bayesian learning bounds. Contrary to previous works that focused on upper bounds on the estimation error of M with various matrix norms, we are able to derive from this analysis a PAC bound on the prediction error of our algorithm. We focus essentially on convex relaxation through the hinge loss, for which we present a complete analysis, a complete simulation study and a test on the MovieLens data set. We also discuss a variational approximation to deal with the logistic loss.

Keywords Matrix completion · PAC-Bayesian bounds · Variational Bayes · Supervised classification · Risk convexification · Oracle inequalities

1 Introduction

Motivated by modern applications like recommendation systems and collaborative filtering, video analysis or quantum statistics, the matrix completion problem has been widely studied

Editor: Peter Auer.

P. Alquier gratefully acknowledges financial support from the research programme *New Challenges for New Data* from LCL and GENES, hosted by the *Fondation du Risque*, and from Labex ECODEC (ANR-11-LABEX-0047).

✉ Pierre Alquier
pierre.alquier@ensae.fr

¹ CREST, ENSAE, Université Paris Saclay, Paris, France

over the recent years. Recovering a matrix is, without any additional information, an impossible task. However, under some assumptions on the structure of the matrix to be recovered, it might become feasible, as shown by Candès and Tao (2010) and Candès and Recht (2012) where the assumption is that the matrix has a small rank. This assumption is natural in many applications. For example, in recommendation systems, it is equivalent to the existence of a small number of hidden features that explain the users' preferences. While Candès and Tao (2010) and Candès and Recht (2012) focused on matrix completion without noise, many authors extended these techniques to the case of noisy observations, see Candès and Plan (2010) and Chatterjee (2015) among others. The main idea in Candès and Plan (2010) is to minimize the least squares criterion, penalized by the rank. This penalization is then relaxed by the nuclear norm, which is the sum of the singular values of the matrix at hand. An efficient algorithm is described in Recht and Ré (2013).

All the aforementioned papers focused on real-valued matrices. However, in many applications, the matrix entries are binary, that is in the set $\{0, 1\}$. For example, in collaborative filtering, we have often only access to a binary choice: the (i, j) th entry being 1 means that user i is satisfied by object j while this entry being 0 means that he/she is not satisfied by it. The problem of recovering a binary matrix from partial observations is usually referred to as 1-bit matrix completion. To deal with binary observations requires specific estimation methods. Most works on this problem usually assume a generalized linear model (GLM): the observations Y_{ij} for $1 \leq i \leq m_1$, $1 \leq j \leq m_2$, are Bernoulli distributed with parameter $f(M_{ij})$, where f is a link function which maps from $\mathbb{R}$ to $[0, 1]$, for example the logistic function $f(x) = \exp(x)/[1 + \exp(x)]$, and M is a $m_1 \times m_2$ real matrix, see Cai and Zhou (2013), Davenport et al. (2014) and Klopp et al. (2015). In these works, the goal is to recover the matrix M and a convergence rate is then derived. For example, Klopp et al. (2015) provides an estimate $\hat{M}$ for which, under suitable assumptions and when the data are generated according to the true model with $M = M_0$,

$$\frac{1}{m_1 m_2} \|\hat{M} - M_0\|_F^2 \leq C \max \left(\sqrt{\frac{\log(m_1 + m_2)}{n}}, \frac{\max(m_1, m_2) \text{rank}(M_0) \log(m_1 + m_2)}{n} \right)$$

for some constant C that depends on the assumptions and the sampling scheme, and where $\|\cdot\|_F$ stands for the Frobenius norm [we refer the reader to Corollary 2 page 2955 in Klopp et al. (2015) for the exact statement]. While this result ensures the consistency of $\hat{M}$ when M_0 is low-rank, it does not provide any guarantee on the probability of a prediction error. Moreover, the results rely on the assumption that the model (in particular the function f) is well specified. In practice, this assumption is unrealistic, and it is important to provide generalization error bounds that hold even in case of misspecification.

Here, we adopt a machine learning point of view: in machine learning, dealing with binary output is called a classification problem, for which methods are known that do not assume any model on the observations. That is, instead of focusing on a parametric model for $Y_{i,j}$, we will only define a set of prediction matrices M and seek for the one that leads to the best prediction error. Using the zero-one loss function, we could actually directly use Vapnik (1998) theory to propose a classifier $\hat{M}$ risk would be controlled by a PAC inequality. However, it is known that this approach usually is computationally intractable. A popular approach is to replace the zero-one loss by a convex surrogate (Zhang 2004), namely, the hinge loss. Our approach is as follows: we propose a pseudo-Bayesian approach, where we define a pseudo-posterior distribution on a set of matrices M . This pseudo-posterior distribution does not have a simple form, however, thanks to a variational approximation, we manage to approximate it by a tractable distribution. Thanks to the PAC-Bayesian theory

(McAllester 1998; Herbrich and Graepel 2002; Shawe-Taylor and Langford 2003; Catoni 2004, 2007; Seldin et al. 2012; Dalalyan and Tsybakov 2008), we are able to provide a PAC bound on the prediction risk of this variational approximation. We then show that, due to the convex relaxation of the zero-one loss, the computation of this variational approximation is actually a bi-convex minimization problem. As a consequence, efficient algorithms are available.

Other settings for 1-bit matrix completion have also been studied. For example, in some real-life applications, only positive instances are available. This setting is studied in details in Hsieh et al. (2015). It requires a different approach. Here, we stick to the classification approach where positive and negative instances are observed. We refer the reader to Hsieh et al. (2015) and the references therein for the positive-only case.

The rest of the paper is as follows. In Sect. 2 we provide the notations used in the paper, the definition of the pseudo-posterior and of its variational approximation. In Sect. 3 we give the PAC analysis of the variational approximation. This yields an empirical and a theoretical upper bound on the prediction risk of our method. Sect. 4 provides details on the implementation of our method. Note that in the aforementioned sections, the convex surrogate of the zero-one loss used is the hinge loss. An extension to the logistic loss is briefly discussed in Sect. 5, together with an algorithm to compute the variational approximation. Finally, Sect. 6 is devoted to an empirical study and Sect. 6.3 to an application to the MovieLens data set. The proof of the theorems of Sect. 3 are provided in Sect. 1.

2 Estimation procedure

For any integer m we define $[m] = \{1, \dots, m\}$; for two real numbers a and b we write $\max(a, b) = a \vee b$ and $\min(a, b) = a \wedge b$. We define, for any integers m_1 and m_2 and any matrix $M \in \mathbb{R}^{m_1 \times m_2}$, $\|M\|_{\max} = \max_{(i,j) \in [m_1] \times [m_2]} M_{ij}$. Let $\mathbb{R}^+$ stand for the set of non-negative real numbers, and $\mathbb{R}^{+*}$ for the positive real numbers. For any real number a , $(a)_+$ is the positive part of a and is equal to $\max(0, a)$.

For a pair of matrices (A, B) , we write $\ell(A, B) = \|A\|_{\max} \vee \|B\|_{\max}$. Finally, when an $m_1 \times m_2$ matrix M has $\text{rank}(M) = r$ then it can be written as $M = LR^T$ where L is $m_1 \times r$ and R is $m_2 \times r$. This decomposition is obviously not unique; we put $\ell(M) = \inf_{(L,R)} \ell(L, R)$ where the infimum is taken over all such possible pairs (L, R) such that $LR^T = M$. In frequentist approaches like Klopp et al. (2015), it is common that the upper bound depends on the infinite norm of the entries. This quantity is replaced in our analysis by $\ell(M)$.

2.1 1-Bit matrix completion as a classification problem

We formally describe the 1-bit matrix completion problem as a classification problem: we observe $(X_k, Y_k)_{k \in [n]}$ that are n i.i.d pairs from a distribution $\mathbf{P}$. The X_k 's take values in $\mathcal{X} = [m_1] \times [m_2]$ and the Y_k 's take values in $\mathcal{Y} = \{-1, +1\}$. Hence, the k -th observation of an entry of the matrix is Y_k and the corresponding position in the matrix is provided by $X_k = (i_k, j_k)$. In this setting, a predictor is a function $[m_1] \times [m_2] \rightarrow \mathbb{R}$ and thus can be represented by a matrix M and for any X , M_X is the entry of M at location X . It is natural to use M in the following way: when $(X, Y) \sim \mathbf{P}$, M predicts Y by $\text{sign}(M_X)$. The ability of this predictor to predict a new entry of the matrix is then assessed by the risk

$$\mathbf{R}(M) = \mathbb{E}_{\mathbf{P}} [1(YM_X < 0)],$$

and its empirical counterpart is:

$$r_n(M) = \frac{1}{n} \sum_{k=1}^n \mathbb{1}(Y_k M_{X_k} < 0) = \frac{1}{n} \sum_{k=1}^n \mathbb{1}(Y_k M_{i_k, j_k} < 0).$$

It is then possible to use the standard approach in classification theory (Vapnik 1998). For example, the best possible classifier is the Bayes classifier and it relies on the regression function:

$$\eta(x) = \mathbb{E}(Y|X = x) \quad \text{or equivalently} \quad \eta(i, j) = \mathbb{E}[Y|X = (i, j)],$$

and therefore we have an optimal matrix

$$M_{ij}^B = \text{sign}[\eta(i, j)] = \text{sign}\{\mathbb{E}[Y|X = (i, j)]\}.$$

We define $\overline{\mathbf{R}} = \inf_M \mathbf{R}(M) = \mathbf{R}(M^B)$, and $\overline{r}_n = r_n(M^B)$. Note that, clearly, if two matrices M^1 and M^2 are such as, for every (i, j) , $\text{sign}(M_{ij}^1) = \text{sign}(M_{ij}^2)$ then $\mathbf{R}(M^1) = \mathbf{R}(M^2)$, and obviously,

$$\forall M, \forall (i, j) \in [m_1] \times [m_2], \quad \text{sign}(M_{ij}) = M_{ij}^B \Rightarrow r_n(M) = \overline{r}_n.$$

While the risk $\mathbf{R}(M)$ has a clear interpretation, its empirical counterpart $r_n(M)$ usually leads to intractable problems, as it is non-smooth and non-convex. Hence, it is standard to replace the empirical risk by a convex surrogate (Zhang 2004). In this paper, we will mainly deal with the hinge loss, which leads to the following so-called hinge risk and hinge empirical risk:

$$\begin{aligned} R^h(M) &= \mathbb{E}_{\mathbf{P}}[(1 - Y M_X)_+], \\ r_n^h(M) &= \frac{1}{n} \sum_{k=1}^n (1 - Y_k M_{X_k})_+. \end{aligned}$$

The hinge loss was also used by Srebro et al. (2004) and in Herbster et al. (2016) in the 1-bit matrix completion problem, with a different approach leading to different algorithms. Moreover, here, we provide an analysis of the rate of convergence of our method, that is not provided in Srebro et al. (2004). Note that our analysis can be extended to any Lipschitz and convex surrogate, and indeed, we study also briefly the logistic loss in Sect. 5. Still, we prefer to focus only on the hinge loss in the main part of the paper, for its good algorithmic and theoretical properties, c.f. Zhang (2004).

Contrary to many recent papers on matrix completion, our approach leads to distribution-free bounds. The marginal distribution of X is not an issue and we do not have to assume a uniform sampling scheme. Following standard notations in matrix completion, we define Ω as the set of indices of observed entries: $\Omega = \{X_1, \dots, X_n\}$. We will use in the following the sub-sample of $\{1, \dots, n\}$ for a specified line i : $\Omega_{i,\cdot} = \{l \in [n] : (i, j_l) \in \Omega\}$ and the counterpart for a specified column j : $\Omega_{\cdot,j} = \{l \in [n] : (i_l, j) \in \Omega\}$.

2.2 Pseudo-Bayesian estimation

The Bayesian framework has been used several times for matrix completion [see Salakhutdinov and Mnih (2008) and Lim and Teh (2007) and the references therein]. The PAC-Bayesian approach has been well used on different models [see Mai and Alquier (2015) and Sect. 6 in Seldin and Tishby (2010)]. A common idea in all of these papers is to factorize the matrix

into two parts in order to define a prior on low-rank matrices. It needs an additional parameter and a hierarchical model in order to be rank-adaptive and we explain here the idea. Every matrix whose rank is r can be factorized:

$$M = LR^\top, L \in \mathbb{R}^{m_1 \times r}, \quad R \in \mathbb{R}^{m_2 \times r}.$$

As mentioned in the introduction, the Bayes matrix M^B is expected to be low-rank, or at least well approximated by a low-rank matrix. However, in practice, we do not know what would be the rank of this matrix. So, we actually write $M = LR^\top$ with $L \in \mathbb{R}^{m_1 \times K}$, $R \in \mathbb{R}^{m_2 \times K}$ for some large enough K . Adaptation with respect to $r \in [K]$ is obtained by shrinking some columns of L and R to 0. In order to do so, we will scale parameters γ_k for the columns of L and R , and let $\gamma := (\gamma_1, \dots, \gamma_K)$. We then define the following hierarchical probability distribution:

$$\forall k \in [K], \quad \gamma_k \stackrel{iid}{\sim} \pi^\gamma, \quad (1)$$

$$\forall (i, j, k) \in [m_1] \times [m_2] \times [K], \quad L_{i,k}, R_{j,k} | \gamma \stackrel{indep.}{\sim} \mathcal{N}(0, \gamma_k), \quad (2)$$

$$\text{and } M = LR^\top, \quad (3)$$

where the prior distribution on the variances π^γ is yet to be specified. It means that the entries of L and R are normally distributed but the variance depends on the column index: a large γ_k leads to spread values and a small γ_k leads to almost null entries of the column k . In most papers π^γ is chosen as an inverse-Gamma distribution because it is conjugate in this model. This kind of hierarchical prior distribution is also very similar to the Bayesian Lasso developed in [Park and Casella \(2008\)](#) and especially of the form of the Bayesian Group Lasso developed in [Kyung et al. \(2010\)](#) in which the variance term is Gamma distributed. We will show that the Gamma distribution is a possible alternative in matrix completion, both for theoretical results and practical considerations. Thus all the results in this paper are stated under the assumption that π^γ is either the Gamma or the inverse-Gamma distribution: $\pi^\gamma = \Gamma(\alpha, \beta)$, or $\pi^\gamma = \Gamma^{-1}(\alpha, \beta)$.

Let θ denote the parameter $\theta = (L, R, \gamma)$ and π denote the prior distribution defined in (1). Following the aforementioned papers in PAC-Bayesian theory, we define the pseudo-posterior as follows:

$$\hat{\rho}_\lambda(d\theta) = \frac{\exp[-\lambda r_n^h(LR^\top)]}{\int \exp[-\lambda r_n^h] d\pi} \pi(d\theta)$$

where $\lambda > 0$ is a parameter to be fixed by the statistician. The calibration of λ is discussed below. This distribution is close to a classic posterior distribution but the likelihood has been replaced by the pseudo-likelihood $\exp[-\lambda r_n^h(LR^\top)]$ based on the hinge empirical risk.

2.3 Variational Bayes approximations

Unfortunately, the pseudo-posterior is intractable and MCMC methods may be too expensive because of the dimension of the parameter. We decide to use a Variational Bayes approximation, that is to seek an approximation of the pseudo-posterior by efficient optimization algorithms ([Bishop 2006](#)). First, we fix a subset $\mathcal{F}$ of the set of all distributions on the parameter space. The class $\mathcal{F}$ should be large enough to contain a good enough approximation of $\hat{\rho}_\lambda$, but not too large, in order to keep optimization in $\mathcal{F}$ feasible. The VB approximation is

then defined by

$$\arg \min_{\rho \in \mathcal{F}} \mathcal{H}(\rho, \hat{\rho}_\lambda), \text{ where } \mathcal{H}(\rho, \hat{\rho}_\lambda) = \int \log \left(\frac{d\rho}{d\hat{\rho}_\lambda} \right) d\rho$$

is the Kullback–Leibler divergence between ρ and $\hat{\rho}_\lambda$. When $\mathcal{H}(\rho, \hat{\rho}_\lambda)$ is not available in closed form, it is usual to replace it by an upper bound. Following the classical approach with matrix factorization priors, as in [Lim and Teh \(2007\)](#), we define here the class $\mathcal{F}$ as follows:

$$\mathcal{F} = \left\{ \rho(dL, dR, d\gamma) = \prod_{k=1}^K \left[\prod_{i=1}^{m_1} \varphi(L_{i,k}; L_{i,k}^0, v_{i,k}^L) dL_{i,k} \prod_{j=1}^{m_2} \varphi(R_{j,k}; R_{j,k}^0, v_{j,k}^R) dR_{j,k} \rho^{\gamma_k}(d\gamma_k) \right], \right. \\ \left. L^0 \in \mathbb{R}^{m_1 \times K}, R^0 \in \mathbb{R}^{m_2 \times K}, v^L \in \mathbb{R}_+^{m_1 \times K}, v^R \in \mathbb{R}_+^{m_2 \times K} \right\},$$

where $\varphi(\cdot; \mu, v)$ is the density of the Gaussian distribution with parameters (μ, v) and ρ^{γ_k} ranges over all possible probability distributions for $\gamma_k \in \mathbb{R}^+$. The VB approximations are referred as *parametric* when $\mathcal{F}$ is finite dimensional and as *mean-field* otherwise. Here we actually use a mixed approach. Informally, under $\rho \in \mathcal{F}$, all the coordinates are independent and the variational distribution of the entries of L and R is specified. The free variational parameters to be optimized are the means and the variances. We will show below that the optimization with respect to ρ^{γ_k} is available in close form. Also, note that any probability distribution $\rho \in \mathcal{F}$ is uniquely determined by L^0, R^0, v^L, v^R and $\rho^{\gamma_1}, \dots, \rho^{\gamma_K}$. We could actually use the notation $\rho = \rho_{L^0, R^0, v^L, v^R, \rho^{\gamma_1}, \dots, \rho^{\gamma_K}}$, but it would be too cumbersome, so we will avoid it as much as possible. Conversely, once ρ is given in $\mathcal{F}$, we can define $L^0 = \mathbb{E}_\rho[L]$, $R^0 = \mathbb{E}_\rho[R]$ and so on.

It is well-known that the Kullback divergence can be decomposed as

$$\mathcal{H}(\rho, \hat{\rho}_\lambda) = \lambda \int r_n^h d\rho + \mathcal{H}(\rho, \pi) + \log \int \exp[-\lambda r_n^h] d\pi \quad (4)$$

but the first term in the right-hand side is not tractable here. We then use the Lipschitz property of the loss and derive an upper bound of the Kullback divergence for any $\rho \in \mathcal{F}$ which is explicit in the parameters of ρ . It is this quantity that we will optimize in the algorithm and we will see in the next section that this estimate enjoys good properties. We remind the reader that all the proofs are postponed to Sect. 1.

Proposition 1 For any $\rho = \rho_{L^0, R^0, v^L, v^R, \rho^{\gamma_1}, \dots, \rho^{\gamma_K}} \in \mathcal{F}$,

$$\int r_n^h d\rho + \frac{1}{\lambda} \mathcal{H}(\rho, \pi) \leq r_n^h \left(\mathbb{E}_\rho[L] \mathbb{E}_\rho[R]^\top \right) + \mathcal{R}(\rho, \lambda) \quad (5)$$

where

$$\mathcal{R}(\rho, \lambda) = \frac{1}{n} \sum_{h=1}^n \sum_{k=1}^K \left[\sqrt{v_{i_h,k}^L \frac{2}{\pi}} \sqrt{v_{j_h,k}^R \frac{2}{\pi}} + |R_{j_h,k}^0| \sqrt{v_{i_h,k}^L \frac{2}{\pi}} + |L_{i_h,k}^0| \sqrt{v_{j_h,k}^R \frac{2}{\pi}} \right] \\ + \frac{1}{\lambda} \left\{ \frac{1}{2} \sum_{k=1}^K \mathbb{E}_\rho \left[\frac{1}{\gamma_k} \right] \left(\sum_{i=1}^{m_1} (v_{i,k}^L + (L_{i,k}^0)^2) + \sum_{j=1}^{m_2} (v_{j,k}^R + (R_{j,k}^0)^2) \right) \right\}$$

$$-\frac{1}{2} \sum_{k=1}^K \left(\sum_{i=1}^{m_1} \log v_{i,k}^L + \sum_{j=1}^{m_2} \log v_{j,k}^R \right) + \sum_{k=1}^K [\mathcal{H}(\rho^{\gamma_k}, \pi^{\gamma_k}) + \frac{m_1 + m_2}{2} (\mathbb{E}_{\rho} [\log \gamma_k] - 1)] \Big\}.$$

Note that the explicit expression of our upper bound $\mathcal{R}(\rho, \lambda)$ is very cumbersome to say the least. A few comments are in order. First, this upper bound is explicit and can be computed easily. Hence, instead of minimizing the Kullback divergence, this is this term that we minimize in practice. Then, our theoretical analysis will show that the upper bound is acceptable in the sense that its minimization will lead to a small generalization error. But it is actually possible to understand at first sight why the minimization of $r_n^h(\mathbb{E}_{\rho}[L]\mathbb{E}_{\rho}[R]^{\top}) + \mathcal{R}(\rho, \lambda)$ works well. Indeed, assume that a matrix M with $r_n^h(M) = 0$ satisfies $\text{rank}(M) = r \ll K$. Then, it is possible to decompose M as a product $M = L^0(R^0)^{\top}$ with $L_{i,k}^0 = R_{j,k}^0 = 0$ when $r < k \leq K$. So, the sum

$$\frac{1}{2\lambda} \sum_{k=1}^K \mathbb{E}_{\rho} \left[\frac{1}{\gamma_k} \right] \left(\sum_{i=1}^{m_1} (L_{i,k}^0)^2 + \sum_{j=1}^{m_2} (R_{j,k}^0)^2 \right)$$

has actually only $r(m_1 + m_2) = \text{rank}(M)(m_1 + m_2)$ non-null terms. To minimize $r_n^h(\mathbb{E}_{\rho}[L]\mathbb{E}_{\rho}[R]^{\top}) + \mathcal{R}(\rho, \lambda)$ is thus related to penalized risk minimization with a penalty proportional to the rank, as in most frequentist approaches (Klopp et al. 2015).

The quantity $r_n^h(\mathbb{E}_{\rho}[L]\mathbb{E}_{\rho}[R]^{\top}) + \mathcal{R}(\rho, \lambda)$ will be referred as the Approximate Variational Bound (AVB) of ρ in the following. We are now able to define our estimate.

Definition 1 For a fixed $\lambda > 0$ we put

$$\begin{aligned} AVB(\rho, \lambda) &= r_n^h(\mathbb{E}_{\rho}[L]\mathbb{E}_{\rho}[R]^{\top}) + \mathcal{R}(\rho, \lambda), \\ \tilde{\rho}_{\lambda} &= \arg \min_{\rho \in \mathcal{F}} AVB(\rho, \lambda). \end{aligned} \quad (6)$$

Also, when explicit notations involving $L^0, R^0, v^L, v^R, \rho^{\gamma_1}, \dots, \rho^{\gamma_k}$ are necessary we will use the notation

$$AVB(L^0, R^0, v^L, v^R, \rho^{\gamma_1}, \dots, \rho^{\gamma_k}, \lambda) = AVB(\rho_{L^0, R^0, v^L, v^R, \rho^{\gamma_1}, \dots, \rho^{\gamma_k}}, \lambda).$$

In the next section, we study the theoretical properties of our estimate. The main result is that the minimizer $\tilde{\rho}_{\lambda}$ of the $AVB(\rho, \lambda)$ has a small prediction risk for a well chosen λ . We also provide an algorithm that computes $\tilde{\rho}_{\lambda}$ and show on simulations that it behaves well in practice.

3 PAC analysis of the variational approximation

Alquier et al. (2015) propose a general framework for analyzing the prediction properties of VB approximations of pseudo-posteriors based on PAC-Bayesian bounds. In this section, we apply this method to derive a control of the out-of-sample prediction risk $\mathbf{R}$ for our approximation $\tilde{\rho}_{\lambda}$.

3.1 Empirical bound

The first result is a so-called empirical bound: it provides an upper bound on the prediction risk of the pseudo-posterior $\tilde{\rho}_\lambda$ that depends only on the data and on quantities defined by the statistician.

Lemma 1 *For any $\epsilon \in (0, 1)$, with probability at least $1 - \epsilon$ on the drawing of the sample, for any $\rho \in \mathcal{F}$,*

$$\int \mathbf{R} d\rho \leq r_n^h \left(\mathbb{E}_\rho[L] \mathbb{E}_\rho[R]^\top \right) + \mathcal{R}(\rho, \lambda) + \frac{\lambda}{2n} + \frac{\log \frac{1}{\epsilon}}{\lambda} = AVB(\rho, \lambda) + \frac{\lambda}{2n} + \frac{\log \frac{1}{\epsilon}}{\lambda}.$$

This shows that our strategy to minimize $AVB(\rho, \lambda)$ is indeed the minimization of an empirical upper bound on the prediction risk, a standard approach in PAC-Bayesian theory. An immediate consequence of Lemma 1 and of the definition of $\tilde{\rho}_\lambda$ is the following theorem.

Theorem 1 *For any $\epsilon \in (0, 1)$, with probability at least $1 - \epsilon$ on the drawing of the sample,*

$$\int \mathbf{R} d\tilde{\rho}_\lambda \leq \inf_{\rho \in \mathcal{F}} AVB(\rho, \lambda) + \frac{\lambda}{2n} + \frac{\log \frac{1}{\epsilon}}{\lambda}$$

Even though the bound in the right-hand side may be evaluated in practice, and thus may provide a numerical guarantee on the out-of-sample prediction risk, it is not very clear how it depends on the parameters. The following corollary of Theorem 1 will clarify things. It is obtained by deriving upper bounds of $AVB(\rho, \lambda)$ (once again, the proof is provided explicitly in Sect. 1).

Corollary 1 *Assume that $\lambda \leq n$. For any $\epsilon \in (0, 1)$, with probability at least $1 - \epsilon$:*

$$\int \mathbf{R} d\tilde{\rho}_\lambda \leq \inf_M \left[r_n^h(M) + \mathcal{C}_{\pi^\gamma} \frac{\text{rank}(M)(m_1 + m_2)[\log n + \ell^2(M)]}{\lambda} \right] + \frac{\lambda}{2n} + \frac{\log \frac{1}{\epsilon}}{\lambda}$$

where the constant $\mathcal{C}_{\pi^\gamma}$ is explicitly known, and depends only on the form of prior π^γ (Gamma, or Inverse-Gamma) and of its hyperparameters.

An exact value for $\mathcal{C}_{\pi^\gamma}$ can be deduced from the proof. It is thus clear that the algorithm performs a trade-off between the fit to the data, through the term $r_n^h(M)$, and the rank of M .

In addition to empirical bounds, it is necessary to provide so-called theoretical bounds, that will prove that the risk of $\tilde{\rho}_\lambda$ will indeed converge to the Bayes risk when the sample size grows. It is the goal of the next subsection.

3.2 Theoretical bound

For this type of theoretical analysis, it is common in classification to make an additional assumption on $\mathbf{P}$ which leads to an easier task and therefore to better rates of convergence. We propose a definition adapted from Mammen and Tsybakov (1999).

Definition 2 Mammen and Tsybakov margin assumption is satisfied when there is a constant C such that, for any matrix M :

$$\mathbb{E} \left[\left(\mathbb{1}_{Y M_X \leq 0} - \mathbb{1}_{Y M_X^B \leq 0} \right)^2 \right] \leq C[\mathbf{R}(M) - \bar{\mathbf{R}}].$$

It is known that if there is a constant $t > 0$ such that $\mathbb{P}(0 < |\eta(X)| < t) = 0$ then the margin assumption is satisfied with some C that depends on t . For example, in the noiseless case where $Y = M_X^B$ almost surely, which corresponds to $t = 1$, then

$$\mathbb{E} \left[\left(\mathbb{1}_{YM_X \leq 0} - \mathbb{1}_{YM_X^B \leq 0} \right)^2 \right] = \mathbb{E} [\mathbb{1}_{YM_X \leq 0}^2] = \mathbb{E} [\mathbb{1}_{YM_X \leq 0}] = \mathbf{R}(M) = \mathbf{R}(M) - \bar{\mathbf{R}},$$

so the margin assumption is satisfied with $C = 1$.

We are now ready to state our theoretical bound. It makes a link between the integrated risk of the estimator and the lowest possible risk, which is reached by the Bayes classifier M^B . In opposition to the empirical bound, it involves non-observable quantities, depending on M^B , in the right-hand side.

Theorem 2 Assume that Mammen and Tsybakov assumption is satisfied for a given constant $C > 0$. Then, for any $\epsilon \in (0, 1)$ and for $\lambda = sn/C$, $s \in (0, 1)$, with probability at least $1 - 2\epsilon$,

$$\int \mathbf{R} d\tilde{\rho}_\lambda \leq 2(1 + 3s)\bar{\mathbf{R}} + \mathcal{C}_{C,s,\pi^\gamma} \left(\frac{\text{rank}(M^B)(m_1 + m_2)[\log n + \ell^2(M^B)] + \log(\frac{1}{\epsilon})}{n} \right)$$

where $\mathcal{C}_{C,s,\pi^\gamma}$ is known and depends only on the s , C and π^γ .

Note the adaptive nature of this result, in the sense that the estimator does *not* depend on $\text{rank}(M^B)$. Clearly, when $\text{rank}(M^B)$ is small, the prediction error will be close to the Bayes error $\bar{\mathbf{R}}$ even for small sample size. This type of inequality is often referred to as an ‘oracle inequality’ in the sense that our estimator behaves as well as if we knew the rank of M^B through an oracle.

Corollary 2 In the noiseless case $Y = \text{sign}(M_X^B)$ a.s., for any $\epsilon > 0$ and for $\lambda = 2n$, with probability at least $1 - 2\epsilon$,

$$\int \mathbf{R} d\tilde{\rho}_\lambda \leq \mathcal{C}'_{\pi^\gamma} \left[\frac{\text{rank}(M^B)(m_1 + m_2)[\log n + \ell^2(M^B)] + \log \frac{1}{\epsilon}}{n} \right] \quad (7)$$

where $\mathcal{C}'_{\pi^\gamma} = \mathcal{C}_{1, \frac{1}{4}, 1, \pi^\gamma}$.

Remark 1 Note that an empirical inequality comparable to Corollary 1 appears in [Srebro et al. \(2004\)](#). In both cases, the dependance of the bounds with respect to n is $1/\sqrt{n}$ (take $\lambda = \sqrt{n}$ in Corollary 1). One notable difference is that our bound also provides an explicit dependance to the rank, which is not the case in [Srebro et al. \(2004\)](#).

In addition to this, theoretical inequalities like Theorem 2 and Corollary 2 are completely new results. They allow to compare the out-of-sample error of our predictor to the optimal one. They show that the rate is $\text{rank}(M^B)(m_1 + m_2)/n$ up to log terms. This can not be improved as this rate is known to be minimax optimal ([Alquier et al. 2017](#)).

Remark 2 Determining the tuning parameter λ is not an easy task in practice: even though there are values that lead to the theoretical bounds, it is more efficient in practice to use cross-validation. We used this technique in the empirical results section.

4 Algorithm

4.1 General algorithm

The minimization problem (6) that defines our VB approximation is not straightforward:

$$\min_{L^0, R^0, v^L, v^R, \rho^{\gamma_1}, \dots, \rho^{\gamma_K}} AVB(L^0, R^0, v^L, v^R, \rho^{\gamma_1}, \dots, \rho^{\gamma_K}, \lambda).$$

When v^L, v^R and all the ρ^{γ_k} 's are fixed, this is actually the canonical example of so-called biconvex problems: it is convex with respect to L^0 , and with respect to R^0 , but not with respect to the pair (L^0, R^0) . Such problems are notoriously difficult. In this case, alternating blockwise optimization seems to be an acceptable strategy. While there is no guarantee that the algorithm will not get stuck in a local minimum (or even in a singular point that is actually not a minimum), it seems to give very good results in practice, and no efficient alternative is available. We refer the reader to the discussion in Subsection 9.2 page 76 of [Boyd et al. \(2011\)](#) for more details on this problem.

Our strategy is as follows. We update iteratively $L^0, R^0, v^L, v^R, \rho^{\gamma_1}, \dots, \rho^{\gamma_K}$: for L^0 and R^0 we use a gradient step, while for $v^L, v^R, \rho^{\gamma_1}, \dots, \rho^{\gamma_K}$ an explicit minimization is available. The details for the mean-field optimization (that is, w.r.t. $\rho^{\gamma_1}, \dots, \rho^{\gamma_K}$) are given in Sect. 4.2. See Algorithm 1 for the general version of the algorithm.

Algorithm 1 Variational Approximation with Hinge Loss

Require: $\epsilon, (\eta_t)_{t \in \mathbb{N}}, L_0^0, R_0^0, v_0^L, v_0^R, \rho_0^{\gamma}$

$t \leftarrow 0$

while $\|L_t^0(R_t^0)^\top - L_{t-1}^0(R_{t-1}^0)^\top\|_F^2 > \epsilon$ **do**

$t \leftarrow t + 1$

$L_t^0 \leftarrow L_{t-1}^0 - \eta_t \frac{\partial AVB}{\partial L^0}(L_{t-1}^0, R_{t-1}^0, v_{t-1}^L, v_{t-1}^R, \rho_{t-1}^{\gamma_1}, \dots, \rho_{t-1}^{\gamma_K}, \lambda)$

$R_t^0 \leftarrow R_{t-1}^0 - \eta_t \frac{\partial AVB}{\partial R^0}(L_{t-1}^0, R_{t-1}^0, v_{t-1}^L, v_{t-1}^R, \rho_{t-1}^{\gamma_1}, \dots, \rho_{t-1}^{\gamma_K}, \lambda)$

$v_t^L \leftarrow \arg \min_{v^L} AVB(L_t^0, R_t^0, v^L, v_{t-1}^R, \rho_{t-1}^{\gamma_1}, \dots, \rho_{t-1}^{\gamma_K}, \lambda)$

$v_t^R \leftarrow \arg \min_{v^R} AVB(L_t^0, R_t^0, v_t^L, v^R, \rho_{t-1}^{\gamma_1}, \dots, \rho_{t-1}^{\gamma_K}, \lambda)$

$(\rho_t^{\gamma_1}, \dots, \rho_t^{\gamma_K}) \leftarrow \arg \min_{\rho^{\gamma_1}, \dots, \rho^{\gamma_K}} AVB(L_t^0, R_t^0, v_t^L, v_t^R, \rho^{\gamma_1}, \dots, \rho^{\gamma_K}, \lambda)$

end while

4.2 Mean field optimization

As the pseudo-likelihood does not involve the parameters $(\gamma_1, \dots, \gamma_K)$, the variational distribution can be optimized in the same way as in [Lim and Teh \(2007\)](#) where the noise is Gaussian. The general update formula is:

$$\begin{aligned} \rho^{\gamma_k}(\gamma_k) &\propto \exp \mathbb{E}_{\rho^{\gamma_k}} (\log \pi(L, R, \gamma)) \\ &\propto \exp \mathbb{E}_{\rho^{\gamma_k}} (\log [\pi(L|\gamma) \pi(R|\gamma) \pi^\gamma(\gamma)]) \\ &\propto \exp \left\{ \sum_{i=1}^{m_1} \mathbb{E}_{\rho^L} [\log \pi(L_{i,k}|\gamma_k)] + \sum_{j=1}^{m_2} \mathbb{E}_{\rho^R} [\log \pi(R_{j,k}|\gamma_k)] + \log \pi^\gamma(\gamma_k) \right\} \\ &\propto \exp \left\{ -\frac{m_1 + m_2}{2} \log \gamma_k - \frac{1}{\gamma_k} \mathbb{E}_\rho \left[\frac{\sum_{i=1}^{m_1} L_{i,k}^2 + \sum_{j=1}^{m_2} R_{j,k}^2}{2} \right] + \log \pi^\gamma(\gamma_k) \right\} \end{aligned}$$

where $\rho^{-\gamma_k}$ stands for the marginal distribution of $(\gamma_{k'})_{k' \neq k}$ under ρ . The solution then depends on π^γ . In what follows we derive explicit formulas for ρ^{γ_k} according to the choice of π^γ : we remind the reader that π^γ could be either a Gamma distribution, or an Inverse-Gamma distribution.

4.2.1 Inverse-gamma prior

The conjugate prior for this part of the model is the inverse-Gamma distribution. The prior of γ_k is $\pi^\gamma = \Gamma^{-1}(\alpha, \beta)$ and its density is:

$$\pi^\gamma(\gamma_k; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \gamma_k^{-\alpha-1} \exp\left(-\frac{\beta}{\gamma_k}\right) \mathbb{1}_{\mathbb{R}^+}(\gamma_k).$$

The moments we need to develop the algorithm and to compute the empirical bound are:

$$\mathbb{E}_{\pi^\gamma}(\log \gamma_k) = \log \beta - \psi(\alpha), \text{ and } \mathbb{E}_{\pi^\gamma}(1/\gamma_k) = \frac{\alpha}{\beta},$$

where ψ is the digamma function. Therefore, we get:

$$\rho^{\gamma_k}(\gamma_k) \propto \exp\left\{-\left(\frac{m_1 + m_2}{2} + \alpha + 1\right) \log \gamma_k - \frac{1}{\gamma_k} \left(\mathbb{E}_\rho \left[\frac{\sum_{i=1}^{m_1} L_{i,k}^2 + \sum_{j=1}^{m_2} R_{j,k}^2}{2} \right] + \beta\right)\right\},$$

so we can conclude that:

$$\rho^{\gamma_k} = \Gamma^{-1}\left(\frac{m_1 + m_2}{2} + \alpha, \mathbb{E}_\rho \left[\frac{\sum_{i=1}^{m_1} L_{i,k}^2 + \sum_{j=1}^{m_2} R_{j,k}^2}{2} \right] + \beta\right). \quad (8)$$

4.2.2 Gamma prior

Even though it seems that this fact was not used in prior works on Bayesian matrix estimation, it is also possible to derive explicit formulas when the prior π^γ on γ_k 's is a $\Gamma(\alpha, \beta)$ distribution. In this case, ρ^{γ_k} is given by

$$\rho^{\gamma_k}(\gamma_k) \propto \exp\left\{\left(\alpha - \frac{m_1 + m_2}{2} - 1\right) \log \gamma_k - \beta \gamma_k - \frac{1}{\gamma_k} \mathbb{E}_\rho \left[\frac{\sum_{i=1}^{m_1} L_{i,k}^2 + \sum_{j=1}^{m_2} R_{j,k}^2}{2} \right]\right\}.$$

We remind the reader that the Generalized Inverse Gaussian distribution is a three-parameter family of distributions over $\mathbb{R}^{+*}$, written $GIG(a, b, \eta)$. Its density is given by:

$$f(x; a, b, \eta) = \frac{(a/b)^{\eta/2}}{2K_\eta(\sqrt{ab})} x^{\eta-1} \exp\left(-\frac{1}{2}(ax + bx^{-1})\right),$$

where K_λ is the modified Bessel function of second kind.

The variational distribution ρ^{γ_k} is in consequence $GIG(a_k, b_k, \eta_k)$ with:

$$a_k = 2\beta, \quad b_k = \mathbb{E}_\rho \left[\frac{\sum_{i=1}^{m_1} L_{i,k}^2 + \sum_{j=1}^{m_2} R_{j,k}^2}{2} \right], \quad \eta_k = \alpha - \frac{m_1 + m_2}{2}.$$

The moment we need in order to compute the variational distribution of L, R is:

$$\mathbb{E}_{\rho^{\gamma_k}}\left(\frac{1}{\gamma_k}\right) = \frac{K_{\eta_k-1}(\sqrt{a_k b_k})}{K_{\eta_k}(\sqrt{a_k b_k})} \sqrt{\frac{a_k}{b_k}}.$$

5 Logistic model

As mentioned in [Zhang \(2004\)](#), the hinge loss is not the only possible convex relaxation of the zero-one loss. The logistic loss $\text{logit}(u) = \log[1 + \exp(-u)]$ can also be used [even though it might lead to a loss in the rate of convergence of the risk the Bayes risk ([Zhang 2004](#))]. This leads to the definitions:

$$\mathbf{R}^\ell(M) = \mathbb{E}_{\mathbf{P}} [\text{logit}(YM_X)],$$

$$r_n^\ell(M) = \frac{1}{n} \sum_{k=1}^n \text{logit}(Y_k M_{X_k}).$$

In this case, the pseudo-likelihood $\exp(-\lambda r_n^\ell(M))$, if $\lambda = n$, is exactly equal to the likelihood of the logistic model:

$$Y|X = \begin{cases} 1 & \text{with probability } \sigma(M_X) \\ -1 & \text{with probability } 1 - \sigma(M_X) \end{cases}$$

where σ is the link function $\sigma(x) = \frac{\exp(x)}{1 + \exp(x)}$. The likelihood is written $\Lambda(L, R) = \prod_{l=1}^n \sigma(Y_l (LR^\top)_{X_l})$. The prior distribution is exactly the same as in the previous sections and the object of interest is the posterior distribution:

$$\hat{\rho}_l(d\theta) = \frac{\Lambda(L, R)\pi(d\theta)}{\int \Lambda(L, R)\pi(d\theta)}.$$

In order to deal with large matrices, it remains interesting to develop a variational Bayes algorithm. However it is not as simple as in the quadratic loss model, see [Lim and Teh \(2007\)](#) in which the authors develop a mean field approximation, because the logistic likelihood leads to intractable update formulas. A common way to deal with this model is to maximize another quantity which is very close to the one we are interested in. The principle, coming from [Jaakkola and Jordan \(2000\)](#), is well explained in [Bishop \(2006\)](#) and an extended example can be found in [Latouche et al. \(2015\)](#).

We consider the mean field approximation so the approximation is sought among the distributions ρ that are factorized $\rho(d\theta) = \prod_{i=1}^{m_1} \rho^{L_i}(dL_{i,\cdot}) \prod_{j=1}^{m_2} \rho^{R_j}(dR_{j,\cdot}) \prod_{k=1}^K \rho^{\gamma_k}(d\gamma_k)$. We have the following decomposition, for all distribution ρ :

$$\log \int \Lambda(L, R)\pi(d\theta) = \mathcal{L}(\rho) + \mathcal{K}(\rho, \hat{\rho}_l)$$

$$\text{with } \mathcal{L}(\rho) = \int \log \left(\frac{\Lambda(L, R)\pi(\theta)}{\rho(\theta)} \right) \rho(d\theta).$$

Since the left-hand side (called log-evidence) is fixed, minimizing the Kullback divergence w.r.t. ρ is the same as maximizing $\mathcal{L}(\rho)$. Unfortunately, this quantity is intractable. But a lower bound, which corresponds to a Gaussian approximation, is much more easier to optimize. We introduce the additional parameter $\xi = (\xi_l)_{l \in [n]}$.

Proposition 2 For all $\xi \in \mathbb{R}^n$ and for all ρ ,

$$\mathcal{L}(\rho) \geq \int \log \frac{H(\theta, \xi) \pi(\theta)}{\rho(\theta)} \rho(d\theta) := \mathcal{L}(\rho, \xi)$$

$$\text{where } \log H(\theta, \xi) = \sum_{l \in [n]} \left\{ \log \sigma(\xi_l) + \frac{Y_l(LR^\top)_{X_l} - \xi_l}{2} - \tau(\xi_l) \left[(LR^\top)_{X_l}^2 - \xi_l^2 \right] \right\}$$

$$\text{and } \tau(x) = 1/(2x)(\sigma(x) - 1/2).$$

Hence the estimator is (even though ξ is not the parameter of interest):

$$(\tilde{\rho}, \tilde{\xi}) = \arg \min_{\rho \in \mathcal{F}} \mathcal{L}(\rho, \xi). \quad (9)$$

5.1 Bayes algorithm

The lower bound $\mathcal{L}(\rho, \xi)$ is maximized with respect to ρ by the mean field algorithm. A direct calculation shows that the optimal distribution of each site (written with a star subscript) is given by:

$$\forall i \in [m_1], \log \rho_{\star}^{L_{i,\cdot}}(L_{i,\cdot}) = \int \log [H(\theta, \xi) \pi(\theta)] \rho^R(dR) \rho^\gamma(d\gamma) \prod_{i' \neq i} \rho(dL_{i',\cdot}) + \text{const}$$

$$\forall j \in [m_2], \log \rho_{\star}^{R_{j,\cdot}}(R_{j,\cdot}) = \int \log [H(\theta, \xi) \pi(\theta)] \rho^L(dL) \rho^\gamma(d\gamma) \prod_{j' \neq j} \rho(dR_{j',\cdot}) + \text{const}$$

As $\log H(\theta, \xi)$ is a quadratic form in $(L_{i,\cdot})_{i \in [m_1]}$ and $(R_{j,\cdot})_{j \in [m_2]}$, the variational distribution of each parameter is Gaussian and a direct calculation gives:

$$\begin{aligned} \rho_{\star}^{L_{i,\cdot}} &= \mathcal{N}(\mathcal{M}_i^L, \gamma_i^L), \quad \rho_{\star}^{R_{j,\cdot}} = \mathcal{N}(\mathcal{M}_j^R, \gamma_j^R) \quad \text{where} \\ \mathcal{M}_i^L &= \left(\frac{1}{2} \sum_{l \in \Omega_{i,\cdot}} Y_l \mathbb{E}_\rho [R_{jl,\cdot}] \right) \gamma_i^L, \quad \gamma_i^L = \left(2 \sum_{l \in \Omega_{i,\cdot}} \tau(\xi_l) \mathbb{E}_\rho [R_{jl,\cdot}^\top R_{jl,\cdot}] + \mathbb{E}_\rho \left[\text{diag} \left(\frac{1}{\gamma} \right) \right] \right)^{-1} \\ \mathcal{M}_j^R &= \left(\frac{1}{2} \sum_{l \in \Omega_{\cdot,j}} Y_l \mathbb{E}_\rho [L_{il,\cdot}] \right) \gamma_j^R, \quad \gamma_j^R = \left(2 \sum_{l \in \Omega_{\cdot,j}} \tau(\xi_l) \mathbb{E}_\rho [L_{il,\cdot}^\top L_{il,\cdot}] + \mathbb{E}_\rho \left[\text{diag} \left(\frac{1}{\gamma} \right) \right] \right)^{-1} \end{aligned}$$

The variational optimization for γ is exactly the same as in the Hinge Loss setting (with both possible prior distributions Γ and Γ^{-1}). The optimization of the variational parameters is given by:

$$\forall l \in [n], \quad \hat{\xi}_l = \sqrt{\mathbb{E}_\rho \left[(LR^\top)_{X_l}^2 \right]}.$$

6 Empirical results

In this section we compare our methods to the other 1-bit matrix completion techniques on simulated and real datasets. It is worth noting that the low rank decomposition does not involve the same matrix: in our model, it affects the Bayesian classifier matrix; in logistic model, it concerns the parameter matrix. The estimate from our algorithm is $\hat{M} = \mathbb{E}_{\tilde{\rho}_\lambda}(L) \mathbb{E}_{\tilde{\rho}_\lambda}(R)^\top$ and we focus on the zero-one loss in prediction. We first test the performances on simulated

Table 1 Type of noise

Type	Name	B	Z	Y
1	No noise	$B = 1$ a.s.	$Z = 0$ a.s.	$Y_l = \text{sign}(M_{i_l, j_l})$ a.s.
2	Switch	$B \sim 0.9\delta_1 + 0.1\delta_{-1}$	$Z = 0$ a.s.	$Y_l = \text{sign}(M_{i_l, j_l})$ w.p. 0.9
3	Logistic	$B = 1$ a.s.	$Z \sim \text{Logistic}$	$Y_l = 1$ w.p. $\sigma(M_{i_l, j_l})$

matrices and then experiment them on a real data set. We compare the four following models: (a) hinge loss with variational approximation (referred as *HL*), (b) Bayesian logistic model with variational approximation (referred as *Logis.*), (c) the frequentist logistic model from Davenport et al. (2014) (referred as *freq. Logis.*) and (d) the frequentist least squares model from Mazumder et al. (2010) (referred as *SI* for SoftImpute). The former two are tested with both Gamma and Inverse-Gamma prior distributions. The hyperparameters are all tuned by cross validation. The parameter of the frequentist methods is a regularization parameter that is also tuned by cross-validation.

The choice of K in our methods is more difficult. A large K leads to more parameters to be estimated. This considerably slows down our algorithms. In the end, some (very large) values of K are not feasible in practice. Still, what we observe is that the prior leads to an adaptive estimator, in accordance with the theoretical results: when K is taken too large (but still small enough in order to keep the computations feasible), the additional parameters are shrunk to zero. Having observed this fact, we keep $K = 10$ in many simulations. Still, we added simulations with a larger value, $K = 50$, in order to show that this shrinkage effect indeed takes place.

From a theoretical perspective, the complexity of each step of Algorithm 1 is of order $(m_1 + m_2)K$. Each step only involves very simple calculations, no matrix operations. On the opposite, the methods that use the nuclear norm are very time-consuming because the complexity of the SVD is of order $m_1 m_2 \min(m_1, m_2)$. It is possible to use approximate SVD, but the method is more difficult to tune.

6.1 Simulated data: small matrices

The goal is to assess the models with different scenarios of data generation. The general scheme of the simulations is as follows: the observations come from a 200×200 matrix and we pick randomly 20% of its entries. We set $K = 10$ in our algorithms. The observations are generated as:

$$Y_l = \text{sign}(M_{i_l, j_l} + Z_l) B_l, \quad \text{where } M \in \mathbb{R}^{m \times m}, \quad (B_l, Z_l)_{l \in [n]} \text{ are iid.}$$

The noise term (B, Z) is such that $\mathbf{R}(M) = \bar{\mathbf{R}}$ and M has low rank noted r in the followings. The predictions are directly compared to M . Two types of matrices M are built: the *type A* corresponds to the favorable case to the hinge loss; the entries of M lie in $\{-1, +1\}$.¹ The *type B* corresponds to the a more difficult classification problem because many entries of M are around 0: M is a product of two matrices with r columns where the entries are iid $\mathcal{N}(0, 1)$. The noise term is specified in Table 1. Note that the example A3 may also be seen as a switch noise with probability $\frac{e}{1+e} \approx 0.73$. The experiments are done one time for each.

¹ The matrices are built by drawing r independent columns with only $\{-1, 1\}$. The remaining columns are randomly equal to one of the first r columns multiplied by a factor in $\{-1, 1\}$.

Table 2 Prediction error on simulated observations—rank 3

Type	Logis.-G (%)	Logis.-IG (%)	HL-G (%)	HL-IG (%)	Freq. logis. (%)	SI (%)
A1	0.0	0.0	0.0	0.0	0.0	0.0
A2	0.5	0.9	0.1	0.0	0.5	0.4
A3	16.0	15.9	8.5	8.5	17.3	17.3
B1	4.1	4.0	5.3	5.8	5.1	5.6
B2	10.1	10.1	10.8	10.6	10.7	10.8
B3	16.0	16.0	22.1	21.3	19.8	19.8

Table 3 Prediction error on simulated matrices—rank 5

Type	Logis.-G (%)	Logis.-IG (%)	HL-G (%)	HL-IG (%)	Freq. logis. (%)	SI (%)
A1	0.01	0.01	0.0	0.0	0.01	0.0
A2	4.4	3.1	0.54	0.55	3.1	2.8
A3	32.5	33.1	27.0	26.7	30.1	30.0
B1	7.8	7.8	9.4	10.4	9.0	9.6
B2	17.3	17.3	17.9	18.1	18.3	18.4
B3	21.5	21.4	24.4	22.9	22.1	22.2

For rank 3 (see Table 2) and rank 5 matrices (see Table 3), the results of the Bayesian algorithms are very similar for both prior distributions and there is no evidence to favor a particular one. The results are better for the hinge loss method on type A observations and the difference of performance between models is very large for A3. On the opposite, the performance of the logistic model is better when the observations are generated from this model and when the parameter matrix is not separable. In comparison with the results from the frequentist approach, the variational approximation seems very good even though we have not at all any theoretical properties. For rank 5 matrices, the performances are worse but the meanings are the same as the rank 3 experiment.

The last simulation is a focus on the influence of the level of switch noise. On A2 type on rank 3 (Table 2), we see that 10% of corrupted entries is not enough to not almost perfectly recover the Bayesian classifier matrix. We challenge the frequentist program as well. The results are clear, see Fig. 1: the hinge loss method is better almost everywhere. For a noise up to 25%, which means that one fourth of observed entries are corrupted, it is possible to get a very good predictor with less than 10% of misclassification error. It is getting worse when the level of noise increases and the problem becomes almost impossible for noise greater than 30%.

6.2 Simulated data: large matrices

The second experiment involves larger matrices in order to assess the efficiency of the Bayesian methods on large dataset. The observations come now from a 2000×2000 matrix, 10% are observed randomly. The base matrix has now rank 10. It is worth noting that the matrix to be recovered is 100 times larger than in the first example. For the Bayesian methods, we fix $K = 50$. On the other hand, the frequentist methods need a singular value decompo-

Fig. 1 Results on simulated A2 matrices with different levels of noise—rank 3

Table 4 Prediction error on simulated matrices—rank 10

Type	Logis.-G (%)	Logis.-IG (%)	HL-G (%)	HL-IG (%)	Freq. logis. (%)	SI (%)
A1	0	0	0	0	0	0
A2	0.1	0.3	0	0	0.1	0.1
A3	9.6	9.9	1.8	1.8	9.9	9.9
B1	3.7	4.1	8.2	6.6	6.2	7.4
B2	11	11	11.3	10.6	11.8	12.0
B3	10.4	10.2	11.6	11.3	11.0	11.2

sition, which is very time consuming for such a large matrix. Exactly the same observation generation procedure (Table 1) is used and six experiments are done (Table 4).

The results are in the line with the previous section: all the methods give very similar results except for high level of switch noise (matrix A3). The gap between the proposed methods and the logistic model is quite large and the algorithms using hinge loss perform almost perfectly. On the other hand, the hinge loss models perform well on logistic data and it is worth noting that the Bayesian logistic models perform better for logistic model with low level of noise (B1).

6.3 Real data set: MovieLens

The last experiment, presented in Table 5, involves the well known MovieLens.² data set. It has already been used by Davenport et al. (2014) and we follow them for the study. The ratings lie between 1 to 5 so we split them into binary data between good ratings (above the mean which is 3.5) and bad ones. The low rank assumption is usual in this case because it is expected that the taste of a particular user is related to only few hidden parameters. The smallest data set contains 100,000 ratings from 943 users and 1682 movies so we use 95,000 of them as a training set and the 5000 remaining as the test set. The performances are very similar between the frequentist logistic model from Davenport et al. (2014) and the hinge loss model. The performances seem slightly worse for the Bayesian logistic model but it is hard to favor a particular model at this stage (note that the difference between 0.28 and 0.27 is not significant on 5000 observations).

² Available at <http://grouplens.org/datasets/movielens/100k/>.

Table 5 Misclassification rate on MovieLens 100 k data set

Algorithm	HL-IG	HL-G	Logis.-G	Logis.-IG	Freq. logis.
Misclassif. rate	0.28	0.29	0.32	0.32	0.27

7 Discussion

We undertake the 1-bit matrix completion problem with classification tools and we are able to derive PAC-bounds on the risk and an efficient algorithm to compute the estimator. The previous works only focused on GLM models, which is not the right way to establish distribution free risk bounds. This work relies on PAC-Bayesian framework and the pseudo-posterior distribution is approximated by a variational algorithm. In practice, it is able to deal with large matrices. We also derive a variational approximation of the posterior distribution in the Bayesian logistic model and it works very well in our examples.

The variational approximations look very promising in order to build algorithm which are able to deal with large data and this framework may be extended to more general models and other Machine Learning tools.

Acknowledgements We would like to thank Vincent Cottet's Ph.D. supervisor Professor Nicolas Chopin, for his kind support during the project and the three anonymous referees for their helpful and constructive comments.

Appendix: Proofs

Proofs of Proposition 1 from Sect. 2

Proof of Proposition 1 Let $\rho = \rho_{L^0, R^0, v^L, v^R, \rho^{\gamma_1}, \dots, \rho^{\gamma_k}}$ be a distribution from $\mathcal{F}$. The first term in (4) is upper bounded by the Lipschitz property of the hinge loss:

$$\begin{aligned}
 \int r_n^h(LR^\top)\rho(dL, dR, d\gamma) &= \frac{1}{n} \sum_{h=1}^n \int \left(1 - Y_h L_{ih, \cdot} R_{jh, \cdot}^\top\right)_+ \rho(dL, dR, d\gamma) \\
 &\leq \frac{1}{n} \sum_{h=1}^n \left(\left(1 - Y_h L_{ih, \cdot}^0 R_{jh, \cdot}^{0\top}\right)_+ + \int |L_{ih, \cdot} R_{jh, \cdot}^\top - L_{ih, \cdot}^0 R_{jh, \cdot}^{0\top}| \rho(dL, dR) \right) \\
 &\leq r_n^h \left(L^0 R^{0\top} \right) + \frac{1}{n} \sum_{h=1}^n \sum_{k=1}^K \left[\sqrt{v_{ih, k}^L \frac{2}{\pi}} \sqrt{v_{jh, k}^R \frac{2}{\pi}} + |R_{jh, k}^0| \sqrt{v_{ih, k}^L \frac{2}{\pi}} + |L_{ih, k}^0| \sqrt{v_{jh, k}^R \frac{2}{\pi}} \right].
 \end{aligned}$$

The second part (KL-divergence) can be explicitly calculated. Let $\rho^{L_{i,k}}$ denote the marginal distribution of $L_{i,k}$ under ρ . We define in the same way $\rho^{R_{j,k}}$. Also, $\pi \rho^{L_{i,k}|\gamma_k}$ denote the distribution of $L_{i,k}$ given γ_k under π , and we define in the same way $\pi \rho^{R_{j,k}|\gamma_k}$. Then we have

$$\begin{aligned}
 \mathcal{K}(\rho, \pi) &= \sum_{k=1}^K \left[\sum_{i=1}^{m_1} \mathbb{E}_{\rho^{\gamma_k}} \left[\mathcal{K}(\rho^{L_{i,k}}, \pi^{L_{i,k}|\gamma_k}) \right] + \sum_{j=1}^{m_2} \mathbb{E}_{\rho^{\gamma_k}} \left[\mathcal{K}(\rho^{R_{j,k}}, \pi^{R_{j,k}|\gamma_k}) \right] \right. \\
 &\quad \left. + \mathcal{K}(\rho^{\gamma_k}, \pi^{\gamma_k}) \right]
 \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2} \sum_{k=1}^K \mathbb{E}_{\rho^{\gamma_k}} \left[\frac{1}{\gamma_k} \right] \left(\sum_{i=1}^{m_1} (v_{i,k}^L + (L_{i,k}^0)^2) + \sum_{j=1}^{m_2} (v_{j,k}^R + (R_{j,k}^0)^2) \right) \\
&\quad - \frac{1}{2} \sum_{k=1}^K \left(\sum_{i=1}^{m_1} \log v_{i,k}^L + \sum_{j=1}^{m_2} \log v_{j,k}^R \right) + \sum_{k=1}^K [\mathcal{K}(\rho^{\gamma_k}, \pi^{\gamma_k}) \\
&\quad + \frac{m_1 + m_2}{2} (\mathbb{E}_{\rho^{\gamma_k}} [\log \gamma_k] - 1)] \\
&= \frac{1}{2} \sum_{k=1}^K \left[\sum_{i=1}^{m_1} \left(\mathbb{E}_{\rho^{\gamma_k}} \left[\frac{1}{\gamma_k} \right] (v_{i,k}^L + (L_{i,k}^0)^2) + \mathbb{E}_{\rho^{\gamma_k}} [\log \gamma_k] - \log v_{i,k}^L - 1 \right) \right. \\
&\quad + \sum_{j=1}^{m_2} \left(\mathbb{E}_{\rho^{\gamma_k}} \left[\frac{1}{\gamma_k} \right] (v_{j,k}^R + (R_{j,k}^0)^2) + \mathbb{E}_{\rho^{\gamma_k}} [\log \gamma_k] - \log v_{j,k}^R - 1 \right) \\
&\quad \left. + 2\mathcal{K}(\rho^{\gamma_k}, \pi^{\gamma_k}) \right].
\end{aligned}$$

□

Proofs of the results in Sect. 3.1

Proof of Theorem 1 As the indicator function is uniformly bounded by 1, we can use Lemma 5.1 in [Alquier et al. \(2015\)](#):

$$\int \mathbb{E} \exp\{\lambda[\mathbf{R}(LR^\top) - r_n(LR^\top)]\} d\pi(R, L, \gamma) \leq \exp \left[\frac{\lambda^2}{2n} \right].$$

So, the assumptions of Theorem 4.1 in [Alquier et al. \(2015\)](#) are satisfied and we obtain that, for any $\epsilon \in (0, 1)$, with probability at least $1 - \epsilon$ on the drawing of the sample, for any ρ in $\mathcal{F}$:

$$\begin{aligned}
\int \mathbf{R} d\rho &\leq \int r_n d\rho + \frac{\mathcal{K}(\rho, \pi)}{\lambda} + \frac{\lambda}{2n} + \frac{\log \frac{1}{\epsilon}}{\lambda} \\
&\leq \int r_n^h d\rho + \frac{\mathcal{K}(\rho, \pi)}{\lambda} + \frac{\lambda}{2n} + \frac{\log \frac{1}{\epsilon}}{\lambda} \text{ (as } r_n^h \geq r_n), \\
&\leq r_n^h (L^0 R^{0\top}) + \mathcal{R}(\rho, \lambda) + \frac{\lambda}{2n} + \frac{\log \frac{1}{\epsilon}}{\lambda} \text{ (thanks to Proposition 1)}.
\end{aligned}$$

We end the proof by minimizing the right-hand-side w.r.t $\rho \in \mathcal{F}$.

□

In order to prove Corollary 1, we need a preliminary result. For any $m_1 \times m_2$ matrix M with rank $r \in [K]$, we can write $M = LR^\top$ where L is $m_1 \times K$, R is $m_2 \times K$ and, up to a reordering of the columns, $L_{\cdot, r+1} = \dots = L_{\cdot, K} = 0$ and $R_{\cdot, r+1} = \dots = R_{\cdot, K} = 0$. We denote by $\mathcal{B}(M)$ the set of such pairs of matrices (L, R) and

$$\mathcal{F}(M) = \{\rho \in \mathcal{F} : (\mathbb{E}_\rho(L), \mathbb{E}_\rho(R)) \in \mathcal{B}(M)\}.$$

Lemma 2 *There is a constant $\mathcal{C}_{\pi^\gamma}$ that depends only on the choice of the prior π^γ such that for any $\lambda \leq n$,*

$$\inf_{\rho \in \mathcal{F}(M)} \mathcal{R}(\rho, \lambda) \leq \mathcal{C}_{\pi^\gamma} \frac{\text{rank}(M)(m_1 + m_2)(\ell(M)^2 + \log n)}{\lambda},$$

$\mathcal{C}_{\pi^\gamma}$ is explicitly known and depends only on the choice of the prior π^γ (Gamma, or Inverse-Gamma) and of its parameters.

It is obvious that when we combine Lemma 2 with Theorem 1 we obtain Corollary 1. It remains to prove the lemma.

Proof Let M be fixed and for short let r denote $\text{rank}(M)$. We remind that, by definition:

$$\begin{aligned} \mathcal{R}(\rho, \lambda) = & \frac{1}{n} \sum_{h=1}^n \sum_{k=1}^K \left[\sqrt{v_{ih,k}^L \frac{2}{\pi}} \sqrt{v_{jh,k}^R \frac{2}{\pi}} + |R_{jh,k}^0| \sqrt{v_{ih,k}^L \frac{2}{\pi}} + |L_{ih,k}^0| \sqrt{v_{jh,k}^R \frac{2}{\pi}} \right] \\ & + \frac{1}{\lambda} \left(\frac{1}{2} \sum_{k=1}^K \mathbb{E}_\rho \left[\frac{1}{\gamma_k} \right] \left(\sum_{i=1}^{m_1} (v_{i,k}^L + (L_{i,k}^0)^2) + \sum_{j=1}^{m_2} (v_{j,k}^R + (R_{j,k}^0)^2) \right) \right. \\ & - \frac{1}{2} \sum_{k=1}^K \left(\sum_{i=1}^{m_1} \log v_{i,k}^L + \sum_{j=1}^{m_2} \log v_{j,k}^R \right) \\ & \left. + \sum_{k=1}^K \left[\mathcal{K}(\rho^{\gamma_k}, \pi^\gamma) + \frac{m_1 + m_2}{2} (\mathbb{E}_\rho [\log \gamma_k] - 1) \right] \right). \end{aligned}$$

We will now upper bound the infimum for a special choice for $\rho = \rho_{L^0, R^0, v^L, v^R, \rho^{\gamma_1}, \dots, \rho^{\gamma_K}}$ with $(L^0, R^0) \in \mathcal{B}(M)$: for all pairs (i, k) and (j, k') $v_{i,k}^L = v_{j,k'}^R = v^0$ when $k, k' \leq r$ $v_{i,k}^L = v_{j,k'}^R = v^1$ otherwise. The choice for v^0 and v^1 will be given below. For γ , we fix two distributions ρ_γ^0 and ρ_γ^1 and fix $\rho^{\gamma_k} = \rho_\gamma^0$ for $k \leq r$ and $\rho^{\gamma_k} = \rho_\gamma^1$ otherwise. Then:

$$\begin{aligned} \inf_{\rho \in \mathcal{F}(M)} \mathcal{R}(\rho, \lambda) \leq & \frac{2}{\pi} ((K-r)v^1 + rv_0) + 2r\ell(M) \sqrt{\frac{2v^0}{\pi}} \\ & + \frac{1}{\lambda} \left(r \mathcal{K}(\rho_\gamma^0, \pi^\gamma) + (K-r) \mathcal{K}(\rho_\gamma^1, \pi^\gamma) \right) \\ & + \frac{m_1 + m_2}{2\lambda} \underbrace{\left\{ r \left[\mathbb{E}_{\rho_\gamma^0} \left[\frac{1}{\gamma_k} \right] (v^0 + \ell^2(M)) + \mathbb{E}_{\rho_\gamma^0} \log \gamma_k - \log v^0 - 1 \right] \right.}_{A_1} \\ & \left. + (K-r) \left[\mathbb{E}_{\rho_\gamma^1} \left[\frac{1}{\gamma_k} \right] v_1 + \mathbb{E}_{\rho_\gamma^1} \log \gamma_k - \log v^1 - 1 \right] \right\}. \end{aligned}$$

By actually choosing $\rho_\gamma^0 = \pi^\gamma|_{[1, 1+\delta]}$ for some $\delta > 0$ and $\rho_\gamma^1 = \pi^\gamma|_{[v^1, v_1+\delta]}$, we obtain

$$\begin{aligned} A_2 & \leq 1 - 1 + \log \frac{v^1 + \delta}{v^1} \leq \frac{\delta}{v^1}; \\ A_1 & \leq v^0 + l^2 + \delta - \log v^0. \end{aligned}$$

At this stage, we can set the free parameters v_0 , v_1 and δ in order to reach the desired rate. The choices are: $v_1 = \frac{1}{n}$, $v_0 = \frac{1}{n^2}$, $\delta = \frac{r}{Kn}$. We finally have to upper bound $r \mathcal{K}(\rho_\gamma^0, \pi^\gamma) + (K-r) \mathcal{K}(\rho_\gamma^1, \pi^\gamma)$. The upper bound actually depends on the choice for π^γ . We consider three cases: the Gamma prior with $\alpha \geq 1$, with $\alpha < 1$ and then the inverse-Gamma prior.

Let us deal with the $\Gamma^{-1}(\alpha, \beta)$ prior first:

$$\begin{aligned} & r\mathcal{K}(\rho_\gamma^0, \pi^\gamma) + (K-r)\mathcal{K}(\rho_\gamma^1, \pi^\gamma) - K \log \frac{1}{\delta} - K \log \frac{\Gamma(\alpha)}{\beta^\alpha} \\ & \leq r[(\alpha+1)\log(1+\delta) + \beta] + (K-r) \left[(\alpha+1)\log(v_1+\delta) + \frac{\beta}{v_1} \right] \\ & \leq r[(\alpha+1)\delta + \beta] + (K-r) \left[(\alpha+1)(\log v_1 + \frac{\delta}{v_1}) + \frac{\beta}{v_1} \right] \\ & \leq r[(\alpha+1)\frac{r}{Kn} + \beta] + K[(\alpha+1)(-\log n + 1) + n\beta] \end{aligned}$$

Let's turn to the $\Gamma(\alpha, \beta)$ distribution with $\alpha \geq 1$:

$$\begin{aligned} & r\mathcal{K}(\rho_\gamma^0, \pi^\gamma) + (K-r)\mathcal{K}(\rho_\gamma^1, \pi^\gamma) - K \log \frac{1}{\delta} - K \log \frac{\Gamma(\alpha)}{\beta^\alpha} \\ & \leq r[\beta(1+\delta)] + (K-r)[-(\alpha-1)\log v_1 + \beta(v_1+\delta)] \\ & \leq r[\beta(1+\delta)] + (K-r)[(\alpha-1)\log \frac{1}{v_1} + \beta(v_1+\delta)] \\ & \leq 2r\beta + K[(\alpha-1)\log n + \frac{2\beta}{n}] \end{aligned}$$

The last case is the $\Gamma(\alpha, \beta)$ distribution with $0 < \alpha < 1$:

$$\begin{aligned} & r\mathcal{K}(\rho_\gamma^0, \pi^\gamma) + (K-r)\mathcal{K}(\rho_\gamma^1, \pi^\gamma) - K \log \frac{1}{\delta} - K \log \frac{\Gamma(\alpha)}{\beta^\alpha} \\ & \leq r[-(\alpha-1)\log(1+\delta) + \beta(1+\delta)] + (K-r)[-(\alpha-1)\log(v_1+\delta) + \beta(v_1+\delta)] \\ & \leq r[(1-\alpha)\delta + \beta(1+\delta)] + (K-r)[(1-\alpha)(\log v_1 + \frac{\delta}{v_1}) + \beta(v_1+\delta)] \\ & \leq 2r\beta + r(1-\alpha)\frac{r}{Kn} + K \left[(1-\alpha)(-\log n + 1) + \frac{2\beta}{n} \right] \end{aligned}$$

In any case, as $\lambda \leq n$, when α and β are constant, the leading term is in $\frac{r(m_1+m_2)(\ell^2(M)+\log n)}{\lambda}$ so we can upper bound the whole by $\mathcal{C}_{\pi^\gamma} \frac{r(m_1+m_2)(\ell^2(M)+\log n)}{\lambda}$ where $\mathcal{C}_{\pi^\gamma}$ depends on α and β (and takes a different form depending on the case: Gamma or inverse-Gamma). $\square$

Note actually that from the previous proof we can provide more explicit forms for the bound in the three cases. We did not include this in the core of the paper, but we prove the following lemmas for the sake of completeness.

Lemma 3 When $\pi^\gamma = \Gamma(\alpha, \beta)$,

$$\begin{aligned} \inf_{\rho \in \mathcal{F}(M)} \mathcal{R}(\rho, \lambda) & \leq \frac{1}{n} \left[\frac{4}{\pi} K + \sqrt{\frac{8}{\pi}} r \ell(M) \right] + \frac{r(m_1+m_2)}{2\lambda} [3 + \ell^2(M) + 2\log n] \\ & \quad + \frac{K}{\lambda} \left[\log \frac{Kn}{r} + \log \frac{\Gamma(\alpha)}{\beta^\alpha} + \frac{r}{K} \left[(\alpha+1)\frac{r}{Kn} + \beta \right] \right. \\ & \quad \left. + [(\alpha+1)(-\log n + 1) + n\beta] \right]. \end{aligned}$$

Lemma 4 When $\pi^\gamma = \Gamma^{-1}(\alpha, \beta)$ with $\alpha \geq 1$,

$$\inf_{\rho \in \mathcal{F}(M)} \mathcal{R}(\rho, \lambda) \leq \frac{1}{n} \left[\frac{4}{\pi} K + \sqrt{\frac{8}{\pi}} r \ell(M) \right] + \frac{r(m_1 + m_2)}{2\lambda} [3 + \ell^2(M) + 2 \log n] \\ + \frac{K}{\lambda} \left[\log \frac{Kn}{r} + \log \frac{\Gamma(\alpha)}{\beta^\alpha} + \frac{2r\beta}{K} + \left[(\alpha - 1) \log n + \frac{2\beta}{n} \right] \right].$$

Lemma 5 When $\pi^\gamma = \Gamma^{-1}(\alpha, \beta)$ with $0 < \alpha < 1$,

$$\inf_{\rho \in \mathcal{F}(M)} \mathcal{R}(\rho, \lambda) \leq \frac{1}{n} \left[\frac{4}{\pi} K + \sqrt{\frac{8}{\pi}} r \ell(M) \right] + \frac{r(m_1 + m_2)}{2\lambda} [3 + \ell^2(M) + 2 \log n] \\ + \frac{K}{\lambda} \left[\log \frac{Kn}{r} + \log \frac{\Gamma(\alpha)}{\beta^\alpha} + \frac{2r\beta}{K} + (1 - \alpha) \frac{r^2}{K^2 n} \right. \\ \left. + \left[(1 - \alpha)(-\log n + 1) + \frac{2\beta}{n} \right] \right].$$

Proofs of the results in Sect. 3.2

We first start with preliminary lemmas.

Lemma 6 For $\epsilon > 0$, with probability at least $1 - \epsilon$ and for every $s \in (0, 1)$,

$$\overline{r}_n \leq (1 + s)\overline{\mathbf{R}} + \frac{1}{ns} \log \frac{1}{\epsilon}$$

Proof Let $s \in (0, 1)$, then

$$\mathbb{E}(\exp[sn\overline{r}_n]) = \prod_{h=1}^n \mathbb{E} \left(\exp \left[s \mathbb{1}(Y_h M_{X_h}^B < 0) \right] \right) \\ = \prod_{h=1}^n \mathbb{E} \left(\exp \left[s \mathbb{1}(Y_h M_{X_h}^B < 0) + 0 \left(1 - \mathbb{1}(Y_h M_{X_h}^B < 0) \right) \right] \right) \\ \leq \prod_{h=1}^n \left(\left(1 - \mathbb{E} \left[\mathbb{1}(Y_h M_{X_h}^B < 0) \right] \right) + e^s \mathbb{E} \left[\mathbb{1}(Y_h M_{X_h}^B < 0) \right] \right) \\ \leq \prod_{h=1}^n ((1 - \overline{\mathbf{R}}) + e^s \overline{\mathbf{R}}) = \prod_{h=1}^n (1 + \overline{\mathbf{R}}(e^s - 1)) \\ \leq \prod_{h=1}^n \exp(\overline{\mathbf{R}}(e^s - 1)) = \exp(n\overline{\mathbf{R}}(e^s - 1)).$$

Therefore, for $\epsilon \in (0, 1)$:

$$\mathbb{E} \left[\exp \left(sn\overline{r}_n - n\overline{\mathbf{R}}(e^s - 1) - \log \frac{1}{\epsilon} \right) \right] \leq \epsilon.$$

We use the fact that $\mathbb{E}[\exp U] \geq \mathbb{P}(U > 0)$ (Markov's inequality) for any U so, with probability at least $1 - \epsilon$:

$$\overline{r}_n \leq \frac{e^s - 1}{s} \overline{\mathbf{R}} + \frac{\log \frac{1}{\epsilon}}{sn}$$

On $[0, 1]$, $e^x \leq 1 + x + x^2$, thus ending the proof. $\square$

Lemma 7 Assume that Mammen and Tsybakov assumption is satisfied for a certain constant C . Assume that $\lambda < 2n/C$. Then, for $\epsilon > 0$, with probability at least $1 - \epsilon$:

$$\int \mathbf{R} d\tilde{\rho}_\lambda \leq \bar{\mathbf{R}} + \frac{1}{1 - C\lambda/(2n)} \left\{ \inf_{\rho \in \mathcal{F}} \left[r_n^h \left(L^0 R^{0\top} \right) + \mathcal{R}(\rho, \lambda) \right] - \bar{r}_n + \frac{1}{\lambda} \log \left(\frac{1}{\epsilon} \right) \right\} \quad (10)$$

Proof Assume that the Mammen and Tsybakov assumption is satisfied for a certain constant C . The zero-one loss is bounded then, from Bernstein's inequality [Theorem 10 page 37 in Boucheron et al. (2013)] we get:

$$\int \mathbb{E} \exp \{ \lambda [\mathbf{R}(LR^\top) - \bar{\mathbf{R}}] - \lambda [r_n(LR^\top) - \bar{r}_n] \} d\pi(L, R, \gamma) \leq \int \exp [C\lambda^2/(2n) [\mathbf{R}(LR^\top) - \bar{\mathbf{R}}]] d\pi(L, R, \gamma).$$

Apply Fubini's theorem to the inequality:

$$\mathbb{E} \int \exp \{ (\lambda - C\lambda^2/(2n)) [\mathbf{R}(LR^\top) - \bar{\mathbf{R}}] - \lambda [r_n(LR^\top) - \bar{r}_n] \} \pi(d\theta) \leq 1$$

(we remind that $\theta = (L, R, \gamma)$ for short).

$$\mathbb{E} \exp \left\{ \sup_{\rho} \int [\lambda [\mathbf{R}(LR^\top) - \bar{\mathbf{R}}] - \lambda [r_n(LR^\top) - \bar{r}_n] - C\lambda^2/(2n) [\mathbf{R}(LR^\top) - \bar{\mathbf{R}}]] \rho(d\theta) - \mathcal{H}(\rho, \pi) \right\} \leq 1.$$

Using Markov's inequality,

$$\mathbb{P} \left(\sup_{\rho} \int [(\lambda - C\lambda^2/(2n)) [\mathbf{R}(LR^\top) - \bar{\mathbf{R}}] - \lambda [r_n(LR^\top) - \bar{r}_n]] \rho(d\theta) - \mathcal{H}(\rho, \pi) + \log \epsilon > 0 \right) \leq \epsilon.$$

Then take the complementary of this event and we get that with probability at least $1 - \epsilon$:

$$\forall \rho, \quad (\lambda - C\lambda^2/(2n)) \int [\mathbf{R}(LR^\top) - \bar{\mathbf{R}}] \rho(d\theta) \leq \lambda \int [r_n(LR^\top) - \bar{r}_n] \rho(d\theta) + \mathcal{H}(\rho, \pi) + \log \frac{1}{\epsilon}$$

Now, note that

$$\begin{aligned} (\lambda - C\lambda^2/(2n)) \left[\int \mathbf{R} d\rho - \bar{\mathbf{R}} \right] &\leq \lambda \left[\int r_n d\rho - \bar{r}_n \right] + \mathcal{H}(\rho, \pi) + \log \left(\frac{1}{\epsilon} \right) \\ \Rightarrow (\lambda - C\lambda^2/(2n)) \left[\int \mathbf{R} d\rho - \bar{\mathbf{R}} \right] &\leq \lambda \left[\int r_n^h d\rho + \frac{1}{\lambda} \mathcal{H}(\rho, \pi) \right] - \lambda \bar{r}_n + \log \left(\frac{1}{\epsilon} \right) \\ \Rightarrow (\lambda - C\lambda^2/(2n)) \left[\int \mathbf{R} d\rho - \bar{\mathbf{R}} \right] &\leq \lambda \left[r_n^h \left(L^0 R^{0\top} \right) + \mathcal{R}(\rho, \lambda) - \bar{r}_n + \frac{1}{\lambda} \log \left(\frac{1}{\epsilon} \right) \right] \end{aligned}$$

As it stands for all ρ then the right hand side can be minimized and the minimizer over $\mathcal{F}$ is $\tilde{\rho}_\lambda$. Thus we get, when $\lambda < 2n/C$,

$$\int \mathbf{R} d\tilde{\rho}_\lambda \leq \bar{\mathbf{R}} + \frac{1}{1 - C\lambda/(2n)} \left\{ \inf_{\rho \in \mathcal{F}} \left[r_n^h \left(L^0 R^{0\top} \right) + \mathcal{R}(\rho, \lambda) \right] - \bar{r}_n + \frac{1}{\lambda} \log \left(\frac{1}{\epsilon} \right) \right\}$$

□

We are now ready for the proofs.

Proof of Theorem 2 We apply Lemma 7 and, as we have $\mathcal{F}(M^B) \subset \mathcal{F}$,

$$\int \mathbf{R} d\tilde{\rho}_\lambda \leq \bar{\mathbf{R}} + \frac{1}{1 - C\lambda/(2n)} \left\{ \inf_{\rho \in \mathcal{F}(M)} \left[r_n^h \left(L^0 R^{0\top} \right) + \mathcal{R}(\rho, \lambda) \right] - \bar{r}_n + \frac{1}{\lambda} \log \left(\frac{1}{\epsilon} \right) \right\} \quad (11)$$

As by definition, all the entries of M^B are in $\{-1, 1\}$, $r_n^h(M^B) = 2\bar{r}_n$ and then, by Lemma 2:

$$\int \mathbf{R} d\tilde{\rho}_\lambda \leq \bar{\mathbf{R}} + \frac{1}{1 - C\lambda/(2n)} \left\{ 2\bar{r}_n + \mathcal{C}_{\pi^\gamma} \frac{\text{rank}(M^B)(m_1 + m_2)(\ell^2(M) + \log n)}{\lambda} + \frac{1}{\lambda} \log \left(\frac{1}{\epsilon} \right) \right\}$$

Then, we use Lemma 7 to get, with probability at least $1 - 2\epsilon$,

$$\int \mathbf{R} d\tilde{\rho}_\lambda \leq \bar{\mathbf{R}} + \frac{1}{1 - C\lambda/(2n)} \left\{ 2(1 + s)\bar{\mathbf{R}} + \frac{1}{ns} \log \frac{1}{\epsilon} + \mathcal{C}_{\pi^\gamma} \frac{\text{rank}(M^B)(m_1 + m_2)(\ell^2(M) + \log n)}{\lambda} + \frac{1}{\lambda} \log \left(\frac{1}{\epsilon} \right) \right\}$$

To end up the proof, we have to take $\lambda = \frac{2cn}{C}$ with $c \in (0, 1/2)$. We thus have:

$$\frac{1}{1 - C\lambda/(2n)} = \frac{1}{1 - c} \leq 1 + 2c,$$

this ends the proof by taking $c = s/2$. □

Proof of Corollary 2 As we are in the noiseless case, the margin assumption is satisfied with $C = 1$, and $\bar{\mathbf{R}} = 0$. □

Detailed calculations for Sect. 5

Proof of Proposition 2 From Jaakkola and Jordan (2000), we have the following lower bound:

$$\forall (x, \xi) \in \mathbb{R}^2$$

$$\log \sigma(x) \geq \log \sigma(\xi) + \frac{x - \xi}{2} - \tau(\xi)(x^2 - \xi^2) \quad \text{where} \quad \tau(\xi) = \frac{1}{2\xi} \left(\sigma(\xi) - \frac{1}{2} \right)$$

The likelihood of one observation $y \in \{-1, 1\}$ at point x is then lower bounded:

$$\forall \xi \in \mathbb{R}, \sigma(yx) \geq \sigma(\xi) \exp \left\{ \frac{yx - \xi}{2} - \tau(\xi)(x^2 - \xi^2) \right\} := h(yx, \xi).$$

Therefore, the likelihood of the model is lower bounded:

$$\begin{aligned} \forall \xi \in \mathbb{R}^n, \quad \Lambda(L, R) &= \prod_{l=1}^n \sigma(Y_l(LR^\top)_{X_l}) \\ &\geq \prod_{l=1}^n \sigma(\xi_l) \exp \left\{ \frac{(LR^\top)_{X_l} - \xi_l}{2} - \tau(\xi_l) \left[(LR^\top)_{X_l}^2 - \xi_l^2 \right] \right\} := H(\theta, \xi). \end{aligned}$$

□

It is now easy to optimize $\mathcal{L}(\rho, \xi)$ with respect to ξ elementwise, which is the same as maximizing $H(\theta, \xi)$ elementwise and then each part $h(Y_l(LR^\top)_{X_l}, \xi_l)$:

$$\begin{aligned} \hat{\xi}_l &= \arg \max_{\xi_l} \int \log H(\theta, \xi) \rho(d\theta) = \arg \max_{\xi_l} \mathbb{E}_\rho \left[\log h(Y_l(LR^\top)_{X_l}, \xi_l) \right] \\ &= \arg \max_{\xi_l} \left\{ \log \sigma(\xi_l) - \frac{\xi_l}{2} - \tau(\xi_l) \left(\mathbb{E}_\rho \left[(LR^\top)_{X_l}^2 \right] - \xi_l^2 \right) \right\} \end{aligned}$$

The maximum is reached at the zero of the derivative and we can conclude that:

$$\hat{\xi}_l^2 = \mathbb{E}_\rho \left[(LR^\top)_{X_l}^2 \right].$$

References

- Alquier, P., Cottet, V., & Lecué, G. (2017). Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions. arXiv preprint [arXiv:1702.01402](https://arxiv.org/abs/1702.01402).
- Alquier, P., Ridgway, J., & Chopin, N. (June 2015). On the properties of variational approximations of Gibbs posteriors. arXiv e-prints.
- Bishop, C. M. (2006). *Pattern recognition and machine learning (information science and statistics)*. New York: Springer.
- Boucheron, S., Lugosi, G., & Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford: OUP.
- Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1), 1–122.
- Cai, T., & Zhou, W.-X. (2013). A max-norm constrained minimization approach to 1-bit matrix completion. *Journal of Machine Learning Research*, 14, 3619–3647.
- Candès, E. J., & Plan, Y. (2010). Matrix completion with noise. *Proceedings of the IEEE*, 98(6), 925–936.
- Candès, E. J., & Recht, B. (2012). Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6), 111–119.
- Candès, E. J., & Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5), 2053–2080.
- Catoni, O. (2004). Statistical learning theory and stochastic optimization. In J. Picard (Ed.), *Saint-Flour Summer School on probability theory 2001*, Lecture notes in mathematics Berlin: Springer.
- Catoni, O. (2007). *PAC-Bayesian supervised classification: The thermodynamics of statistical learning* (Vol. 56). Institute of mathematical statistics lecture notes—Monograph series Beachwood, OH: Institute of Mathematical Statistics.
- Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *Annals of Statistics*, 43(1), 177–214.
- Dalalyan, A., & Tsybakov, A. B. (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72(1), 39–61.
- Davenport, M. A., Plan, Y., van den Berg, E., & Wooters, M. (2014). 1-Bit matrix completion. *Information and Inference*, 3(3), 189–223.
- Herbrich, R., & Graepel, T. (2002). A PAC-Bayesian margin bound for linear classifiers. *IEEE Transactions on Information Theory*, 48(12), 3140–3150.

- Herbster, M., Pasteris, S., & Pontil, M. (2016). Mistake bounds for binary matrix completion. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, R. Garnett, & R. Garnett (Eds.), *Proceedings of the 29th conference on neural information processing systems (NIPS 2016)*. Barcelona, Spain: NIPS Proceedings.
- Hsieh, C.-J., Natarajan, N., & Dhillon, I. S. (2015). PU learning for matrix completion. In *Proceedings of the 32nd international conference on machine learning*, pp. 2445–2453.
- Jaakkola, T. S., & Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1), 25–37.
- Klopp, O., Lafond, J., Moulines, É., & Salmon, J. (2015). Adaptive multinomial matrix completion. *Electronic Journal of Statistics*, 9(2), 2950–2975.
- Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5(2), 369–412.
- Latouche, P., Robin, S., & Ouadah, S. (2015). Goodness of fit of logistic models for random graphs. arXiv preprint [arXiv:1508.00286](https://arxiv.org/abs/1508.00286).
- Lim, Y. J. & Teh, Y. W. (2007). Variational Bayesian approach to movie rating prediction. In *Proceedings of KDD cup and workshop*.
- Mai, T. T., & Alquier, P. (2015). A bayesian approach for noisy matrix completion: Optimal rate under general sampling distribution. *Electronic Journal of Statistics*, 9(1), 823–841.
- Mammen, E., & Tsybakov, A. (1999). Smooth discrimination analysis. *The Annals of Statistics*, 27(6), 1808–1829.
- Mazumder, R., Hastie, T., & Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11(Aug), 2287–2322.
- McAllester, D. A. (1998). Some PAC-Bayesian theorems. In *Proceedings of the eleventh annual conference on computational learning theory* (pp. 230–234). New York, ACM.
- Park, T., & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482), 681–686.
- Recht, B., & Ré, C. (2013). Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2), 201–226.
- Salakhutdinov, R. & Mnih, A. (2008). Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th international conference on machine learning*, pp. 880–887.
- Seldin, Y., & Tishby, N. (2010). PAC-Bayesian analysis of co-clustering and beyond. *Journal of Machine Learning Research*, 11(Dec), 3595–3646.
- Seldin, Y., Laviolette, F., Cesa-Bianchi, N., Shawe-Taylor, J., & Auer, P. (2012). PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12), 7086–7093.
- Shawe-Taylor, J., & Langford, J. (2003). PAC-Bayes and margins. *Advances in Neural Information Processing Systems*, 15, 439.
- Srebro, N., Rennie, J., & Jaakkola, T. S. (2004). Maximum-margin matrix factorization. In *Advances in neural information processing systems*, pp. 1329–1336.
- Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.
- Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32(1), 56–85.