Appendix: Convolutional Imputation of Matrix Networks

Qingyun Sun [*1] **Mengyuan Yan** [*2] **David Donoho** [3] **Stephen Boyd** [2]

Exact recovery guarantee

Theorem 1. *We assume that A is a matrix network on a graph G, and its graph Fourier transform $\hat{A}(k)$ are a sequence of matrices, each of them is at most rank r, and $\hat{A}$ satisfy the incoherence condition with coherence μ. And we observe a matrix network A^Ω on the graph G, for a subset of node in Ω random sampled from the network, node i on the network is sampled with probability p_i, we define the average sampling rate $p = \frac{1}{N}\sum_{i=1}^N p_i = |\Omega|/(Nn^2)$, and define $\mathcal{R} = \frac{1}{p}P_\Omega \mathcal{U}^*$.*

Then we prove that for any sampling probability distribution $\{p_i\}$, as long as the average sampling rate $p > C\mu\frac{r}{n}\log^2(Nn)$ for some constants C, the solution to the optimization problem

$$\begin{array}{ll} \underset{\hat{M}}{\textit{minimize}} & \|\hat{M}\|_{*,1}, \\ \textit{subject to} & A^\Omega = \mathcal{R}\hat{M} \end{array}$$

is unique and is exactly $\hat{A}$ with probability $1 - (Nn)^{-\gamma}$,where $\gamma = \frac{\log(Nn)}{16}$.

Proof. We define a inner product: $\langle \hat{M}_1, \hat{M}_2\rangle = \sum_k \langle \hat{M}_1(k), \hat{M}_2(k)\rangle$. Then we have the following two inequalities

$$\|\hat{M}(k)\|_* = \mathbf{Tr}(\text{sgn}(\hat{M}(k))\hat{M}(k)) = \langle \text{sgn}(\hat{M}(k)), \hat{M}(k)\rangle.$$

Therefore,

$$\|\hat{M}\|_{*,1} = \langle \text{sgn}(\hat{M}), \hat{M}\rangle.$$

Here $\text{sgn}(\hat{M}) = V_1 V_2^*$ is the sign matrix of the singular values of $\hat{M}$ under the singular vector basis.

We consider $\Delta = \hat{M} - \hat{A}$, then either $\mathcal{R}\Delta \neq 0$, or $\|\hat{A} + \Delta\|_{*,1} > \|\hat{A}\|_{*,1}$.

First we define a decomposition $\Delta = \Delta_T + \Delta_T^\perp = P_T\Delta + P_{T^\perp}\Delta$.

For $\mathcal{R}\Delta = 0$, we compute

$$\begin{array}{ll} & \|\hat{A} + \Delta\|_{*,1} \\ \geq & \|P_1(\hat{A}+\Delta)P_2\|_{*,1} + \|P_1^\perp(\hat{A}+\Delta)P_2^\perp\|_{*,1} \\ = & \|\hat{A} + P_1\Delta P_2\|_{*,1} + \|\Delta_T^\perp\|_{*,1} \\ \geq & \langle \text{sgn}(\hat{A}), \hat{A} + P_1\Delta P_2\rangle + \langle \text{sgn}(\Delta_T^\perp), \Delta_T^\perp\rangle \\ = & \|\hat{A}\|_{*,1} + \langle \text{sgn}(\hat{A}), P_1\Delta P_2\rangle + \langle \text{sgn}(\Delta_T^\perp), \Delta_T^\perp\rangle \\ = & \|\hat{A}\|_{*,1} + \langle \text{sgn}(\hat{A}) + (\Delta_T^\perp), \Delta\rangle. \end{array}$$

Now we want to estimate $\langle(\hat{A}) + (\Delta_T^\perp), \Delta\rangle$. We make two assumptions, which we will prove later.

First, we assume that for all $\Delta \in \text{range}(\mathcal{R})^\perp$, with probability $1 - (Nn)^{-\gamma}$,

$$\|\Delta_T\|_2 < 2nN\|\Delta_T^\perp\|_2.$$

Second, we want to construct a dual certificate $K \in \text{range}(\mathcal{R})$, such that for $k = 3 + \frac{1}{2}\log_2(r) + \log_2(n) + \log_2(N)$, with probability $1 - (Nn)^{-\gamma}$,

$$\begin{array}{rl} \|P_T(K) - \mathbf{sign}(\hat{A})\|_2 \leq & (\frac{1}{2})^k\sqrt{r}, \\ \|P_{T^\perp}(K)\| \leq & \frac{1}{2}. \end{array}$$

Then

$$\begin{array}{ll} & \langle \text{sgn}(\hat{A}) + (\Delta_T^\perp), \Delta\rangle \\ = & \langle \text{sgn}(\hat{A}) + (\Delta_T^\perp) - K, \Delta\rangle \\ = & \langle \text{sgn}(\hat{A}) - K, \Delta_T\rangle + \langle(\Delta_T^\perp) - K, \Delta_T^\perp\rangle \\ \geq & \dfrac{1}{2}\|\Delta_T^\perp\|_2 - (\dfrac{1}{2})^k\sqrt{r}\|\Delta_T\|_2 \\ \geq & \dfrac{1}{4}\|\Delta_T^\perp\|_2. \end{array}$$

When $\hat{M}$ is a minimizer, we must have $\Delta_T^\perp = 0$, otherwise $\|\hat{A} + \Delta\|_{*,1} < \|\hat{A}\|_{*,1}$. By assumption, $\|\Delta_T\|_2 < n^2\|\Delta_T^\perp\|_2.$, $\Delta_T = 0$, then $\Delta = 0$. Therefore, under the two assumption, $\hat{M}$ is the unique mininizer, and $\hat{M} = \hat{A}$.

Now we prove the above assumption and construct dual certificate.

[*]Equal contribution [1]Department of Mathematics, Stanford University, California, USA [2]Department of Electrical Engineering, Stanford University, California, USA [3]Department of Statistics, Stanford University, California, USA. Correspondence to: Qingyun Sun <qysun@stanford.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

First, we show that if

$$\|\Delta_T\|_2 \geq (2nN)\|\Delta_T^{\perp}\|_2,$$

then $\|\mathcal{R}\Delta_T\|_2 > \|\mathcal{R}\Delta_T^{\perp}\|_2$,

$$\begin{aligned} & \|\mathcal{R}\Delta\|_2 \\ = & \|\mathcal{R}\Delta_T + \mathcal{R}\Delta_T^{\perp}\|_2 \\ \geq & \|\mathcal{R}\Delta_T\|_2 - \|\mathcal{R}\Delta_T^{\perp}\|_2 \\ > & 0. \end{aligned}$$

We have a lower bound on $\|\mathcal{R}\Delta_T\|_2$ and upper bound on $\|\mathcal{R}\Delta_T^{\perp}\|_2$.

$$\|\mathcal{R}\Delta_T^{\perp}\|_2^2 \leq \|\mathcal{R}\|^2\|\Delta_T^{\perp}\|_2^2.$$

Here $\|\mathcal{R}\|$ is the operator norm of $\mathcal{R}$.

$$\begin{aligned} \|\mathcal{R}\Delta_T\|_2^2 &= \langle \mathcal{R}\Delta_T, \mathcal{R}\Delta_T \rangle \\ &\geq \|\mathcal{R}\|^2/(nN)^2(1 - \|P_T - P_T\mathcal{R}P_T\|)\|\Delta_T\|_2^2. \end{aligned}$$

Since $E(P_T\mathcal{R}P_T) = P_T$, we only need to control the deviation, we could use a concentration inequality called operator-Bernstein inequality (1),

$$\mathbf{P}[\|P_T - P_T\mathcal{R}P_T\| > t] \leq \exp(-\frac{npt^2}{4\mu r}).$$

Using the condition that $p = C\mu\frac{r}{n}\log^2(Nn)$, let $t = 1/4$, we have

$$\begin{aligned} & \mathbf{P}[\|P_T - P_T\mathcal{R}P_T\| > t] \\ \leq & \exp(-\tfrac{n\mu\frac{r}{n}\log^2(Nn)}{16\mu r}) \\ = & \exp(-\tfrac{\log^2(Nn)}{16}) \\ = & (nN)^{-\gamma}, \end{aligned}$$

where $\gamma = \frac{\log(Nn)}{16}$. Therefore, with probability $1 - (nN)^{-\gamma}$, the the inequality holds for $t = 1/2$. When the inequality holds, $\|P_T - P_T\mathcal{R}P_T\| < 1/2$, $\mathcal{R}\Delta \neq 0$.

Second, we construct the dual certificate K by the following construction: We decompose Ω as the union of k subset Ω_t, where each entry is sampled independently so that $E(|\Omega_t| = p_t = 1 - (1-p)^{1/k}$, and define $R_t = \frac{1}{p_t}P_{\Omega_t}\mathcal{U}^*$. Define

$$H_0 = (\hat{A}), K_t = \sum_{j=1}^{t} R_j H_{j-1}, H_t = (\hat{A}) - P_T K_t.$$

Then the dual certificate is defined as $K = K_k$.

This construction is called golfing scheme, which is invented in (1). Since $p_t = p/k = C\mu\frac{r}{nk}\log^2(Nn)$, we can assume $\|P_T - P_T\mathcal{R}_jP_T\| < 1/2$, which is true with probability $1 - \exp(\frac{Cnpt^2}{\mu kr})$.

$$\|H_t\|_2 \leq \|P_T - P_T\mathcal{R}P_T\|\|H_{t-1}\|_2 \leq \frac{1}{2}\|H_{t-1}\|_2.$$

And

$$\|P_T(K) - (\hat{A})\|_2 = \|H_k\| \leq (\frac{1}{2})^k\|(\hat{A})\| \leq (\frac{1}{2})^k\sqrt{r}.$$

Then

$$\|P_T(K) - (\hat{A})\|_2 \leq (\frac{1}{2})^k\sqrt{r}.$$

Also, $\|P_{T^{\perp}}(K)\| \leq \sum_{j=1}^{k}\|P_{T^{\perp}}R_jH_{j-1}\|$, use the operator-Bernstein inequality for a sequence of $t_j = 1/(4\sqrt{r})$, we have $\|P_{T^{\perp}}R_jH_{j-1}\| \leq t_i\|H_{j-1}\|_2$, and since $\|H_j\|_2 \leq \sqrt{r}2^{-j}$, then

$$\|P_{T^{\perp}}(K)\| \leq \sum_{j=1}^{k} t_i\|H_{j-1}\|_2 \leq \frac{1}{4}\sum_{j=1}^{k} 2^{-(j-1)} < 1/2.$$

Therefore, K is the dual certificate, the whole proof is done.

□

Imputation algorithm convergence Now we show that the solution of our imputation algorithm converges asymptotically to a minimizer of the previously defined objective $L_\lambda(\hat{M})$.

Each step of our imputation algorithm is minimizing a surrogate $Q_\lambda(\hat{M}|\hat{M}^{\text{old}})$ of the above objective function as

$$\|A^{\Omega} + P_{\Omega}^{\perp}\mathcal{U}^{-1}\hat{M}^{\text{old}} - \mathcal{U}^{-1}\hat{M}\|^2 + \sum_{k=1}^{N}\lambda_k\|\hat{M}(k)\|_*.$$

The resulting minimizer forms a sequence $\hat{M}_\lambda^t$ with starting point $\hat{M}_\lambda^0$

$$\hat{M}_\lambda^{t+1} = \quad \text{argmin} \quad Q_\lambda(\hat{M}|\hat{M}_\lambda^t).$$

Theorem 2. *The imputation algorithm produces a sequence of iterates $\hat{M}_\lambda^t$ that converges to the minimizer of $L_\lambda(\hat{M})$.*

The main idea of the proof is to show that Q_λ decreases after every iteration and $\hat{M}_\lambda^t$ is a Cauchy sequence, and the limit point is a stationary point of L_λ.

Proof. For each iteration in our algorithm, we are solving for a surrogate of the objective function as

$$Q_\lambda(\hat{M}|\hat{M}^{\text{old}}) = \|A^{\Omega} + P_{\Omega}^{\perp}\mathcal{U}^{-1}\hat{M}^{\text{old}} - \mathcal{U}^{-1}\hat{M}\|^2 + \sum_{k=1}^{N}\lambda_k\|\hat{M}(k)\|_*.$$

And the sequence $\hat{M}_\lambda^t$ with any starting point $\hat{M}_\lambda^0$ is given by

$$\hat{M}_\lambda^{t+1} = \quad \text{argmin} \quad Q_\lambda(\hat{M}|\hat{M}_\lambda^t).$$

The sequence satisfies

$$L_\lambda(\hat{M}_\lambda^{t+1}) \leq Q_\lambda(\hat{M}_\lambda^{t+1}|\hat{M}_\lambda^t) \leq L_\lambda(\hat{M}_\lambda^t).$$

Because

$$Q_\lambda(\hat{M}_\lambda^{t+1}|\hat{M}_\lambda^{t+1}) = L_\lambda(\hat{M}_\lambda^{t+1})$$

and

$$\begin{aligned} & Q_\lambda(\hat{M}|\hat{M}^{\text{old}}) \\ = \quad & \|P_\Omega(A) + P_\Omega^\perp \mathcal{U}^{-1}\hat{M}^{\text{old}} - \mathcal{U}^{-1}\hat{M}\|^2 + \sum_{k=1}^{N}\lambda_k\|\hat{M}(k)\|_* \\ \geq \quad & \|P_\Omega(A) + P_\Omega^\perp \mathcal{U}^{-1}\hat{M} - \mathcal{U}^{-1}\hat{M}\|^2 + \sum_{k=1}^{N}\lambda_k\|\hat{M}(k)\|_* \\ = \quad & Q_\lambda(\hat{M}|\hat{M}) \end{aligned}$$

Below we prove the following successive differences are monotonically decreasing

$$\|\hat{M}_\lambda^{t+1} - \hat{M}_\lambda^{t}\|^2 \leq \|\hat{M}_\lambda^{t} - \hat{M}_\lambda^{t-1}\|^2.$$

and the difference sequence converges to zero,

$$\hat{M}_\lambda^{t+1} - \hat{M}_\lambda^{t} \to 0.$$

The successive differences are monotonically decreasing because the soft threshold operator is a contraction in L_2 norm (2). And when there are positive singular values smaller than the threshold, the successive differences will strictly decrease until the algorithm converges.

Then $\hat{M}_\lambda^t$ is a Cauchy sequence, therefore we have a set of limit points. Also by monotonic convergence theorem, since $\hat{M}_\lambda^{t+1} - \hat{M}_\lambda^t$ converges to zero monotonically, the Cauchy sequence $\hat{M}_\lambda^t$ has an unique limit $\hat{M}_\lambda^\infty$. Moreover, we can verify that $\hat{M}_\lambda^\infty$ is a solution to the fixed point equation $\nabla L_\lambda = 0$, and a stationary point of $L_\lambda(\hat{M}_\lambda)$. Since $L_\lambda(\hat{M}_\lambda)$ is convex, each stationary point is a minimizer. Therefore, the convergence is proved. □

References

[1] David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

[2] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322, 2010.