Matrix completion*

22.04., 28.04.2024 (lectures)

David Steurer (lecture), Rares Buhai (scribe)

Contents

1 Introduction

In matrix completion we consider a low-rank matrix model with missing data.

For instance, recall the movie rating example from the low-rank matrix factorization lectures. In that example, entry (i, j) of the matrix represents the rating given to movie i by user j. However, in reality, most users will only rate a small subset of the movies. The question that we are concerned with is: can we estimate the missing data?

*students – staff – index — versions: web / pdf / gitlab

2 Formalization

The ground truth consists of an unknown matrix $\Theta^\star \in \mathbb{R}^{d\times d}$ with $\operatorname{rank}(\Theta^\star) \le r$. For instance, $\Theta^\star$ may represent the ratings given to d movies by d users. For simplicity we assume that $\Theta^\star$ is a square matrix, but the case of a rectangular matrix is similar.

We observe n noisy entries of $\Theta^\star$ at random locations. Specifically, we observe $\boldsymbol{X}_i$ and $\boldsymbol{y}_i$ pairs such that

$$\boldsymbol{y}_i = \langle \boldsymbol{X}_i, \Theta^\star\rangle + \boldsymbol{w}_i,$$

where

$$\boldsymbol{X}_i = e_{\boldsymbol{a}_i} e_{\boldsymbol{b}_i}^T,$$

$$\boldsymbol{a}_i, \boldsymbol{b}_i \overset{i.i.d.}{\sim} \operatorname{Uniform}(\{1, \ldots, d\}),$$

$$\boldsymbol{w}_i \sim \mathcal{N}(0, \sigma^2).$$

Note that, using the notation above, $\langle \boldsymbol{X}_i, \Theta^\star\rangle = \Theta^\star_{\boldsymbol{a}_i, \boldsymbol{b}_i}$.

We want to focus on the case $\Omega(d \log d) \le n \ll d^2$. The lower bound of $d \log d$ ensures that with high probability every row and column contains at least one observation (see the «coupon collector's problem»). Note that, on the other hand, with d^2 samples most of the entries of $\Theta^\star$ are observed, which makes the problem much easier. Also note that, because $\Theta^\star$ is a rank-r matrix, it is specified by $2 \cdot r \cdot d$ unknowns, so we also expect to need $r \cdot d \ll n$.

The maximum likelihood estimator (MLE) for this problem is

$$\hat{\boldsymbol{\Theta}}_{\text{MLE}} = \arg\min_{\Theta} \left\{ \sum_{i=1}^{n} (\boldsymbol{y}_i - \langle X_i, \Theta\rangle)^2 \mid \operatorname{rank}(\Theta) \le r \right\}.$$

Unfortunately, computing the MLE is NP-hard in the worst case.

3 What we want

First, we want to have a polynomial-time computable estimator $\hat{\boldsymbol{\Theta}}$.

Second, we want this estimator to achieve small error $\frac{1}{d^2}\left\|\hat{\boldsymbol{\Theta}} - \Theta^\star\right\|_F^2$. What upper bound can we hope to achieve on this error?

For a reasonable estimator, we might expect that

$$\frac{1}{d^2}\left\|\hat{\boldsymbol{\Theta}} - \Theta^\star\right\|_F^2 \approx \frac{1}{n}\sum_{i=1}^{n} (\langle \boldsymbol{X}_i, \hat{\boldsymbol{\Theta}}\rangle - \langle \boldsymbol{X}_i, \Theta^\star\rangle)^2.$$

That is, we might expect the average error over all the entries to be approximately equal to the average error over all the observed entries. Note the similarity of the right-hand side to the least-squares error! Then, we might expect to obtain error

$$\frac{1}{d^2}\left\|\hat{\mathbf{\Theta}} - \Theta^\star\right\|_F^2 \leq O\left(\frac{\sigma^2 \cdot r \cdot d}{n}\right),$$

which is the least-squares error if one of the two rank-r factors of $\Theta^\star$ was known, where $r \cdot d$ comes from the number of unknowns that specifies the other rank-r factor.

However, as we will see, something is missing in this bound.

3.1 Obstruction

Consider the following example. Let $r = 1$ and $\sigma = 0$, and suppose that

$$\Theta^\star \in \{\Theta \in \mathbb{R}^d \mid \Theta_{1,1} = \pm M, \Theta_{i,j} = 0 \text{ for } (i,j) \neq (1,1)\}.$$

That is, $\Theta^\star$ has a single non-zero entry at $(1,1)$ with value M or $-M$ for some known M. For instance, think of $M = 2^d$. The goal is to determine the sign of $\Theta_{1,1}$.

Then, if $n \ll d^2$, with high probability entry $(1,1)$ is not observed in any of the samples. In this case $\boldsymbol{y}_i = 0$ for all i, so we have no information about the sign of $\Theta^\star_{1,1}$. All we can do is guess. Then, with probability at least $\frac{1}{2}$, the error is

$$\frac{1}{d^2}\|\hat{\mathbf{\Theta}} - \Theta^\star\|_F^2 \geq \frac{M^2}{d^2}.$$

More precisely,

$$\max_{\Theta^\star \in \{\pm M \cdot e_1 e_1^T\}} \mathbb{E}\frac{1}{d}\|\hat{\mathbf{\Theta}} - \Theta^\star\|_F^2 \geq \Omega(\frac{M^2}{d^2}).$$

To fix this, we will allow the error to depend on $\|\Theta^\star\|_{\max} = \max_{a,b}|\Theta^\star_{a,b}|$.

4 Polynomial-time estimator

We will obtain a suitable estimator by changing the objective function of the MLE optimization so that it becomes polynomial-time computable.

Concretely, we consider an estimator of the following form for a d-by-d random matrix $\boldsymbol{Y}$,

$$\hat{\mathbf{\Theta}} = \operatorname*{arg\,min}_{\Theta \in \mathbb{R}^{d\times d}} \left\{\|\Theta - \boldsymbol{Y}\|_F^2 \mid \operatorname{rank}(\Theta) \leq r\right\}.$$

Note that given $\boldsymbol{Y}$, we can efficiently compute the estimator $\hat{\boldsymbol{\Theta}}$ by truncating all but the first r terms of a singular-value decomposition of $\boldsymbol{Y}$.

How do we want to choose $\boldsymbol{Y}$? Some criteria are:

- we want $\boldsymbol{Y}$ to be a simple function of $\{(\boldsymbol{y}_i, \boldsymbol{X}_i)\}_{i=1}^n$
- we want $\boldsymbol{Y}$ to satisfy $\mathbb{E}\boldsymbol{Y} = \Theta^\star$

Choose

$$\boldsymbol{Y} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{y}_i \cdot d^2 \cdot \boldsymbol{X}_i.$$

Note that this satisfies

$$\mathbb{E}\boldsymbol{y}_i \cdot d^2 \cdot \boldsymbol{X}_i = \frac{1}{d^2}\sum_{a,b=1}^d \Theta^\star_{a,b} \cdot d^2 \cdot e_a e_b^T = \Theta^\star.$$

Theorem 1. If $n \geq r \cdot d \log d$, then with high probability

$$\frac{1}{d^2}\|\hat{\boldsymbol{\Theta}} - \Theta^\star\|_F^2 \leq O((\sigma^2 + \|\Theta^\star\|_{\max}^2) \cdot \frac{r \cdot d \log d}{n}).$$

Note that the upper bound in Theorem 1 contains the $\frac{\sigma^2 \cdot r \cdot d}{n}$ component that we guessed earlier could be the desired error. In addition, the upper bound contains another component that scales with $\|\Theta^\star\|_{\max}^2$.

The proof of the theorem will require a matrix version of the Bernstein tail bound, which is introduced below. Then we prove the theorem above.

5 Matrix Bernstein tail bound

Let $\boldsymbol{Z}$ be a random $d \times d$ matrix with $d \geq 2$. Assume that $\mathbb{E}\boldsymbol{Z} = 0$ and $\{\boldsymbol{Z}\} = \{\boldsymbol{Z}^T\}$. The latter means that the distribution of $\boldsymbol{Z}$ and its transpose are identical.

Definition. The random matrix $\boldsymbol{Z}$ satisfies the (σ_Z, b_Z)-Bernstein condition if

$$\forall j \in \mathbb{N}_{\geq 2}, \quad \left\|\mathbb{E}\|\boldsymbol{Z}\|^{j-2} \cdot \boldsymbol{Z}\boldsymbol{Z}^T\right\| \leq j! \cdot b_Z^{j-2} \cdot \sigma_Z^2.$$

Theorem 2. If $\boldsymbol{Z}$ satisfies the (σ_Z, b_Z)-Bernstein condition, then with probability $1 - d^{-100}$, for $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n \overset{i.i.d.}{\sim} \{\boldsymbol{Z}\}$,

$$\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i\right\| \lesssim \sigma_Z \cdot \sqrt{\frac{\log d}{n}} + b_Z \cdot \frac{\log d}{n}.$$

Proof. See Section 5.4 in [Wainwright, 2019].

6 Proof of Theorem 1

Let $\boldsymbol{U} = \hat{\boldsymbol{\Theta}} - \Theta^\star$ and $\boldsymbol{W} = \boldsymbol{Y} - \Theta^\star$. Note that the latter is equivalent to $\boldsymbol{Y} = \Theta^\star + \boldsymbol{W}$, so $\boldsymbol{W}$ can be interpreted as non-Gaussian noise.

Claim 1. $\|\boldsymbol{U}\|_F^2 \leq 8 \cdot r \cdot \|\boldsymbol{W}\|^2$.

The proof of Claim 1 is the same as for the multi-spike matrix model (see this note). Then, what remains is to bound $\|\boldsymbol{W}\|$.

Noise decompositon. We have that

$$\begin{aligned}
\boldsymbol{W} &= \boldsymbol{Y} - \Theta^\star \\
&= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i \cdot d^2 \cdot \boldsymbol{X}_i - \Theta^\star) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\langle \boldsymbol{X}_i, \Theta^\star\rangle \cdot d^2 \cdot \boldsymbol{X}_i - \Theta^\star + \boldsymbol{w}_i \cdot d^2 \cdot \boldsymbol{X}_i).
\end{aligned}$$

Let $\boldsymbol{Z}_i^{(1)} = \langle \boldsymbol{X}_i, \Theta^\star\rangle \cdot d^2 \cdot \boldsymbol{X}_i - \Theta^\star$ and $\boldsymbol{Z}_i^{(2)} = \boldsymbol{w}_i \cdot d^2 \cdot \boldsymbol{X}_i$. Then we write

$$\boldsymbol{W} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{Z}_i^{(1)} + \boldsymbol{Z}_i^{(2)}).$$

Finally, let $\boldsymbol{W}^{(1)} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i^{(1)}$ and $\boldsymbol{W}^{(2)} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i^{(2)}$.

Claim 2. If $n \geq d \log d$, then with high probability

(1) $\|\boldsymbol{W}^{(1)}\|^2 \leq O(\|\Theta^\star\|_{\max}^2 \cdot \frac{d^3 \log d}{n})$
(2) $\|\boldsymbol{W}^{(2)}\|^2 \leq O(\sigma^2 \cdot \frac{d^3 \log d}{n})$

Together, these two bounds give

$$\begin{aligned}
\|\boldsymbol{U}\|_F^2 &\leq 8r \cdot \|\boldsymbol{W}\|^2 \\
&\leq 8r \cdot (\|\boldsymbol{W}^{(1)}\| + \|\boldsymbol{W}^{(2)}\|)^2 \\
&\leq 16r \cdot (\|\boldsymbol{W}^{(1)}\|^2 + \|\boldsymbol{W}^{(2)}\|^2) \\
&\leq O((\sigma^2 + \|\Theta^\star\|_{\max}^2) \cdot \frac{r \cdot d^3 \log d}{n}).
\end{aligned}$$

Dividing both sides by $\frac{1}{d^2}$ completes the proof of Theorem 2. We prove now the two parts of Claim 2. In both parts, the goal is to apply the matrix Bernstein tail bound.

Proof of Claim 2, part 1 Let $\boldsymbol{Z} = \langle \boldsymbol{X}, \Theta^\star \rangle \cdot d^2 \cdot \boldsymbol{X} - \Theta^\star$, with $\boldsymbol{X} = e_{\boldsymbol{a}} e_{\boldsymbol{b}}^T$ for $\boldsymbol{a}$ and $\boldsymbol{b}$ uniform in $\{1, ..., d\}$. Note that $\boldsymbol{Z}_1^{(1)}, ..., \boldsymbol{Z}_n^{(1)} \overset{i.i.d.}{\sim} \boldsymbol{Z}$.

Now, we prove that $\boldsymbol{Z}$ satisfies Berstein condition with parameter $\sigma_Z^2 = d^3 \cdot \|\Theta^\star\|_{\max}^2$ and $b_Z = 2\|\Theta^\star\|_{\max} \cdot d^2$.

First, we have

$$\begin{aligned}
\|\boldsymbol{Z}\| &\leq \|\langle \boldsymbol{X}, \Theta^\star \rangle \cdot d^2 \cdot \boldsymbol{X}\| + \|\Theta^\star\| \\
&\leq 2\|\Theta^\star\|_{\max} \cdot d^2
\end{aligned}$$

where we used that $\|\langle \boldsymbol{X}, \Theta^\star \rangle \cdot d^2 \cdot \boldsymbol{X}\| \leq \|\Theta^\star\|_{\max} \cdot d^2$ and $\|\Theta^\star\| \leq \|\Theta^\star\|_{\max} \cdot d^2$.

Let $b_Z = 2\|\Theta^\star\|_{\max} \cdot d^2$, to be used in the Bernstein tail bound.

Second, we have the following. Note that below we use the Löwner order notation, where for matrices A and B, $A \preceq B$ denotes the fact that $B - A$ is positive semidefinite.

$$\begin{aligned}
\mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^T] &= \mathbb{E}[\langle \boldsymbol{X}, \Theta^\star \rangle^2 \cdot d^4 \cdot e_{\boldsymbol{a}} e_{\boldsymbol{b}}^T e_{\boldsymbol{b}} e_{\boldsymbol{a}}^T] - \Theta^\star(\Theta^\star)^T \\
&\overset{(1)}{\preceq} \mathbb{E}[\langle \boldsymbol{X}, \Theta^\star \rangle^2 \cdot d^4 \cdot e_{\boldsymbol{a}} e_{\boldsymbol{a}}^T] \\
&\overset{(2)}{\preceq} d^4 \cdot \|\Theta^\star\|_{\max}^2 \cdot \mathbb{E}[e_{\boldsymbol{a}} e_{\boldsymbol{a}}^T] \\
&= d^3 \cdot \|\Theta^\star\|_{\max}^2 \cdot I_d
\end{aligned}$$

where in (1) we used that $e_{\boldsymbol{b}}^T e_{\boldsymbol{b}} = 1$ and $\Theta^\star(\Theta^\star)^T \succeq 0$, in (2) we used that $\langle \boldsymbol{X}, \Theta^\star \rangle^2 \leq \|\Theta^\star\|_{\max}^2$, and in (3) we used that $\mathbb{E}[e_{\boldsymbol{a}} e_{\boldsymbol{a}}^T] = \frac{1}{d} I_d$.

Let $\sigma_Z^2 = d^3 \cdot \|\Theta^\star\|_{\max}^2$, to be used in the Bernstein tail bound.

We now verify the Bernstein condition:

$$0 \preceq \mathbb{E}[\|\boldsymbol{Z}\|^{j-2} \cdot \boldsymbol{Z}\boldsymbol{Z}^T] \preceq b_Z^{j-2} \cdot \sigma_Z^2 \cdot I_d,$$

so

$$\|\mathbb{E}[\|\boldsymbol{Z}\|^{j-2} \cdot \boldsymbol{Z}\boldsymbol{Z}^T]\| \leq b_Z^{j-2} \cdot \sigma_Z^2.$$

Finally, apply the Bernstein bound:

$$\begin{aligned}
\|\boldsymbol{W}^{(1)}\| &\leq O(\sigma_Z \cdot \sqrt{\frac{\log d}{n}} + b_Z \cdot \frac{\log d}{n}) \\
&= O\left(\|\Theta^\star\|_{\max} \cdot (\sqrt{\frac{d^3 \log d}{n}} + \frac{2d^2 \log d}{n})\right),
\end{aligned}$$

where the term with $\sqrt{\frac{d^3 \log d}{n}}$ dominates for $n \geq d \log d$.

Proof of Claim 2, part 2 Let $\boldsymbol{Z} = \boldsymbol{w} \cdot d^2 \cdot \boldsymbol{X}$, with $\boldsymbol{X} = e_{\boldsymbol{a}} e_{\boldsymbol{b}}^T$ for $\boldsymbol{a}$ and $\boldsymbol{b}$ uniform in $\{1, ..., d\}$, and with $\boldsymbol{w} \sim \mathcal{N}(0, \sigma^2)$. Note that $\boldsymbol{Z}_1^{(1)}, ..., \boldsymbol{Z}_n^{(1)} \overset{i.i.d.}{\sim} \boldsymbol{Z}$.

We have

$$\|\boldsymbol{Z}\| = |\boldsymbol{w}| \cdot d^2 \cdot \|\boldsymbol{X}\| = |\boldsymbol{w}| \cdot d^2.$$

We also have $\mathbb{E}|\boldsymbol{w}|^j \leq \sigma^j \cdot j!$ (see «Wikipedia»).

We now verify the Bernstein condition with $\sigma_Z^2 = \sigma^2 \cdot d^3$ and $b_Z = \sigma \cdot d^2$:

$$0 \leq \mathbb{E}\|\boldsymbol{Z}\|^{j-2} \cdot \boldsymbol{Z}\boldsymbol{Z}^T = (d^2)^j \cdot \mathbb{E}|\boldsymbol{w}|^j \cdot \mathbb{E} e_{\boldsymbol{a}} e_{\boldsymbol{b}}^T e_{\boldsymbol{b}} e_{\boldsymbol{a}}^T,$$

with $\mathbb{E}|\boldsymbol{w}|^j \leq \sigma^j \cdot j!$ and $\mathbb{E} e_{\boldsymbol{a}} e_{\boldsymbol{a}} e_{\boldsymbol{b}}^T e_{\boldsymbol{b}} e_{\boldsymbol{a}}^T = \mathbb{E} e_{\boldsymbol{a}} e_{\boldsymbol{a}}^T = \frac{1}{d} I_d$. Then

$$\begin{aligned}\left\|\mathbb{E}\|\boldsymbol{Z}\|^{j-2} \cdot \boldsymbol{Z}\boldsymbol{Z}^T\right\| &\leq (d^2)^j \cdot \sigma^j \cdot j! \cdot \frac{1}{d} \\ &= j! \cdot (\sigma \cdot d^2)^{j-2} \cdot \sigma^2 \cdot d^3 \\ &= j! \cdot b_Z^{j-2} \cdot \sigma_Z^2.\end{aligned}$$

Finally, apply the Bernstein bound:

$$\begin{aligned}\|\boldsymbol{W}^{(2)}\| &\leq O(\sigma_Z \cdot \sqrt{\frac{\log d}{n}} + b_Z \cdot \frac{\log d}{n}) \\ &= O\left(\sigma \cdot (\sqrt{\frac{d^3 \log d}{n}} + \frac{d^2 \log d}{n})\right),\end{aligned}$$

where the term with $\sqrt{\frac{d^3 \log d}{n}}$ dominates for $n \geq d \log d$.

7 Chapter notes

One of the first works that established rigorous guarantees for matrix completion was Candès and Recht [2009].

The exposition in this note is based mainly on a subsequent work Koltchinskii et al. [2011].

References

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009. doi: 10.1007/s10208-009-9045-5. URL https://doi.org/10.1007/s10208-009-9045-5.

Vladimir Koltchinskii, Karim Lounici, and Alexandre B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5), Oct 2011. ISSN 0090-5364. doi: 10.1214/11-aos894. URL http://dx.doi.org/10.1214/11-aos894.

Martin J. Wainwright. *High-dimensional statistics*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019. ISBN 978-1-108-49802-9. doi: 10.1017/9781108627771. URL https://doi.org/10.1017/9781108627771. A non-asymptotic viewpoint.