A Zero-imputation Approach in Recommendation Systems with Data Missing Heterogeneously

Jiashen Lu, Kehui Chen

University of Pittsburgh

Supplementary Material

Proof of Theorem 1, Corollary 2, Theorem 2, and Theorem 3.

S1 Proof of Theorem 1

Proof of Theorem 1. First consider

$$\frac{1}{mn}\|\hat{\mathrm{A}} - \mathrm{P}\|_F^2, \tag{S1.1}$$

where $\hat{\mathrm{A}}$ is soft-threshold estimator, $\mathrm{P} = \mathbb{E}(\mathrm{A})$ is the population parameter matrix, and $\|\cdot\|_F$ denotes the matrix Frobenius norm. Here A is a general notation for the truncation matrix $A^{(k)}$ or $A_{(k)}$. The proof of this part mainly follows Lemma 1 in Xu (2018). Let the error matrix $\mathrm{E} = \mathrm{A} - \mathrm{P}$ and let $\|\mathrm{E}\|$ denote the spectral norm of E. With the notation that $\mathrm{P} = \delta_{n,m}\tilde{\mathrm{P}}_{i,j}$, we have $\mathrm{Var}(\mathrm{E}_{i,j}) = \delta_{n,m}\tilde{\mathrm{P}}_{i,j} - \delta_{n,m}^2\tilde{\mathrm{P}}_{i,j}^2 \le \delta_{n,m}$.

Let σ_r and $\sigma_r(\mathrm{A})$ be the r-th singular values of $\tilde{\mathrm{P}}$ and A. By Lemma 2 in Xu (2018), we know that there exist some positive constants c_1 and η', such that the following event happens with probability at least 1-n^{-c_1}.

$$Event = \{\|\mathrm{E}\| \leq \eta' \sqrt{\delta_{n,m} n}\}. \tag{S1.2}$$

Note that to apply this lemma, we need the assumption that $\delta_{n,m}$ is lowered bounded by $c_2 \frac{\log(n)}{n}$ for some positive constant c_2, i.e., $\delta_{n,m} \geq c_2 \frac{\log(n)}{n}$.

On Equation (S1.2), consider the singular value threshold for some positive constant c_0,

$$\lambda = (1 + c_0)\eta' \sqrt{\delta_{n,m} n}, \tag{S1.3}$$

which means we only keep the singular values of A that are greater than λ for the soft-threshold procedure and $\|\mathrm{E}\| \leq \frac{1}{1+c_0}\lambda$. Consider

$$\ell = \sup\{r : \delta_{n,m}\sigma_r \geq \frac{c_0}{1 + c_0}\lambda\}. \tag{S1.4}$$

If $\ell = m$, it is easy to check the result. Now assume $\ell < m$, by Weyl's Theorem,

$$\sigma_{\ell+1}(\mathrm{A}) \leq \delta_{n,m}\sigma_{\ell+1} + \|\mathrm{E}\| < \lambda,$$

which implies the rank of $\hat{\mathrm{A}}$ is bounded by ℓ. Let P_ℓ denote the best rank ℓ approximation to P, then

$$\begin{aligned}
\|\hat{\mathrm{A}} - \mathrm{P}\|_F^2 &\leq 2\|\hat{\mathrm{A}} - \mathrm{P}_\ell\|_F^2 + 2\|\mathrm{P}_\ell - \mathrm{P}\|_F^2 \\
&\leq 4\ell\|\hat{\mathrm{A}} - \mathrm{P}_\ell\|^2 + 2\delta_{n,m}^2 \sum_{i=\ell+1} \sigma_i^2 \\
&\leq 16\ell\lambda^2 + 2\delta_{n,m}^2 \sum_{i=\ell+1} \sigma_i^2 \\
&\leq 16 \min_{0\leq r\leq m} \{r\lambda^2 + (\frac{1+c_0}{c_0})^2 \sum_{i=r+1} \delta_{n,m}^2\sigma_i^2\}.
\end{aligned}$$

The second to last inequality holds since

$$\|\hat{\mathrm{A}} - \mathrm{P}_\ell\| \leq \|\hat{\mathrm{A}} - \mathrm{A}\| + \|\mathrm{A} - \mathrm{P}\| + \|\mathrm{P} - \mathrm{P}_\ell\| \leq 2\lambda. \tag{S1.5}$$

The last inequality holds since $\delta_{n,m}\sigma_{\ell+1} \leq \frac{c_0}{1+c_0}\lambda$ and by the definition that the last line in inequality has minimum value at ℓ. Therefore, on event Equation (S1.2), there exist some constants C_1, C_2, such that

$$\frac{1}{mn}\|\hat{\mathrm{A}} - \mathrm{P}\|_F^2 \leq \frac{C_1 \min_{0\leq r\leq m} \{r\lambda^2 + C_2 \sum_{i=r+1} \delta_{n,m}^2\sigma_i^2\}}{mn}. \tag{S1.6}$$

Recall that for each rating k, we recover the upper probability

$$\begin{aligned}
\hat{S}_{i,j}^k =& \frac{\hat{A}_{i,j}^{(k)}}{\max\{\hat{A}_{i,j}^{(k)} + \hat{A}_{(k)i,j}, \varepsilon_{n,m}\}} \\
=& \frac{\mathbb{E}(A_{i,j}^{(k)})}{\mathbb{E}(A_{i,j}^{(k)}) + \mathbb{E}(A_{(k);i,j})} + f_x(\xi,\eta)(\hat{A}_{i,j}^{(k)} - \mathbb{E}(A_{i,j}^{(k)})) + \\
& f_y(\xi,\eta)(\max\{\hat{A}_{i,j}^{(k)} + \hat{A}_{(k);i,j}, \varepsilon_{n,m}\} - \mathbb{E}(A_{i,j}^{(k)}) - \mathbb{E}(A_{(k);i,j})),
\end{aligned}$$

where $f(x,y)=\frac{x}{y}$, $f_x(x,y)=\frac{1}{y}$, $f_y(x,y)=\frac{-x}{y^2}$ and $(\xi,\eta)^T$ is some point in the line segment between the true value and the estimated value, i.e. there exists some value t between 0 and 1 such that $[\xi,\eta]^T = t[\mathbb{E}(A_{i,j}^{(k)}), \max\{\mathbb{E}A_{i,j}^{(k)}+\mathbb{E}\hat{A}_{(k);i,j}, \varepsilon_{n,m}\}]^T+(1-t)[\hat{A}_{i,j}^{(k)}, \max\{\hat{A}_{i,j}^{(k)}+\hat{A}_{(k);i,j}, \varepsilon_{n,m}\}]^T$. The expectation element $\mathbb{E}_{i,j}$ corresponds to $\mathrm{P}_{i,j}$ that appeared previously. The absolute value of two partial derivatives are bounded by $\frac{1}{\eta}$, since $\frac{\xi}{\eta^2}\leq\frac{1}{\eta}$.

Note that η is a point between true observation probability and the estimated probability. By the assumption that the true value is lower bounded by $c\delta_{n,m}$ and assumption that $\varepsilon_{n,m}$ is $c'\delta_{n,m}$ $(c'<c)$, the partial derivatives are upper bounded by $\frac{1}{c_3\delta_{n,m}}$ for some constant c_3. So the overall MSE is

$$\frac{1}{mn}\sum_{i,j}(\hat{S}_{i,j}^k - P(S_{i,j}\geq k))^2 \leq \min_{0\leq r\leq m}\{\frac{C_3 r}{m\delta_{m,n}}+\frac{C_4\sum_{i=r+1}\sigma_i^2}{mn}\}. \tag{S1.7}$$

□

S2 Proof of Corollary 2

Proof of Corollary 2. From the proof in Theorem 1, we know that for the minimum point ℓ, we have $\delta_{n,m}\sigma_\ell \geq c\sqrt{\delta_{n,m}n}$ and $\delta_{n,m}\sigma_{\ell+1} < c'\sqrt{\delta_{n,m}n}$. Use the assumption that $\sigma_\ell \asymp \frac{\sqrt{mn}}{\ell^\alpha}$, we have $\ell \asymp (m\delta_{n,m})^{1/(2\alpha)}$. Therefore the first term $\frac{\ell}{m\delta_{n,m}}$ in MSE is in the order of $(\frac{1}{m\delta_{n,m}})^{1-\frac{1}{2\alpha}}$. For the singular value summation term, using the fact that

$$\sum_{r=\ell+1}^{n\wedge m} r^{-2\alpha} = O(\frac{1}{\ell^{2\alpha-1}}),$$

we conclude that the second term in MSE is in the order of $(\frac{1}{m\delta_{n,m}})^{1-\frac{1}{2\alpha}}$. □

S3 Proof of Theorem 2

Proof of Theorem 2. Suppose the corresponding graphon W admits strong SVD in the form of

$$W(s,t) = \sum_i \lambda_i \phi_i(s)\psi_i(t).$$

Let s and t be i.i.d. $Unif(0,1)$, and let $u(s) = [u_1(s), \ldots, u_r(s), \ldots]^T$ in which $u_r(s) = \sqrt{\lambda_r}\phi_r(s)$, and $v(t) = [v_1(t), \ldots, v_r(t), \ldots]^T$ in which $v_r(t) = \sqrt{\lambda_r}\psi_r(t)$. The norm of each random variable is finite by the strong decomposition assumption. Moreover, $W(s,t) = u(s)^T v(t)$ almost everywhere. The sampling distribution generated by W with dimension n, m is, by Aldous-Hoover Theorem, first samples $s_1, \ldots, s_n$ and $t_1, \ldots, t_m$ from i.i.d. $Unif(0,1)$, then generate Bernoulli random variables with parameters $W(s_i, t_j)$. This is, by the construction, the same as first independently sampling from the BGRD distribution to get $u(s_i)$ and $v(t_j)$, then form the exchangeable arrays by their inner-products, where F_1 is the probability measure induced by $u(s) : [0,1] \to K$ with $s \sim Unif(0,1)$ and F_2 is the probability measure induced by $v(t) : [0,1] \to K$ with $t \sim Unif(0,1)$. □

S4 Proof of Theorem 3

Proof of Theorem 3. ($\Leftarrow$) Since orthogonal transform maintains inner product, this direction is clear.

($\Rightarrow$) By Proposition 3.5 in Lei (2021), for a distribution F_1 on a separable Hilbert space

K, there exists an inverse transform sampling, i.e., a measurable function $u : [0,1] \to K$ such that if $s \sim Unif(0,1) \Rightarrow u(s) \sim F_1$. Therefore, for a sampling point in BGRD $F = F_1 \times F_2$, we can write it as $(u(s), v(t))$, where u and v are inverse transform samplings, and $s, t \sim Unif(0,1)$. By equally-weighted assumption and without loss of generality, we assume that (u, v) have the same diagonal second moment matrix Λ. Analogously, we denote a sample point from G by $(\tilde{u}(s), \tilde{v}(t))$, and their moment matrix $\tilde{\Lambda}$.

Define the graphon W corresponding to F as

$$\begin{aligned} W(s,t) &= \langle u(s), v(t) \rangle \\ &= \sum_j \lambda_j \lambda_j^{-1/2} u_j(s) \lambda_j^{-1/2} v_j(t), \end{aligned}$$

where λ_j is the jth diagonal value in Λ. Note that the above is the SVD decomposition of W. We can define $\tilde{W}$ similarly for G. Since F and G lead to the same sampling distribution of binary arrays, we have

$$W(s,t) \overset{d}{=} \tilde{W}(s,t).$$

By Theorem 4.1 in Kallenberg (1989), we have $\forall\ j,\ \lambda_j = \tilde{\lambda}_j$ and there exists unitary operator Q with $Q_{j,j'} = 0$ for $\lambda_j \neq \lambda_{j'}$, such that for any measurable set A,

$$P(\Lambda^{-1/2} u \in A) = P(Q \Lambda^{-1/2} \tilde{u} \in A),$$

$$P(\Lambda^{-1/2} v \in A) = P(Q \Lambda^{-1/2} \tilde{v} \in A).$$

Therefore

$$P(u \in A) = P(\Lambda^{-1/2}u \in \Lambda^{-1/2}A)$$

$$= P(Q\tilde{u} \in A).$$

The same result holds for v. Therefore $F \overset{o.t.}{=} G$. $\square$

Bibliography

Kallenberg, O. (1989), "On the representation theorem for exchangeable arrays," *Journal of Multivariate Analysis*, 30, 137–154.

Lei, J. (2021), "Network representation using graph root distributions," *The Annals of Statistics*, 49, 745–768.

Xu, J. (2018), "Rates of convergence of spectral methods for graphon estimation," in *International Conference on Machine Learning*, PMLR, pp. 5433–5442.