Matrix Completion

A brief overview of Section 3.8 of Chen et al. (2021)

Zhiling Gu

Iowa State University

November 21, 2022

Motivation I

In the practice, it is extremely common to encounter missing data due to collection difficulty, erroneous data, and etc. And most of the data can be represented in the matrix. For example, if we consider each row of a matrix is the features/ measurements of a single subject, a matrix would represent the features of all the subjects/ population of interest. To tackle the missing data problem, one of the tool is matrix completion.

Motivation II

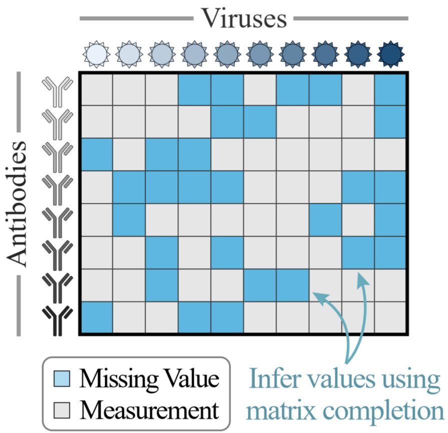

Figure 1: source: `https://www.fredhutch.org/en/news/spotlight/2022/08/bs-einav-cellsys.html`

Preparation I

Norms

- For any vector $\boldsymbol{v}$, we denote by $\|\boldsymbol{v}\|_2$, $\|\boldsymbol{v}\|_1$ and $\|\boldsymbol{v}\|_\infty$ its ℓ_2 norm, ℓ_1 norm and ℓ_∞ norm, respectively.
- For any matrix $\boldsymbol{A} = \left[A_{i,j}\right]_{1\le i\le m, 1\le j\le n}$, we let $\|\boldsymbol{A}\|$, $\|\boldsymbol{A}\|_*$, $\|\boldsymbol{A}\|_{\mathrm{F}}$ and $\|\boldsymbol{A}\|_\infty$ represent respectively its spectral norm (i.e., the largest singular value of $\boldsymbol{A}$), its nuclear norm (i.e., the sum of singular values of $\boldsymbol{A}$), its Frobenius norm (i.e., $\|\boldsymbol{A}\|_{\mathrm{F}} := \sqrt{\sum_{i,j} A_{i,j}^2}$), and its entrywise ℓ_∞ norm (i.e., $\|\boldsymbol{A}\|_\infty := \max_{i,j}\left|A_{i,j}\right|$). We also refer to $\|\boldsymbol{A}\|_{2,\infty}$ as the $\ell_{2,\infty}$ norm of $\boldsymbol{A}$, defined as $\|\boldsymbol{A}\|_{2,\infty} := \max_i \left\|\boldsymbol{A}_{i,\cdot}\right\|_2$. Similarly, we define the $\ell_{\infty,2}$ norm of $\boldsymbol{A}$ as $\|\boldsymbol{A}\|_{\infty,2} := \left\|\boldsymbol{A}^\top\right\|_{2,\infty}$.
- Singular values of $\boldsymbol{M}$ are square roots of the eigenvalues of $\boldsymbol{M}^H\boldsymbol{M}$.
- The largest singular value $\sigma_1(\boldsymbol{M})$= operator norm $\|\boldsymbol{M}\|_{op} := \max_{\|x\|_2=1}\|\boldsymbol{M}x\|_2$.
- In addition, for any matrices $\boldsymbol{A} = \left[A_{i,j}\right]_{1\le i\le m, 1\le j\le n}$ and $\boldsymbol{B} = \left[B_{i,j}\right]_{1\le i\le m, 1\le j\le n}$, the inner product of $\boldsymbol{A}$ and $\boldsymbol{B}$ is defined as and denoted by $\langle\boldsymbol{A}, \boldsymbol{B}\rangle = \sum_{1\le i\le m, 1\le j\le n} A_{i,j}B_{i,j} = \mathrm{Tr}\left(\boldsymbol{A}^\top\boldsymbol{B}\right)$.

Preparation II

Consider $\boldsymbol{M} = \boldsymbol{M}^* + \boldsymbol{E}$ and $\boldsymbol{M}^*$ be two matrices of $\mathbb{R}^{n_1 \times n_2}$, $n_1 \le n_2$. Let $\boldsymbol{M}^* = \boldsymbol{U}^* \boldsymbol{\Sigma}^* \boldsymbol{V}^*$, $\boldsymbol{M} = \boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{V}$ as follows

$$\boldsymbol{M}^* = \sum_{i=1}^{n_1} \sigma_i^\star u_i^\star \boldsymbol{v}_i^{\star\top} = \left[\boldsymbol{U}^* \, \boldsymbol{U}^*{}_\perp\right] \begin{bmatrix} \boldsymbol{\Sigma}^* & 0 & 0 \\ 0 & \boldsymbol{\Sigma}^*{}_\perp & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{V}^{\star\top} \\ \boldsymbol{V}_\perp^{\star\top} \end{bmatrix};$$

$$\boldsymbol{M} = \sum_{i=1}^{n_1} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top = \left[\boldsymbol{U} \, \boldsymbol{U}_\perp\right] \begin{bmatrix} \boldsymbol{\Sigma} & 0 & 0 \\ 0 & \boldsymbol{\Sigma}_\perp & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{V}^\top \\ \boldsymbol{V}_\perp^\top \end{bmatrix}.$$

Here, $\sigma_1 \ge \cdots \ge \sigma_{n_1}$ (resp. $\sigma_1^\star \ge \cdots \ge \sigma_{n_1}^\star$) stand for the singular values of M (resp. $M^\star$) arranged in descending order, $\boldsymbol{u}_i$ (resp. $\boldsymbol{u}_i^\star$) denotes the left singular vector associated with the singular value σ_i (

resp. $\sigma_i^\star$), and $\boldsymbol{v}_i$ (resp. $\boldsymbol{v}_i^\star$) represents the right singular vector associated with σ_i (resp. $\sigma_i^\star$). In addition, we denote

$$
\begin{aligned}
\boldsymbol{\Sigma} &:= \operatorname{diag}\left([\sigma_1, \cdots, \sigma_r]\right), & \boldsymbol{\Sigma}_\perp &:= \operatorname{diag}\left([\sigma_{r+1}, \cdots, \sigma_{n_1}]\right), \\
\boldsymbol{U} &:= [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r] \in \mathbb{R}^{n_1 \times r}, & \boldsymbol{U}_\perp &:= [\boldsymbol{u}_{r+1}, \cdots, \boldsymbol{u}_{n_1}] \in \mathbb{R}^{n_1 \times (n_1 - r)}, \\
\boldsymbol{V} &:= [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_r] \in \mathbb{R}^{n_2 \times r}, & \boldsymbol{V}_\perp &:= [\boldsymbol{v}_{r+1}, \cdots, \boldsymbol{v}_{n_2}] \in \mathbb{R}^{n_2 \times (n_2 - r)}
\end{aligned}
$$

The matrices $\boldsymbol{\Sigma}^\star, \boldsymbol{\Sigma}_\perp^\star, \boldsymbol{U}^\star, \boldsymbol{U}_\perp^\star, \boldsymbol{V}^\star, \boldsymbol{V}_\perp^\star$ are defined analogously. In addition, we define the distance between two matrices as

$$
\operatorname{dist}\left(\boldsymbol{U}, \boldsymbol{U}^\star\right) := \min_{\boldsymbol{R} \in \mathcal{O}^{r \times r}} \left\| \boldsymbol{U}\boldsymbol{R} - \boldsymbol{U}^\star \right\| \tag{1}
$$

Problem formulation and assumption I

Suppose the data matrix $\mathbf{M}^*$ is of dimension $n_1 \times n_2$ with rank r. Assume

$$n_1 \leq n_2.$$

We start with the single value decomposition of $\boldsymbol{M}^*$ as follows

$$\boldsymbol{M}^* = \boldsymbol{U}^* \boldsymbol{\Sigma}^* \boldsymbol{V}^{*\top},$$

where $col(\boldsymbol{U}^*) \in \mathbb{R}^{n_1 \times r}$, $col(\boldsymbol{V}^*) \in \mathbb{R}^{n_2 \times r}$, and $\boldsymbol{\Sigma}^*$ is a diagonal matrix with entries singular values, denoted as $\sigma_1(\boldsymbol{M}^*), \ldots, \sigma_r(\boldsymbol{M}^*)$ in descending order. And we introduce *condition number* of matrix $\boldsymbol{M}^*$ to be

$$\kappa := \sigma_1(\boldsymbol{M}^*)/\sigma_r(\boldsymbol{M}^*),$$

and we define an index subset $\Omega \subset [n_1] \times [n_2]$ such that $(i,j) \in \Omega \iff \boldsymbol{M}^*{}_{ij}$ is observed.

Problem formulation and assumption II

Assumption 1 (Random sampling). In this report, we assume each entry of $\boldsymbol{M}^*$ is observed independently with probability $0 < p < 1$. This corresponds to *missing at random* in statistics terminology.

Example (Incoherence). Here we provide an example that satisfies random sampling but causes unfaithful recovery. Consider $\boldsymbol{M}^*$ being a zero matrix except for 1 entry. If $p = o(1)$, then with high probability, the single nonzero entry would be missing, and any recovery method would be in vain to recover the rank 1 property.

$$\begin{pmatrix} 10000 \\ 00000 \\ 00000 \\ 00000 \\ 00000 \end{pmatrix}$$

Problem formulation and assumption III

μ**-incoherent.** Motivated by the previous example, we define the *incoherence parameter* μ of $\boldsymbol{M}^*$ as follows

$$\mu := \max \left\{ \frac{n_1 \|\boldsymbol{U}^*\|_{2,\infty}^2}{r}, \frac{n_2 \|\boldsymbol{V}^*\|_{2,\infty}^2}{r} \right\}.$$

Recall that $\|\boldsymbol{U}^*\|_{2,\infty} = \max_i \|\boldsymbol{U}^*{}_{i,\cdot}\|_2$ is the largest ℓ_2 norm among rows of $\boldsymbol{U}^*$. Also note by SVD, $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$ are unitary matrices, and thus $\boldsymbol{U}^* \boldsymbol{U}^{*\top} = \mathbf{I}_r$ leading to $\|\boldsymbol{U}^*\|_F^2 = r$.

$$\frac{r}{n_1} = \frac{1}{n_1} \|\boldsymbol{U}^*\|_F^2 \leq \|\boldsymbol{U}^*\|_{2,\infty}^2 \leq \|\boldsymbol{U}^*\|^2 = 1$$

$$\implies 1 \leq \mu \leq \max\{n_1, n_2\}/r = n_2/r.$$

A smaller μ indicates the energy of singular vectors is spread out across different elements.

Algorithm I

Euclidean projection operator: $\mathcal{P}_\Omega : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$. It is now natural to define a projection from original space $\mathbb{R}^{n_1 \times n_2}$ where $\boldsymbol{M}^*$ lies in a subspace of $\mathbb{R}^{n_1 \times n_2}$ as follows:

$$[\mathcal{P}_\Omega(\boldsymbol{M}^*)]_{ij} = \begin{cases} \boldsymbol{M}^*{}_{ij}, & \text{if}(i,j) \in \Omega \\ 0, & \text{else.} \end{cases}$$

And our goal is to recover $\boldsymbol{M}^*$ on the basis of $\mathcal{P}_\Omega(\boldsymbol{M}^*)$.
Example:

$$\text{Observed matrix} = \begin{pmatrix} 1\,?\,?\,2 \\ 2\,1\,3\,1 \\ 4\,1\,1\,? \end{pmatrix}$$

$$\mathcal{P}_\Omega(\boldsymbol{M}^*) = \begin{pmatrix} 1\,0\,0\,2 \\ 2\,1\,3\,1 \\ 4\,1\,1\,0 \end{pmatrix}$$

Algorithm II

Algorithm: Under the assumption of random sampling, we consider an approximation $\boldsymbol{M}^*$, $\boldsymbol{M}$, through *inverse probability weighting* of observed data matrix

$$\boldsymbol{M} := p^{-1}\mathcal{P}_\Omega(\boldsymbol{M}^*). \tag{2}$$

Since the observed data is in the random subspace $\mathcal{P}_\Omega(\boldsymbol{M}^*)$, $\boldsymbol{M}$ is in fact a random approximation matrix. This construction leads to

$$\mathbb{E}_\Omega(\boldsymbol{M}) = \boldsymbol{M}^*.$$

Then we compute rank-r SVD of $\boldsymbol{M} = \boldsymbol{U}\Sigma\boldsymbol{V}^\top$, and $\boldsymbol{U}$, $\boldsymbol{V}$ are employed as the estimates of $\boldsymbol{U}^*$, $\boldsymbol{V}^*$, respectively.

Example of inverse probability weighting

$$\text{True matrix } \boldsymbol{M}^* = \begin{pmatrix} 1222 \\ 2131 \\ 4113 \end{pmatrix}$$

$$\text{Observed matrix} = \begin{pmatrix} 1??2 \\ 2131 \\ 411? \end{pmatrix}$$

$$\mathcal{P}_\Omega(\boldsymbol{M}^*) = \begin{pmatrix} 1002 \\ 2131 \\ 4110 \end{pmatrix}$$

$$\text{Approximation matrix } \boldsymbol{M} := p^{-1}\mathcal{P}_\Omega(\boldsymbol{M}^*)$$

Assume p and r are known. $\boldsymbol{U\Sigma V}$ is the rank-r SVD of $\mathbf{M}$.
We ask: how close is $\boldsymbol{U\Sigma V}$ and $\boldsymbol{M}^*$?

Useful bounds of matrix norms

Lemma 1 (Lemma 3.20 of Chen et al. (2021))

Assume $\boldsymbol{M}^* \in \mathbb{R}^{n_1 \times n_2}$ *is* μ*-coherent. Then the following relations holds*

$$\|\boldsymbol{M}^*\|_{2,\infty} \leq \sqrt{\mu r / n_1} \|\boldsymbol{M}^*\| \tag{3}$$

$$\|\boldsymbol{M}^{*\top}\|_{2,\infty} \leq \sqrt{\mu r / n_2} \|\boldsymbol{M}^*\| \tag{4}$$

$$\|\boldsymbol{M}^*\|_{\infty} \leq \mu r \sqrt{1 / n_1 n_2} \|\boldsymbol{M}^*\|. \tag{5}$$

Perturbation bound of $\boldsymbol{M}$

Lemma 2 (Lemma 3.21 of Chen et al. (2021))

Suppose $n_2 p \geq C\mu r \log n_2$ for some constant $C > 0$, then with probability at least $1 - O(n_2^{-10})$, one has

$$\|\boldsymbol{M} - \boldsymbol{M}^*\| \lesssim \sqrt{\frac{\mu r \log n_2}{n_1 p}} \|\boldsymbol{M}^*\|.$$

The higher the probability of observation p is, the better the bound is.

Recovery of $\boldsymbol{U}$, $\boldsymbol{V}$

Theorem 3 (Theorem 3.22 of Chen et al. (2021))

Suppose $n_1 p \geq C\kappa^2 \mu r \log n_2$ for some constant $C > 0$, then with probability at least $1 - O(n_2^{-10})$, one has

$$\max\left\{\operatorname{dist}\left(\boldsymbol{U}, \boldsymbol{U}^\star\right), \operatorname{dist}\left(\boldsymbol{V}, \boldsymbol{V}^\star\right)\right\} \lesssim \kappa\sqrt{\frac{\mu r \log n_2}{n_1 p}}.$$

Note that when the sample size $p n_1 n_2 \gg \kappa^2 \mu r n_2 \log n_2$, the spectral estimate achieves consistent estimation $\max\left\{\operatorname{dist}\left(\boldsymbol{U}, \boldsymbol{U}^\star\right), \operatorname{dist}\left(\boldsymbol{V}, \boldsymbol{V}^\star\right)\right\} = o_p(1)$.

Recovery of $\boldsymbol{M}$

Theorem 4 (Theorem 3.23 of Chen et al. (2021))

Suppose $n_2 p \geq C\mu r \log n_2$ *for some constant* $C > 0$*, then with probability at least* $1 - O(n_2^{-10})$*, one has*

$$\|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top - \boldsymbol{M}^*\|_F \lesssim \sqrt{\frac{\mu r^2 \log n_2}{n_1 p}}\|\boldsymbol{M}^*\|$$

The theorem above only requires Lemma 2 and characterizes the statistical accuracy of $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$.

Wedin's Theorem

Theorem 5 (Wedin $\sin\Theta$ theorem for singular subspace perturbation, Theorem 3.22 of Chen et al. (2021))

If $\|\boldsymbol{E}\| < \sigma_r^\star - \sigma_{r+1}^\star$, then one has

$$\max\left\{\operatorname{dist}\left(\boldsymbol{U}, \boldsymbol{U}^\star\right), \operatorname{dist}\left(\boldsymbol{V}, \boldsymbol{V}^\star\right)\right\} \leq \frac{\sqrt{2}\max\left\{\left\|\boldsymbol{E}^\top \boldsymbol{U}^\star\right\|, \left\|\boldsymbol{E}\boldsymbol{V}^\star\right\|\right\}}{\sigma_r^\star - \sigma_{r+1}^\star - \|\boldsymbol{E}\|}.$$

Proof of Theorem 3.22 of Chen et al. (2021) I

- Sketch of the proof: (i) Prove $\boldsymbol{E} = \boldsymbol{M} - \boldsymbol{M}^*$ satisfy $\|\boldsymbol{E}\| < \sigma_r(\boldsymbol{M}^*) - \sigma_{r+1}(\boldsymbol{M}^*)$, where $\sigma_r(\boldsymbol{M}^*)$ is the r-th largest singular value of $\boldsymbol{M}^*$. (ii) Apply Wedin's theorem to $\boldsymbol{E}$ and use lemma 2.
- Step (i): recall $n_1 \geq n_2$, thus the condition of lemma 2 is satisfied. Then

$$\|\boldsymbol{E}\| = \|\boldsymbol{M} - \boldsymbol{M}^*\| \lesssim \sqrt{\frac{\mu r \log n_2}{n_1 p}} \|\boldsymbol{M}^*\|.$$

In addition, recall that $\sigma_1(\boldsymbol{M}^*) = \|\boldsymbol{M}^*\| = \kappa \sigma_r(\boldsymbol{M}^*)$ by definition of singular value and κ. Therefore

$$\begin{aligned}\|\boldsymbol{E}\| &\lesssim \sqrt{\frac{\mu r \log n_2}{n_1 p}} \|\boldsymbol{M}^*\| = \sqrt{\frac{\kappa^2 \mu r \log n_2}{n_1 p}} \sigma_r(\boldsymbol{M}^*) \\ &\leq \frac{1}{C} \sigma_r(\boldsymbol{M}^*) \text{ for some large enough } C > 0.\end{aligned}$$

Choose C such that $1/C < 1 - 1/\sqrt{2}$, we have

$$\|\boldsymbol{E}\| \lesssim (1 - \frac{1}{\sqrt{2}}) \sigma_r(\boldsymbol{M}^*).$$

Note we can always choose a large enough C such that the condition of Wedin's theorem $\|\boldsymbol{E}\| < \sigma_r(\boldsymbol{M}^*) - \sigma_{r+1}(\boldsymbol{M}^*)$ holds.

Proof of Theorem 3.22 of Chen et al. (2021) II

- Step (ii): Apply Wedin's theorem to $\boldsymbol{E}$, we have

$$
\begin{aligned}
&\max\left\{\operatorname{dist}\left(\boldsymbol{U}, \boldsymbol{U}^{\star}\right), \operatorname{dist}\left(\boldsymbol{V}, \boldsymbol{V}^{\star}\right)\right\} \\
&\leq \frac{\sqrt{2}\max\left\{\left\|\boldsymbol{E}^{\top}\boldsymbol{U}^{\star}\right\|, \left\|\boldsymbol{E}\boldsymbol{V}^{\star}\right\|\right\}}{\sigma_r(\boldsymbol{M}^{*}) - \sigma_{r+1}(\boldsymbol{M}^{*}) - \|\boldsymbol{E}\|} \text{ by Wedin's theorem} \\
&\leq \frac{\sqrt{2}\|\boldsymbol{E}\|\max\left\{\left\|\boldsymbol{U}^{\star}\right\|, \left\|\boldsymbol{V}^{\star}\right\|\right\}}{\sigma_r(\boldsymbol{M}^{*}) - \|\boldsymbol{E}\|} \text{ by } \|AB\| \leq \|A\|\|B\| \\
&\leq \frac{\sqrt{2}\|\boldsymbol{E}\|}{\sigma_r^{\star} - (1 - \frac{1}{\sqrt{2}})\sigma_r(\boldsymbol{M}^{*})} \text{ by unitary matrix } \boldsymbol{U}^{*}, \boldsymbol{V}^{*} \\
&= 2\|\boldsymbol{E}\|/\sigma_r(\boldsymbol{M}^{*}) = 2\kappa\|\boldsymbol{E}\|/\sigma_1(\boldsymbol{M}^{*}) \\
&\lesssim \kappa\sqrt{\frac{\mu r \log n_2}{n_1 p}} \text{ by Lemma 2.}
\end{aligned}
$$

Proof of Theorem 3.23 of Chen et al. (2021)

- By triangle inequality, we have

$$\|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top - \boldsymbol{M}^*\| \le \|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top - \boldsymbol{M}\| + \|\boldsymbol{M} - \boldsymbol{M}^*\|.$$

  Note that $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is the SVD of $\boldsymbol{M}$ and thus the best rank-r approximation to $\boldsymbol{M}$. Therefore $\|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top - \boldsymbol{M}\| \le \|\boldsymbol{M} - \boldsymbol{M}^*\|$, where $\boldsymbol{M}^*$ is an unknown rank-r matrix.
- In addition, since both $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ and $\boldsymbol{M}^*$ are of rank r, the difference between them would have rank at most $2r$. This leads to

$$\begin{aligned}
\|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top - \boldsymbol{M}^*\| &\le \|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top - \boldsymbol{M}\| + \|\boldsymbol{M} - \boldsymbol{M}^*\| \\
&\le 2\|\boldsymbol{M} - \boldsymbol{M}^*\| \\
\|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top - \boldsymbol{M}^*\|_F &\le \sqrt{2r}\|\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top - \boldsymbol{M}^*\| \text{ by Remark *} \\
&\le 2\sqrt{2r}\|\boldsymbol{M} - \boldsymbol{M}^*\| \\
&\lesssim \sqrt{\frac{\mu r^2 \log n_2}{n_1 p}}\|\boldsymbol{M}^*\| \text{ by Lemma 2.}
\end{aligned}$$

- Remark *: $\|\boldsymbol{M}\|_F \le \sqrt{r}\|\boldsymbol{M}\|$. This can be proved as follows. $\|\boldsymbol{M}\|_F = \sqrt{\sum_i (\boldsymbol{e}_i^\top \boldsymbol{M}^\top \boldsymbol{M}\boldsymbol{e}_i)} = \sqrt{\sum_i (\boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{\Sigma}^2 \boldsymbol{V}^\top \boldsymbol{e}_i)} \le \sqrt{\sum_{i=1}^r \|\boldsymbol{e}_i^\top \boldsymbol{U}\|_2 \|\boldsymbol{M}\|^2 \|\boldsymbol{V}^\top \boldsymbol{e}_i\|_2} = \sqrt{r}\|\boldsymbol{M}\|$.

Proof of Lemma 3.20 of Chen et al. (2021) I

- We start the proofs of auxiliary lemmas with the following basic inequalities of matrix norms.

$$
\begin{aligned}
&(i): \|\boldsymbol{AB}\|_{2,\infty} \leq \|\boldsymbol{A}\|_{2,\infty}\|\boldsymbol{B}\|,\\
&(ii): \|\boldsymbol{AB}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{B}\|,\\
&(iii): \|\boldsymbol{AB}^\top\|_\infty \leq \|\boldsymbol{A}\|_{2,\infty}\|\boldsymbol{B}\|.
\end{aligned}
$$

  - Define $\boldsymbol{e}_j$ as the indicator vector, where the j-th entry is one, zero elsewhere. Consider SVD $\boldsymbol{A} = \boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^\top$, and $\boldsymbol{B} = \boldsymbol{U}_2\boldsymbol{\Sigma}_2\boldsymbol{V}_2^\top$.
  - (i) $\|\boldsymbol{AB}\|_{2,\infty} = \max_i \|\boldsymbol{e}_i^\top\boldsymbol{AB}\|_2 = \max_i \|\boldsymbol{e}_i^\top\boldsymbol{A}\|_2\|\boldsymbol{U}_2\boldsymbol{\Sigma}_2\boldsymbol{V}_2^\top\|_2 \leq \|\boldsymbol{A}\|_{2,\infty}\|\boldsymbol{\Sigma}_2\|_2 \leq \|\boldsymbol{A}\|_{2,\infty}\|\boldsymbol{B}\|$.
  - (ii) $\|\boldsymbol{AB}\| = \|\boldsymbol{AB}\|_{op} = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \|\boldsymbol{AB}x\|_2/\|\boldsymbol{x}\|_2 = \max_{\boldsymbol{Bx}\neq\boldsymbol{0}} \|\boldsymbol{ABx}\|_2/\|\boldsymbol{Bx}\|_2 \max_{\boldsymbol{x}\neq\boldsymbol{0}} \|\boldsymbol{Bx}\|_2/\|\boldsymbol{x}\|_2 = \|\boldsymbol{A}\|_{op}\|\boldsymbol{B}\|_{op} = \|\boldsymbol{A}\|\|\boldsymbol{B}\|$.
  - (iii) $\|\boldsymbol{AB}^\top\|_\infty = \max_{ij} |\boldsymbol{e}_i^\top\boldsymbol{AB}^\top\boldsymbol{e}_j| \leq \max_i \|\boldsymbol{e}_i^\top\boldsymbol{A}\|_2\|\boldsymbol{B}^\top\boldsymbol{e}_j\|_2 = \|\boldsymbol{A}\|_{2,\infty}\|\boldsymbol{B}\|_{2,\infty}$ by Cauchy-Schwartz inequality. In addition, since $\|\boldsymbol{B}\|_{2,\infty} \leq \|\boldsymbol{B}\|$, we proved the third inequality.

Proof of Lemma 3.20 of Chen et al. (2021) II

- Equipped with the inequalities above, we consider

$$\begin{aligned}
\|\boldsymbol{M}^*\|_{2,\infty} &= \|\boldsymbol{U}^*\boldsymbol{\Sigma}^*\boldsymbol{V}^{*\top}\|_{2,\infty} \\
&\le \|\boldsymbol{U}^*\|_{2,\infty}\|\boldsymbol{\Sigma}^*\|\|\boldsymbol{V}^*\| \text{ by (i)\& (ii)} \\
&\le \frac{\sqrt{\mu r}}{\sqrt{n_1}}\|\boldsymbol{M}^*\| \text{ by definition of coherence parameter} \mu \ge n_1\|\boldsymbol{U}^*\|_{2,\infty}^2/r,
\end{aligned}$$

- Secondly,

$$\begin{aligned}
\|\boldsymbol{M}^*\|_{\infty} &= \|\boldsymbol{U}^*\boldsymbol{\Sigma}^*\boldsymbol{V}^{*\top}\|_{\infty} \\
&\le \|\boldsymbol{U}^*\|_{2,\infty}\|\boldsymbol{\Sigma}^*\|\|\boldsymbol{V}^*\|_{2,\infty} \text{ by (i)\& (iii)} \\
&\le \frac{\sqrt{\mu r}}{\sqrt{n_1}}\|\boldsymbol{M}^*\|\frac{\sqrt{\mu r}}{\sqrt{n_2}} = \frac{\mu r}{\sqrt{n_1 n_2}}\|\boldsymbol{M}^*\|.
\end{aligned}$$

Proof of Lemma 3.21 of Chen et al. (2021) I

- Sketch of proof: (i) Decompose elements of $\boldsymbol{E}$ as sum of independent random matrices $\boldsymbol{X}_{ij}$. (ii) Apply matrix Bernstein inequality to $\boldsymbol{E}$.
- Step (i): Recall that $\boldsymbol{E} = p^{-1}\mathcal{P}_{\Omega}(\boldsymbol{M}^*) - \boldsymbol{M}^*$, and can be written as follows

$$p^{-1}\mathcal{P}_{\Omega}(\boldsymbol{M}^*) - \boldsymbol{M}^* = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\boldsymbol{X}_{ij}$$

$$\boldsymbol{X}_{ij} = (p^{-1}\delta_{ij} - 1)M_{ij}^*\boldsymbol{e}_i\boldsymbol{e}_j^\top,$$

where $\delta_{ij} \sim Ber(p)$ is indicator random variable for that (i,j)-th entry is observed; $\boldsymbol{e}_i$ is the i-th standard basis vector of appropriate dimension. It cann be seen that

$$\mathbb{E}(\boldsymbol{X}_{ij}) = \boldsymbol{0}, \quad \|\boldsymbol{X}_{ij}\| \leq \frac{1}{p}\|\boldsymbol{M}^*\|_\infty \leq \frac{\mu r}{p\sqrt{n_1 n_2}}\|\boldsymbol{M}^*\|,$$

by Lemma 1.

Proof of Lemma 3.21 of Chen et al. (2021) II

Theorem 6 (Matrix Bernstein, Corollary 3.3 of Chen et al. (2021))

Let $\{\boldsymbol{X}_i\}_{1\leq i\leq m}$ be a set of independent real random matrices with dimension $n_1 \times n_2$. Suppose that

$$\mathbb{E}\left[\boldsymbol{X}_i\right] = \boldsymbol{0}, \quad \text{and} \quad \|\boldsymbol{X}_i\| \leq L, \quad \text{for all } i.$$

For any $a \geq 2$, with probability exceeding $1 - 2n^{-a+1}$ one has

$$\left\|\sum_{i=1}^{m} \boldsymbol{X}_i\right\| \leq \sqrt{2av\log n} + \frac{2a}{3} L\log n,$$

where $n := \max\{n_1, n_2\}$, and variance statistic

$$\begin{aligned} v := \max\Bigg\{ & \left\|\sum_{i=1}^{m} \mathbb{E}\left[\left(\boldsymbol{X}_i - \mathbb{E}\left[\boldsymbol{X}_i\right]\right)\left(\boldsymbol{X}_i - \mathbb{E}\left[\boldsymbol{X}_i\right]\right)^\top\right]\right\|, \\ & \left\|\sum_{i=1}^{m} \mathbb{E}\left[\left(\boldsymbol{X}_i - \mathbb{E}\left[\boldsymbol{X}_i\right]\right)^\top\left(\boldsymbol{X}_i - \mathbb{E}\left[\boldsymbol{X}_i\right]\right)\right]\right\|\Bigg\}. \end{aligned} \tag{6}$$

Proof of Lemma 3.21 of Chen et al. (2021) III

- Step (ii): Apply matrix Bernstein inequality (Theorem 6), take $L = \frac{\mu r}{p\sqrt{n_1 n_2}} \|\boldsymbol{M}^*\|$,

$$\|\boldsymbol{E}\| \leq \sqrt{2av \log n_2} + \frac{2a}{3} \frac{\mu r}{p\sqrt{n_1 n_2}} \|\boldsymbol{M}^*\| \log n_2, \quad \forall a > 2,$$

where

$$v = \max\left\{ \left\| \sum_{ij} \mathbb{E}\left[(\boldsymbol{X}_{ij}) (\boldsymbol{X}_{ij})^\top \right] \right\|, \left\| \sum_{ij} \mathbb{E}\left[(\boldsymbol{X}_{ij})^\top (\boldsymbol{X}_{ij}) \right] \right\| \right\}.$$

Proof of Lemma 3.21 of Chen et al. (2021) IV

For the first term in v, we have

$$\begin{aligned}
\sum_{ij} \mathbb{E}(\boldsymbol{X}_{ij}\boldsymbol{X}_{ij}^\top) &= \sum_{ij} \mathbb{E}\left\{(p^{-1}\delta_{ij} - 1)^2 (M_{ij}^*)^2 \boldsymbol{e}_i\boldsymbol{e}_j^\top \boldsymbol{e}_j\boldsymbol{e}_i^\top\right\} \\
&= \frac{1-p}{p}\sum_{ij}(M_{ij}^*)^2\boldsymbol{e}_i\boldsymbol{e}_i^\top \text{ by random sampling } \delta_{ij} \sim Ber(p) \\
&= \frac{1-p}{p}\sum_{i=1}^{n_1} \|\boldsymbol{M}^*{}_{i,\cdot}\|_2^2 \boldsymbol{e}_i\boldsymbol{e}_i^\top \\
&\preceq \frac{1-p}{p}\|\boldsymbol{M}^*\|_{2,\infty}^2 \sum_{i=1}^{n_1} \boldsymbol{e}_i\boldsymbol{e}_i^\top \\
&\preceq \frac{\mu r}{n_1 p}\|\boldsymbol{M}^*\|^2 \boldsymbol{I}_{n_1} \text{ by Lemma 1}
\end{aligned}$$

where $\boldsymbol{A} \preceq \boldsymbol{B} \iff \boldsymbol{B} - \boldsymbol{A}$ is positive semidefinite. Similarly, we can derive

$$\sum_{ij} \mathbb{E}(\boldsymbol{X}_{ij}^\top \boldsymbol{X}_{ij}) \preceq \frac{\mu r}{n_2 p}\|\boldsymbol{M}^*\|^2 \boldsymbol{I}_{n_2}.$$

Proof of Lemma 3.21 of Chen et al. (2021) V

Thus, we can bound v as follows

$$v \leq \frac{\mu r}{n_1 p} \|\boldsymbol{M}^*\|^2$$

by noting that $n_1 \leq n_2$. Combine the bound of v and the result of Bernstein inequality, we have

$$\|\boldsymbol{E}\| \lesssim \sqrt{\frac{\mu r \|\boldsymbol{M}^*\|^2 \log n_2}{n_1 p}} + \frac{\mu r \|\boldsymbol{M}^*\| \log n_2}{p\sqrt{n_1 n_2}}$$

with probability at least $1 - O(n_2^{-10})$ by setting $a = 11$. Since $\log n_2 \ll \sqrt{n_2}$, the second term above diminishes as n_2 becomes large. In particular, when $n_2 \gtrsim \mu r \|\boldsymbol{M}^*\|^2 \log n_2$, the first term dominates the second term, which leads to

$$\|\boldsymbol{E}\| \lesssim \sqrt{\frac{\mu r \|\boldsymbol{M}^*\|^2 \log n_2}{n_1 p}}.$$

References

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Spectral methods for data science: A statistical perspective. Foundations and Trends® in Machine Learning, 14(5):566–806, 2021. doi: 10.1561/2200000079. URL https://doi.org/10.1561%2F2200000079.