Adaptive one-bit matrix completion

Joseph Salmon [1] & Jean Lafond [1] & Olga Klopp [2] & Éric Moulines [1]

[1] *Institut Mines-Télécom*
Télécom ParisTech
CNRS LTCI
46 Rue Barrault, 75634 Paris Cedex 13, France
E-mail : prénom.nom@telecom-paristech.fr
[2] *CREST et MODAL'X,Université Paris Ouest 200 avenue de la République, 92001 Nanterre, France E-mail : nomprénom@math.cnrs.fr*

Résumé.
Ces dernières années, de nombreux travaux ont montré les bénéfices d'utiliser des techniques de complétion de matrice pour améliorer les systèmes de recommandation (pour la recommandation de films ou de musique notamment). La plupart des études faites ont considéré le cas oú les coefficients à déterminer sont des scores continus. Dans ce travail nous proposons d'étudier le cas où les observations sont de nature binaire. Plus précisément nous nous intéressons à la complétion de matrices dont les coefficients suivent une distribution logistique avec une fonction de lien concave. Notre travail permet de traiter des schémas d'échantillonage de coefficients variés, et l'estimateur que nous considérons est basé sur une méthode de (log-)vraisemblance pénalisée par la norme nucléaire (encore appelée norme-trace). Plus précisément, nous proposons des bornes contrôlant la divergence de Kullback-Leibler entre la vraie distribution matricielle et notre estimation. En pratique, nous utilisons un algorithme de descente de gradient par coordonnées pour permettre de construire notre estimateur dans un cadre de grande dimension.

Mots-clés. Complétion de matrice, statistique en grande dimension, faible rang, ...

Abstract. In the past few years, a large variety of works has shown the benefits of using matrix completion techniques to improve recommender system (*e.g.,* for movie or music recommendation). Most works have considered cases where the coefficients to be determined are continuous scores. Here, we investigate the case where the observations are binary. More precisely, we deal with the problem of matrix completion when the matrix coefficients follow a logistic distribution with a known concave link function. We assume a general sampling scheme for the acquisition process of the coefficients. We study the performance of a nuclear-norm penalized estimator. More precisely we derive bounds for the Kullback-Leibler divergence between the true and estimated distribution. In practice we propose an algorithm based on coordinate gradient descent in order to tackle potentially high dimensional settings.

Keywords. Matrix completion, high-dimensional statistics, low rank, ...

1 Introduction

The matrix completion problem arises in many practical situations and its study has experienced exciting developments in the past few years. One of the well known applications of matrix completion are recommendation systems. For example, Pandora, a popular online music service, lets you rate tracks as you listen. So we consider a matrix where rows represent users, columns represent tracks and the users' rates are the entries of the matrix. Of course, each user may rate only a small part of the tracks. The question is whether it is possible to infer all the missing rates from the known rates in order to "understand" each user's musical tastes ? The answer is yes if the unknown matrix has low rank.

There exists a rapidly growing literature on matrix completion problem (see, *e.g.,* [2, 3, 5, 7, 6, 8]). In the usual matrix completion setting we observe real-valued entries. However, in some applications, such as recommendation systems we are only able to collect positive or negative feedback. The binary output is generated according to a probability distribution parametrized by the corresponding entry of the unknown matrix M. This setting, known as 1-bit matrix completion, was introduced by Davenport *et al.* [4]. They use a very popular nuclear norm minimization approach and show that the recovery is still possible even if the observations are highly quantized. Davenport *et al.* consider the uniform sampling model where entries are assumed to be sampled uniformly at random. Unfortunately, this condition is not realistic in applications. If we take the example of Pandora, some users are more active than others and some tracks are rated more frequently. Another important point is that their methods requires an upper bound on the nuclear norm or on the rank of the unknown matrix. Usually this kind of bounds is not available.

In the paper by Cai and Zhou [1], 1-bit matrix completion was further considered using a max-norm regularization term. The method of [1] allows more general non-uniform samplings but still requires an upper bound on the max-norm of the unknown matrix M. Such bounds are usually not available and getting an estimation of the max-norm of the unknown (partially) observed matrix M is a challenging problem.

We propose a method based on the constrained nuclear norm minimization which allows us to consider general sampling distribution and requires only an upper bound on the maximum absolute value of the entries of M. This condition is very mild since such a bound is often known in applications. For instance, if the entries of M are user's ratings it is the maximal rating. The previously cited papers on 1-bit matrix completion also require this bound, in addition to the bounds on the nuclear or max norm.

1.1 Notations

We consider the Hilbert space $\mathbb{R}^{m_1 \times m_2}$ with the standard scalar product $\langle A|B\rangle := \mathrm{tr}(A^\top B)$. For a given matrix $A \in \mathbb{R}^{m_1 \times m_2}$ we write $\|A\|_\infty := \max_{i,j} |A_{i,j}|$ and denote its

Schatten p-norm by

$$\|A\|_{\sigma,p} := \left(\sum_{i=1}^{m_1 \wedge m_2} \sigma_i(A)^p \right)^{1/p} ,$$

where $\sigma_i(A)$ are the singular values of A ordered in decreasing order and $m_1 \wedge m_2 := \min(m_1, m_2)$. The operator norm of A, is denoted for consistency by $\|A\|_{\sigma,\infty} := \sigma_1(A)$. We denote by $\mathcal{S}_1(A) \subset \mathbb{R}^{m_1}$ (*resp.* $\mathcal{S}_2(A) \subset \mathbb{R}^{m_2}$) the linear spans generated by left (*resp.* right) singular vectors of A. $P_{\mathcal{S}_1^\perp(A)}$ (*resp.* $P_{\mathcal{S}_2^\perp(A)}$) denote the orthogonal projections on $\mathcal{S}_1^\perp(A)$ (*resp.* $\mathcal{S}_2^\perp(A)$). We then define the following orthogonal projections on $\mathbb{R}^{m_1 \times m_2}$

$$\mathcal{P}_A^\perp : B \to P_{\mathcal{S}_1^\perp(A)} B P_{\mathcal{S}_2^\perp(A)} \text{ and } \mathcal{P}_A \, B \to B - \mathcal{P}_A^\perp(B) .$$

If we consider two matrices $P, Q \in [0,1]^{m_1 \times m_2}$, the square Hellinger distance between P and Q is defined as

$$d_H^2(P,Q) := \frac{1}{m_1 m_2} \sum_{\substack{1 \le i \le m_1 \\ 1 \le j \le m_2}} \left[(\sqrt{P_{i,j}} - \sqrt{Q_{i,j}})^2 + (\sqrt{1 - P_{i,j}} - \sqrt{1 - Q_{i,j}})^2 \right] ,$$

and the Kullback-Liebler divergence is

$$KL(P,Q) := \frac{1}{m_1 m_2} \sum_{\substack{1 \le i \le m_1 \\ 1 \le j \le m_2}} \left[P_{i,j} \log \frac{P_{i,j}}{Q_{i,j}} + (1 - P_{i,j}) \log \frac{1 - P_{i,j}}{1 - Q_{i,j}} \right] .$$

1.2 Model Specification

We consider a parameter matrix $X^* \in \mathbb{R}^{m_1 \times m_2}$ which is not directly observed. For an *i.i.d.* sequence $(\omega_i)_{1 \le i \le n}$ of indexes over $[m_1] \times [m_2]$, we observe n independent random elements $(Y_i)_{1 \le i \le n} \in \{-1, 1\}^n$ which are distributed as :

$$\mathbb{P}(Y_i = 1) = f(X_{\omega_i}^*) \text{ and } \mathbb{P}(Y_i = -1) = 1 - f(X_{\omega_i}^*) ,$$

where f is a link function taking value in $[0,1]$. For ease of notation, we write X_i^* instead of $X_{\omega_i}^*$. The log-likelihood of the observations $X \to \mathrm{L_Y}(X)$ is given by :

$$\mathrm{L_Y}(X) = \sum_{i=1}^{n} \left[\mathbb{1}_{\{Y_i = 1\}} \log(f(X_i)) + \mathbb{1}_{\{Y_i = -1\}} \log(1 - f(X_i)) \right] .$$

If we define the matrices $E_{k,l}$ as the canonical basis in $\mathbb{R}^{m_1 \times m_2}$ then $X_i = \langle X | E_{w_i} \rangle$. With the abuse of notation $E_i := E_{w_i}$, the log-likelihood may be expressed as :

$$\mathrm{L_Y}(X) = \sum_{i=1}^{n} \left[\mathbb{1}_{\{Y_i = 1\}} \log \left(f(\langle X | E_i \rangle) \right) + \mathbb{1}_{\{Y_i = -1\}} \log \left(1 - f(\langle X | E_i \rangle) \right) \right] .$$

Note that by assumption, the matrices E_i for $i = 1, \dots, n$, are supposed to be *i.i.d.* on $\mathscr{E}$, the set of canonical matrices $\mathscr{E} := \{E_{k,l} : \ (k,l) \in [m_1] \times [m_2]\}$. Their distribution is denoted by Π.
We also assume that we know a bound on the coefficients of X^*.

Assumption 1. We know $\gamma > 0$ such that $\|X^*\|_\infty \le \gamma$.

The estimator we study is defined as follows :

$$\hat{X} = \underset{\substack{X \in \mathbb{R}^{m_1 \times m_2} \\ \|X\|_\infty \le \gamma}}{\arg\min} \ \Phi_Y^\lambda(X) \ , \tag{1}$$

where

$$\Phi_Y^\lambda(X) = -\frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{1}_{\{Y_i = 1\}} \log \left(f(\langle X | E_i \rangle) \right) + \mathbb{1}_{\{Y_i = -1\}} \log \left(1 - f(\langle X | E_i \rangle) \right) \right) + \lambda \|X\|_{\sigma,1} \ ,$$

with $\lambda > 0$ a regularization parameter.

2 Main results

In this section we present two results controlling the estimation error of $\hat{X}$. Before doing so, let us introduce some additional notations and assumptions. For a given parameter matrix $X \in \mathbb{R}^{m_1 \times m_2}$ the score function, defined as the gradient of the log-likelihood, is $\Sigma_Y(X) = -\nabla \, \mathrm{L}_\mathrm{Y}(X) \, / n$. For the true parameter matrix X^* we also define $\Sigma^* := \Sigma_Y(X^*)$. We also need the following constants depending on the link function f

$$\begin{aligned} M_\gamma &= \sup_{|x| \le \gamma} 2 |\log(f(x))| \ , \\ L_\gamma &= \max \left(\sup_{|x| \le \gamma} \frac{|f'(x)|}{f(x)}, \sup_{|x| \le \gamma} \frac{|f'(x)|}{1 - f(x)} \right) \ , \\ K_\gamma &= \inf_{|x| \le \gamma} \frac{f'(x)^2}{8 f(x)(1 - f(x))} \ . \end{aligned}$$

In our framework, we consider a general distribution Π for the random matrices $(E_i)_{1 \le i \le n}$. However, we need to control "how far" Π is from the uniform distribution. Denoting

$$\pi_{k,l} := \Pi(E_1 = E_{k,l}) \ , \tag{2}$$

we make the following assumption.

Assumption 2. There exists a constant $\mu > 0$ such that for all indexes $(k,l) \in [m_1] \times [m_2]$

$$\pi_{k,l} \ge \frac{1}{\mu m_1 m_2} \ .$$

Let $C_l = \sum_{k=1}^{m_2} \pi_{k,l}$ and $R_k = \sum_{l=1}^{m_1} \pi_{k,l}$.

Assumption 3. There exists a constant $L > 0$ such that

$$\max_{k,l}(R_k, C_l) \leq \frac{\nu}{m_1 \wedge m_2} \ ,$$

We can now state the result.

Theorem 1. *Suppose that Assumption 1, Assumption 2 and Assumption 3 hold and that* $n \geq 2\log(d)/(9\nu)$. *For*

$$\lambda = 6L_\gamma \sqrt{\frac{2\nu \log(d)}{mn}} \ \textit{and}\ \beta = 8eM_\gamma \sqrt{\frac{\log(d)}{n}} \ ,$$

we have with least probability $1 - 3/d$ *:*

$$KL\left(f(X^*), f(\hat{X})\right) \leq \max\left(c^* r^* \nu \mu^2 L_\gamma^2 \frac{\log(d)}{mn}, \mu\beta\right) \ ,$$

where c^* *is a universal constant and* $r^* := (2m_1 m_2 \operatorname{rank}(X^*))/K_\gamma$.

Références

[1] T. T. Cai and W. Zhou. Matrix completion via max-norm constrained optimization. *CoRR*, abs/1303.0341, 2013.

[2] E. J. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6) :925–936, June 2010.

[3] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6) :717–772, 2009.

[4] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters. 1-bit matrix completion. *CoRR*, abs/1209.3672, 2012.

[5] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *J. Mach. Learn. Res.*, 11 :2057–2078, 2010.

[6] O. Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 2(1) :282–303, 02 2014.

[7] V. Koltchinskii, A. B. Tsybakov, and K. Lounici. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5) :2302–2329, 2011.

[8] S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion : optimal bounds with noise. *J. Mach. Learn. Res.*, 13 :1665–1697, 2012.